Modern Martingale Methods: Theory and Applications

Justin Alexander Whitehouse

CMU-CS-24-159

December 2024

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Zhiwei Steven Wu (Co-chair)
Aaditya Ramdas (Co-chair)
Weina Wang
Aarti Singh
Csaba Szepesvári (University of Alberta)
Emilie Kaufmann (Université de Lille)

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

This research was sponsored by Edge Case Research, Inc., Google under purchase order number 4100275364, Pricewaterhouse Coopers LLP, the Department of Defense under award number FA8702-15-D-0002, and the National Science Foundation under award numbers CMMI-1938909, DMS-1945266, IIS-2125692, CNS-2120611, IIS-2229881, CNS-2339775, DGE-1745016, and DGE-2140739. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

Keywords: Applied Probability, Machine Learning, Martingale Methods, Sequential Inference, Online Learning, Differential Privacy

Abstract

Martingale concentration is at the heart of sequential statistical inference. Due to their time-uniform concentration of measure properties, martingales allow researchers to perform inference on highly correlated data as it is adaptively collected over time. Many state-of-the-art results in areas such as differential privacy, multi-armed bandit optimization, causal inference, and online learning boil down to (a) finding an appropriate, problem-dependent martingale and (b) carefully bounding its growth. Despite the important roles martingales and time-uniform concentration of measure play in modern statistical tasks, applications of martingale concentration is typically ad-hoc. Often, poorly chosen martingale concentration inequalities are applied, which results in suboptimal, even vacuous rates in sequential estimation problems.

The focus of this thesis is twofold. In the first part of this thesis, we provide simple yet powerful frameworks for constructing time-uniform martingale concentration inequalities in univariate, multivariate, and even sometimes infinite-dimensional settings. The inequalities contained herein can be applied to processes with both light-tailed and heavy-tailed increments, and follow from simple geometric arguments. The second part of this thesis is focused on applying martingale methods and time-uniform martingale concentration to practically relevant data science tasks. In particular, we show that, by appropriately applying martingale concentration, one can obtain salient improvements over the state-of-the-art in both differentially private machine learning and kernel bandit optimization tasks. In sum, the hope is to give a reader a start to finish view of how to derive and apply time-uniform martingale concentration in modern statistical research.

Contents

Chapter 1

Introduction

Classical methodology for statistical inference fails to live up to the sequential nature of modern data science. The rigidity of assumptions such as fixed sample sizes, i.i.d. data, and Gaussian noise is fundamentally misaligned with how data is structured in real-world settings. For instance, in contextual bandit learning tasks, a learner may want to sequentially estimate an unknown reward function as data is adaptively collected over time. Likewise, in A/B testing applications, a learner may stop experimentation early if there is sufficiently strong evidence of the efficacy of a treatment. Naively using traditional concentration of measure machinery such as finite sample or asymptotic confidence intervals may fail to yield valid coverage of the target estimand.

For a concrete example, suppose an e-commerce platform wants to measure how the deployment of a new design of a webpage (version 1) impacts user engagement relative to an existing control format (version 0). One way to measure this effect would be to conduct a randomized control trial: fix a sample size of, say, 1,000 individuals and randomly assign participants to each website version. If after measuring the level of interaction across all participants there is significant evidence to indicate that version 1 performs at least as well as version as version 0, it may be preferable to roll out the new design. However, if the new design makes it significantly more difficult for users to purchase goods, the company running the trial may experience significantly decreased revenue from running the experiment. Thus, it is clearly desirable from the perspective of the company to develop testing approaches that allow for early stopping in the case of sufficiently strong evidence against the new version.

How have researchers addressed the problem of calibrating confidence in the presence of highly correlated data and data-dependent stopping conditions? The answer is that they have largely turned to martingale methods. A stochastic process is a martingale if, given the history of the process up to time $n-1$, our best guess for the value of the process is at time n is simply its value at time $n-1$. In short, a martingale is just the generalization of an unbiased random walk. Examples of martingales include sums of independent, mean zero random variables, compensated Poisson processes, and geometric Brownian motion. Due to their central importance throughout this entire thesis, we formally define martingales below.

Definition 1.0.1. Let $(S_n)_{n\geq 0}$ be a real-valued, discrete time process adapted to some filtration $(\mathcal{F}_n)_{n\geq 0}$.[1] We say $(S_n)_{n\geq 0}$ is a martingale with respect to $(\mathcal{F}_n)_{n\geq 0}$ if

1. $\mathbb{E}(S_n \mid \mathcal{F}_{n-1}) = S_{n-1}$ for all $n \geq 1$, and
2. $\mathbb{E}|S_n| < \infty$.

We say $(S_n)_{n\geq 0}$ is a super-martingale if "$=$" in item 1 is replaced by "$\leq$".

The definitions of martingales and super-martingales extend naturally to continuous time processes $(S_t)_{t\geq 0}$ through the first condition being replaced by $\mathbb{E}(S_t \mid \mathcal{F}_s) = S_s$ (resp. $\leq S_s$) for all $s < t$. Likewise, a processes $(S_n)_{n\geq 0}$ taking values in a normed space $(\mathbb{B}, \|\cdot\|)$ is said to be a martingale if the second condition is replaced by $\mathbb{E}\|S_n\| \leq \infty$ for all $n \geq 0$.

Why are martingales so useful in sequential statistical tasks? The answer is that they offer strong, time-uniform concentration of measure properties. Whereas fixed-time concentration results for independent random variables are derived from Markov's inequality, time-uniform concentration results for observations with martingale dependence are derived from Ville's inequality, stated below.

Theorem 1.0.2 (Ville's Inequality)**.** *Let $(S_t)_{t\in\mathcal{T}}$ be a non-negative supermartingale with respect to some filtration $(\mathcal{F}_t)_{t\in\mathcal{T}}$ with $\mathbb{E}[S_0] = 1$ and $\mathcal{T} = \mathbb{N}$ or $\mathcal{T} = [0, \infty)$. Then, for any confidence parameter $\delta \in (0, 1)$,*

$$\mathbb{P}(\exists t \in \mathcal{T} : S_t \geq 1/\delta) \leq \delta.$$

This inequality ensures that the probability that a non-negative supermartingale ever crosses the horizontal line with intercept $1/\delta$ is bounded above by δ. While simple in statement, by carefully constructing a non-negative super-martingales, researchers have derived a variety of non-trivial time-uniform concentration of measure results. In particular, natural analogues of classical concentration results exist in the world of martingales. For instance Hoeffding's inequality for independent, bounded observations is replaced by Azuma's inequality[10, 70]. Likewise, Bernstein's/Bennet's inequality is replaced my Freedman's inequality [13, 64]. We present a detailed discussion of existing martingale concentration inequalities in Chapter 2. In general, researchers can mold Theorem 1.0.2 into an inequality that is useful for whatever scientific task is at hand.

There is thus a somewhat general recipe for performing time-uniform statistical inference. The first step is to find an emergent martingale in the task at hand. For example, to study the composition of differential private algorithms, it is natural to control the growth of *privacy loss martingales*, processes that measure the log-likelihood of the observed sequence of outcomes under two similar datasets. Likewise, in linear/contextual bandit problems, one can study the convergence of ridge regression estimates of the unknown reward function by bounding a "residual process" defined in terms of the noise in observations and the multivariate actions chosen by the learner. We talk about these processes respectively in Chapters 4 and 6. Often, finding the specific martingale to study is a bit of an art form requiring domain expertise, and so we try to provide insight in the chapters below.

We focus most our efforts on the second step of the process: appropriately controlling the growth of emergent martingales via time-uniform martingale concentration. This step is often

[1] A filtration $(\mathcal{F}_n)_{n\geq 0}$ of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is simply an increasing sequence of σ-algebras satisfying $\mathcal{F}_\infty := \bigcup_{n=0}^{\infty} \mathcal{F}_n \subset \mathcal{F}$.

poorly conducted, with statisticians either applying bounds that are not a good fit for the problem at hand or bounds that are derived in ad-hoc manner. Sometimes, the constructed bounds result in vacuous convergence rates, such as in the example of Chowdhury and Gopalan [30], who fail to show that Gaussian process upper confidence bound (GP-UCB) attains sub-linear regret for the practically relevant Matérn class of kernels. Other times, even in settings where researchers obtain optimal rates of concentration, the constants present in inappropriately-optimized bounds are typically prohibitively large. This is the case in Rogers et al. [133], who study the composition of differentially-private algorithms with adaptively-chosen privacy parameters, and obtain a bound with prohibitively large constants. From these few examples, it is clear that there needs to be a unified, simple to apply treatment of time-uniform martingale concentration in both univariate and multivariate settings. That is the goal of this thesis.

1.1 Contributions and Outline

This thesis is concerned with the development of a unified framework for time-uniform martingale concentration. In particular, we develop generic concentration inequalities for univariate, multivariate, and even infinite-dimensional processes. These results all follows from first principles and can applied to observations that are light-tailed (e.g. sub-Gaussian, sub-Gamma, sub-Poisson), heavy-tailed (infinite variance, bounded pth moment for $p < 2$), or somewhere in between (e.g. symmetric observations, observations lacking a moment generating function). A complementary focus of this thesis is the application of time-uniform martingale concentration to practically relevant tasks in statistics and machine learning. In particular, we find applications of time-uniform martingale concentration in differential privacy, private machine learning, and kernel bandit optimization. These applications aim to illustrate how to carefully and appropriately apply generic martingale concentration inequalities (such as those derived in the first part of the thesis) to machine learning tasks. A section-by-section enumeration of contributions is provided below.

Part I: Foundations of Time-Uniform Martingale Concentration

The first part of this thesis focuses on the development of time-uniform concentration inequalities for both finite and infinite-dimensional processes. In the finite-dimensional setting, the general goal is to derive inequalities that link the growth of a process $(S_n)_{n\geq 0}$ to some measure of accumulated variance $(V_n)_{n\geq 0}$ through a "sub-ψ" condition. Heuristically, ψ should roughly be thought of as representing the cumulant generating function (or CGF) of the increments of $(S_n)_{n\geq 0}$, but the presented results can handle some settings of heavy-tailed observations where a CGF may not exist. In the infinite-dimensional setting, we study the problem of heavy-tailed mean estimation instead of concentration, providing novel insight on how martingale methods relate to handling heavy-tailed observations that may lack finite variance. We now provide a more detailed description.

In the **first half of Chapter 2**, we study time-uniform, self-normalized concentration for arbitrary univariate sub-ψ processes. Self-normalized concentration, which aims to control the growth a processes when appropriately normalized by some empirical measure of variance, gen-

eralizes the study of concentration inequalities that depend on the number of observed samples. We prove a general self-normalized inequality that not only holds for processes admitting well-defined CGFs (such as those with sub-Gaussian, sub-Exponential, or sub-Gamma increments), but even holds for some heavy-tailed processes, such as those with finite variance or those with infinite variance but symmetric increments. The time-uniform bounds we prove are tight in that they asymptotically match the lower bound prescribed by the law of the iterated logarithm (LIL) with the correct leading constant. Our results can be viewed as a generalization of the contributions of Howard et al. [75], who prove a similar non-asymptotic LIL in the setting where the underlying process has sub-Gamma increments.

In the **second half of Chapter 2**, we extend the above results to multivariate processes. In particular, through leveraging a novel geometric argument, we are able to derive a time-uniform, self-normalized bound on general multivariate sub-ψ process. The bounds we prove, which depends on the condition number of the covariance matrix V_n, depart significantly from the traditional "method of mixtures" approaches commonly used in online learning tasks [41, 3], which only apply for sub-Gaussian processes. Using our bounds, we prove a corresponding law of the iterated logarithm for vector-valued processes, and we also construct a counterexample showing that this rate is, in general, tight.

Lastly, in **Chapter 3**, we study the problem of estimating an unknown mean associated with heavy-tailed observations in infinite-dimensional Banach spaces. In particular, we study a simple truncation-based estimator influenced by Catoni and Giulini [23]. This estimator first uses a small amount of data to construct a naive mean estimate, next projects the remaining observations onto an appropriately-sized ball centered at the naive mean estimate, and finally averages these truncated observations. The analysis of this estimator involves proving novel, time-uniform concentration results for bounded martingales in smooth Banach spaces, building upon foundational results due to Pinelis [128, 129]. The estimator not only enjoys favorable convergence, matching the rate attained by geometric median-of-means with small multiplicative constants [121], but also obtains rapid empirical convergence in simulations.

Part II: Applications of Martingale Concentration

In the second part of this thesis, we pivot away from the theoretical underpinnings of martingale concentration and instead focus on applying martingale methods to practically relevant data science problems. In this part, we focus on problems related to differentially private machine learning and online kernelized learning. Our the results covered below each involve (a) the identification of martingale structure in the underlying problem and (b) the application and optimization of an appropriately-chosen martingale concentration inequality.

In **Chapter 4**, we study the problem of fully-adaptive composition in differential privacy. Differential privacy provides an information-theoretic framework for protecting the integrity of an individual's data when used in computation. If an algorithm if differentially private, it is statistically hard for an attacker to use the output of an algorithm to determine if any given individual's data was used during computation. Given that private algorithms are often run in sequence, composition is perhaps the most important primitive in the study of differential privacy. Classical composition results assume that the privacy parameters (typically governed by quantities ϵ and δ) are fixed prior to computation. This is a poor fit for modern data science

tasks, in which a researcher may adaptively interact with a data set in order to answer relevant statistical questions. The few results that concern fully-adaptive composition, wherein privacy parameters may be adaptively selected by the statistician, indicate a significant price must be paid for adaptivity [133]. Using advances in line-crossing inequalities for martingales [74], we show that fully-adaptive composition can be obtained at no cost over traditional, fixed-parameter composition.

Next, in **Chapter 5**, we study the problem of differentially private empirical risk minimization under strict accuracy constraints. Traditional approaches to private risk minimization take a "privacy first" perspective. Namely, algorithms first require the practitioner to fix privacy levels in advance via parameters (ϵ, δ). The algorithms then inject noise (often Gaussian or Laplace) of appropriate variance to guarantee the target privacy levels are met. These approaches often provide pessimistic, high-probability utility guarantees. In safety-critical regimes (e.g. medical applications of private learning), practitioners may care primarily about accuracy, with privacy being an important secondary desideratum. In short, for these applications, a learner would like to know in advance the maximal variance of noise that should be added in order to guarantee the strict accuracy targets are met, thus minimum individual information leakage. Using tools from time-uniform martingale concentration alongside the theory of continuous-time stochastic processes, we develop a simple algorithm called the Brownian mechanism for privately meeting strict model accuracy requirements. Our algorithm involves first computing the true, risk minimizing parameter associated with a dataset. Then, it adds multivariate Gaussian noise of a large variance to this unknown parameter, iteratively stripping it away in a correlated manner until the target accuracy is met. We show that, by running this algorithm, the learner can obtain an ex-post privacy guarantee that matches the privacy loss had the optimal variance been known in advance up to multiplicative logarithmic factors in the variance.

Finally, in **Chapter 6**, we consider the kernelized bandit problem, in which an agent must sequentially learn the minimum of an unknown function while minimizing regret. This function is assumed to be of known complexity, lying in a ball in some reproducing kernel Hilbert space (RKHS). The simplest algorithm for the kernel bandit problem is Gaussian process upper confidence bound (GP-UCB) algorithm, which involves maintaining a kernel ridge estimate for the unknown function alongside a corresponding confidence ellipsoid. In the case of linear bandits, this algorithm is optimal in terms of minimax regret [102]. In kernelized setting, existing results indicate that GP-UCBC obtains super-linear regret for commonly-used kernels, such a the Matérn kernel family. Using improved self-normalized martingale concentration inequalities in separable Hilbert spaces alongside a simple regularization argument, we show GP-UCB obtains sub-linear regret for most commonly-used kernels, thus partially addressing an open question on the topic due to Vakili et al. [155].

Part III: Conclusions and Future Research Directions

In **Chapter 7**, we summarize the contributions, presents several interesting open problems, and then then discusses the research direction the author will be pursuing after graduating. In particular, the future topics the author will work on lie at the intersection of causal inference and machine learning, and mark a significant departure from the aforementioned completed work on martingale methods.

Part I

Foundations of Time-Uniform Martingale Concentration

Chapter 2

Time-Uniform Concentration for Self-Normalized Processes

Self-normalized processes arise naturally in many statistical tasks. While self-normalized concentration has been extensively studied for scalar-valued processes, there is less work on multidimensional processes outside of the sub-Gaussian setting. In this work, we construct a general, self-normalized inequality for $\mathbb{R}^d$-valued processes that satisfy a simple yet broad "sub-ψ" tail condition, which generalizes assumptions based on cumulant generating functions. From this general inequality, we derive an upper law of the iterated logarithm for sub-ψ vector-valued processes, which is tight up to small constants. We demonstrate applications in prototypical statistical tasks, such as parameter estimation in online linear regression and auto-regressive modeling, and bounded mean estimation via a new (multivariate) empirical Bernstein concentration inequality.

2.1 Introduction

The first Concentration inequalities are employed in many disparate mathematical fields. In particular, time-uniform martingale concentration has proven itself a critical tool in advancing research areas such as multi-armed bandits [88, 3, 102], differential privacy [163, 162], Bayesian learning [33], and online convex optimization [108, 83]. While martingale concentration inequalities have historically been proved in a largely case-by-case manner, recently Howard et al. [74] provided a unified framework for constructing time-uniform concentration inequalities. By introducing a single "sub-ψ" assumption that carefully controls the tail behavior of martingale increments, Howard et al. [74, 75] prove a master theorem that recovers (in fact improves) many classical examples of concentration inequalities, for example those of Blackwell [16], Hoeffding [70], Freedman [64], Azuma [10], de la Peña et al. [40].

Despite the generality of the framework of Howard et al. [74, 75], their results have not been extended to understanding the growth of "self-normalized" vector-valued processes. If $(S_n)_{n\geq 0}$ is a process evolving in $\mathbb{R}^d$ and $(V_n)_{n\geq 0}$ is a process of $d \times d$ positive semi-definite matrices measuring the "accumulated variance" of $(S_n)_{n\geq 0}$, self-normalized concentration aims to control the growth of the normalized process $(\|V_n^{-1/2} S_n\|)_{n\geq 0}$. Self-normalized processes naturally arise in

a variety of common statistical tasks, examples of which include regression problems [98, 99, 14] and contextual bandit problems [3, 30]. As such, any advances in self-normalized concentration for vector-valued processes could directly yield improvements in methodology and analysis of foundational statistical algorithms.

In this work, we provide a new, general approach for constructing self-normalized concentration inequalities. By naturally generalizing the sub-ψ condition of Howard et al. [74, 75] to d-dimensional spaces, we are able to construct a single "master" theorem that provides time-uniform, self-normalized concentration under a variety of noise settings. We prove our results by first constructing a time-uniform concentration inequality for scalar-valued processes that non-asymptotically matches law of the iterated logarithm and then extending this result to higher dimensions using a geometric argument. From our inequality, we can derive a multivariate analogue of the famed law of the iterated logarithm, which we show to be essentially tight. Lastly, we apply our inequality to common statistical tasks, such as calibrating confidence ellipsoids in online linear regression, estimating model parameters in vector auto-regressive models, and estimating a bounded mean via a new "empirical Bernstein" concentration inequality.

2.1.1 Related Work and History

Martingale concentration arguably originated in the work of Ville [157], who showed that the growth of non-negative supermartingales can be controlled uniformly over time. This result, now known commonly referred to as *Ville's inequality*, acts as a time-uniform generalization of Markov's inequality [50]. This result was later extended to submartingale concentration by Doob [47] in an eponymous result, now called *Doob's maximal inequality*. From these two inequalities, a variety of now classical martingale concentration inequalities were proved, such as Azuma's inequality [10], which serves as a time-uniform, martingale variant of Hoeffding's inequality [70] for bounded random variables, and Freedman's inequality [64], which serves as a martingale variant of Bennett's inequality [13] for sub-Poisson, bounded random variables.

Of particular note are the various self-normalized inequalities of de la Peña [42, 40, 41, 43], which provide time-uniform control of the growth of a process $(S_n)_{n\geq 0}$ in terms of an associated accumulated variance process $(V_n)_{n\geq 0}$. In particular, the authors derive their results using a technique first presented by Robbins called the method of mixtures [38, 37], which involves integrating over a family of parameterized exponential supermartingales to obtain significantly tighter (in terms of asymptotic behavior) inequalities than those mentioned earlier. Bercu and Touati [14] also investigate self-normalized concentration in the style of de la Peña, deriving bounds when the increments of $(S_n)_{n\geq 0}$ may exhibit asymmetric heavy-tailed behavior and, in later work, [15] study the effects of weighing predictable and empirical quadratic variation processes in deriving self-normalized concentration results.

Recently, Howard et al. [74] presented a single "master" theorem that ties together much of the literature surrounding scalar-valued concentration (self-normalized or not). Inspired by the classical Cramer-Chernoff method (see Boucheron et al. [18] for instance), which provides high probability tail bounds for a random variable X in terms of its cumulant generating function (or CGF) ψ, the authors present a unified "sub-ψ" condition on a stochastic process. This condition relates the growth of a process $(S_n)_{n\geq 0}$ to some corresponding accumulated variance process $(V_n)_{n\geq 0}$ through a function ψ which obeys many similar properties to a CGF. In particular, the

authors prove "line-crossing" inequalities for sub-ψ processes, giving a bound on the probability that $(S_n)_{n\geq 0}$ will ever cross a line parameterized by ψ and the accumulated variance $(V_n)_{n\geq 0}$. By strategically picking ψ and $(V_n)_{n\geq 0}$, the master theorem in Howard et al. [74] can be used to reconstruct, unify and even improve a variety of existing self-normalized concentration inequalities (such as those in the preceding paragraph), as well as to prove several new ones. Using these ideas in a followup work, Howard et al. [75] prove a time-uniform concentration inequality for scalar-valued processes whose rate non-asymptotically matches the law of the iterated logarithm (LIL) [50]. The only caveat to this result is that the concentration inequality only applies to sub-ψ processes when ψ is either the CGF of a sub-Gaussian (denoted ψ_N) or sub-Gamma (denoted $\psi_{G,c}$) random variable. While any CGF-like function ψ function can be bounded by $a\psi_{G,c}$ for *some* choice of $a, c > 0$ (see Proposition 1 of Howard et al. [75]), this conversion could in general result in loose constants. As a stepping stone toward proving our multivariate concentration inequalities, we generalize the non-asymptotic LIL results of Howard et al. [75] to arbitrary sub-ψ process, greatly increasing the applicability of the obtained results.

To the best of our knowledge, there are relatively few existing results on the self-normalized concentration of vector-valued processes. De la Peña [42] leverage the above-mentioned method of mixtures alongside Ville's inequality to bound the probability that the self-normalized random vector $V_n^{-1/2}S_n \in \mathbb{R}^d$ belongs to some mixture-dependent convex set. These bounds are, in particular, not closed form, and their asymptotic rate of growth is unclear. Our bounds, instead, directly provide time-uniform bounds on the process $(\|V_n^{-1/2}S_n\|)_{n\geq 0}$ in terms of relatively simple function of the variance process $(V_n)_{n\geq 0}$. In particular, we use our bounds to derive a multivariate law of the iterated logarithm that is tight in terms of dependence on V_n and the ambient dimension d up to small, absolute, known constants. While de la Peña et al. [41, 42] do provide an asymptotic LIL for vector processes, it hides an unknown constant and lacks explicit dependence on the dimension d.

In the case where the increments of $(S_n)_{n\geq 0}$ satisfy a sub-Gaussian condition, significantly more is known about vector-valued self-normalized concentration. Abbasi-Yadkori et al. [3] provide a clean bound on $\|V_n^{-1/2}S_n\|$ in terms of $\log\det(V_n)$ using an argument that directly follows from an earlier, method-of-mixtures based argument of de la Peña et al. [41]. First, our bounds are significantly more general than those of Abbasi-Yadkori et al. [3] and de la Peña et al. [41], because ours apply to *general* sub-ψ processes. Additionally, our bounds grow proportionally to $\log\log\gamma_{\max}(V_n)$ and $\log\kappa(V_n)$ ($\gamma_{\max}$ and κ represent maximum eigenvalue and condition number respectively, defined later). Thus, even in the setting of sub-Gaussian increments with predictable covariance, our results are not directly comparable in general. We believe deriving log-determinant rate inequalities for general sub-ψ processes is an interesting open problem, but leave it for future work.

There exist other concentration inequalities for vector-valued data that are not directly related to the self-normalized bounds presented in this paper. First, there are several existing time-uniform concentration results for Banach space-valued martingales [128, 129, 74]. These results are obtained by placing a smoothness assumption on the norm of the Banach space, and in turn provide time-uniform control on the norm of the martingale. We note that although we are working in a Banach space, we are not trying to control the norm of the underlying process $\|S_n\|$, and instead want to control the self-normalized quantity $\left\|V_n^{-1/2}S_n\right\|$. In particular, the

process $(V_n^{-1/2} S_n)_{n \geq 0}$ is not in general a martingale, so the above results cannot be directly applied. Second, there are many concentration results that involve bounding the operator norm of Hermitian matrix-valued martingales using the matrix Chernoff method [8, 31, 151, 152]. Once again, it does not seem like these bounds for matrix-valued processes can be readily applied to obtain vector-valued concentration of the form presented in this paper. Third, in their work on estimating convex divergences, Manole and Ramdas [116] derive a self-normalized concentration inequality for i.i.d. random vectors drawn from some distribution on $\mathbb{R}^d$. The form of this bound resembles that of the central concentration inequality presented in this paper. However, we note that our result allows for arbitrary martingale dependence between the increments of the process $(S_n)_{n \geq 0}$. Furthermore, the argument used in Manole and Ramdas [116] cannot be generalized to the setting of arbitrary dependence, as the authors derive their results using certain reverse martingale arguments which must be conducted with respect to the exchangeable filtration generated by a sequence of random variables, which implies the increments of $(S_n)_{n \geq 0}$ must, at the very least, be exchangeable random variables.

2.1.2 Our Contributions

We now provide a brief, illustrative summary of our primary contributions. For now, when we refer to a process (S_n) being sub-ψ with variance proxy (V_n), the reader should think of the increments of S_n having associated CGF ψ with weights proportional to V_n. This is not precise, but will be made exact when we provide rigorous definitions of the sub-ψ condition for both scalar and vector-valued processes in Section 2.2 below. We present the primary contributions in the order they appear in the paper.

1. First, in Section 2.3, we show that if $(S_n)_{n \geq 0}$ is a scalar (i.e. $\mathbb{R}$-valued) sub-ψ process with variance proxy $(V_n)_{n \geq 0}$, then, with high probability, it holds that

$$S_n = O\left(V_n \cdot (\psi^*)^{-1}\left(\frac{1}{V_n}\log\log(V_n)\right)\right)$$

for all $n \geq 0$ simultaneously. In the case where $\psi(\lambda) = \psi_{G,c}(\lambda) := \frac{\lambda^2}{2(1-c\lambda)}$ is the CGF associated with a sub-Gamma random variable (see Boucheron et al. [18]), our bound reduces to

$$S_n = O\left(\sqrt{V_n \log\log(V_n)} + c\log\log(V_n)\right).$$

Thus, this result can be reviewed as a direct generalization of the primary contributions of Howard et al. [75], who only provide time-uniform, self-normalized concentration results for sub-Gamma processes (Note that in the special case $c = 0$, sub-Gamma concentration reduces to sub-Gaussian concentration).

2. Next, in Section 2.4, we show that if $(S_n)_{n \geq 0}$ is a vector valued process that is sub-ψ with variance proxy $(V_n)_{n \geq 0}$, then, with high probability, simultaneously for all $n \geq 0$,

$$\|V_n^{-1/2} S_n\| = O\left(\sqrt{\gamma_{\min}(V_n)} \cdot (\psi^*)^{-1}\left(\frac{1}{\gamma_{\min}(V_n)}\left[\log\log(\gamma_{\max}(V_n)) + d\log\kappa(V_n)\right]\right)\right),$$

where $\gamma_{\min}(V_n)$ and $\gamma_{\max}(V_n)$ refer, respectively, to the minimum and maximum eigenvalues of the (proxy) covariance matrix V_n, and $\kappa(V_n) := \frac{\gamma_{\max}(V_n)}{\gamma_{\min}(V_n)}$ is the condition number of matrix V_n. We can compare our bounds to existing results [41, 3] in the subGaussian case where $\psi(\lambda) = \psi_N(\lambda) := \frac{\lambda^2}{2}$ is the CGF of a standard normal random variable. In this case, our bound simplifies to

$$\|V_n^{-1/2} S_n\| = O\left(\sqrt{\log\log(\gamma_{\max}(V_n)) + d\log\kappa(V_n)}\right).$$

Existing results on the self-normalized concentration for processes with sub-Gaussian increments provide upper bounds of the form $\|V_n^{-1/2} S_n\| = O\left(\sqrt{\log\det(V_n)}\right)$, which are, in general, incomparable to our bounds. When $\kappa(V_n)$ is small, our bounds may be tighter, but if $\gamma_{\max}(V_n) \gg \gamma_{\min}(V_n)$, the determinant-based bounds may be tighter.

3. Lastly, in Sections 2.5 and 2.6, we apply our vector-valued self-normalized concentration results to statistical tasks. In Section 2.5, we create non-asymptotically valid confidence ellipsoids for estimating unknown slope parameters in online linear regression with sub-ψ noise in observations. In particular, these results can be viewed as extending the confidence ellipsoids of Abbasi-Yadkori et al. [3], which hold only in the sub-Gaussian setting. We further specialize these bounds to model estimation in vector autoregressive models (i.e. in the $\mathrm{VAR}(p)$ model), generalizing a result of Bercu and Touati [14]. In Section 2.6, we prove a multivariate, self-normalized empirical Bernstein inequality, generalizing a result of Howard et al. [75] to d-dimensional space.

In sum, we provide time-uniform, self-normalized concentration inequalities for both scalar and vector-valued processes that hold under quite general noise conditions. Not only are these bounds of theoretical interest, but they are in fact applicable to common statistical tasks — in particular those that can be framed in the online linear regression framework.

2.2 Background and Sub-ψ Processes

In this section we discuss the key sub-ψ condition leveraged in deriving self-normalized concentration results for vector-valued processes. We arrive at our vector sub-ψ condition by extending the eponymous condition defined in the setting of scalar-valued processes [74, 75], to high dimensional spaces. We first summarize some notation that will be used ubiquitously.

Notation: Throughout, we define $\mathbb{N} = \{0, 1, 2, \cdots\}$ to be the set of natural numbers, which we assume to begin at 0. We let $\langle x, y\rangle = x^\top y$ denote that standard Euclidean inner product on $\mathbb{R}^d$. Additionally, we let $\mathbb{S}^{d-1} := \{x \in \mathbb{R}^d : \|x\| = 1\}$ denote the unit sphere and $\mathbb{B}_d := \{x \in \mathbb{R}^d : \|x\| \leq 1\}$ the unit ball in $\mathbb{R}^d$. By $\mathcal{L}_+(\mathbb{R}^d)$, we denote the set of all $d \times d$ positive semi-definite matrices, with $I_d \in \mathcal{L}_+(\mathbb{R}^d)$ denoting the d-dimensional identity matrix. For $V \in \mathcal{L}_+(\mathbb{R}^d)$, let $\gamma_{\max}(V)$ denote the largest eigenvalue of V, $\gamma_{\min}(V)$ the smallest eigenvalue of V, and let

$$\kappa(V) := \frac{\gamma_{\max}(V)}{\gamma_{\min}(V)}$$

denote the condition number of V. Each such V admits a spectral decomposition of the form $V = \sum_{i=1}^{d} \gamma_i(V) v_i v_i^\top$, where $(\gamma_i(V))_{i\in[d]}$ is the non-increasing sequence of eigenvalues associated with matrix V and $(v_i)_{i\in[d]}$ is the corresponding sequence of unit eigenvectors, which we know forms an orthonormal basis for $\mathbb{R}^d$. For $\rho > 0$, let

$$V \vee \rho I_d := \sum_{i=1}^{d} (\gamma_i(V) \vee \rho) v_i v_i^\top,$$

where for scalars $a, b \in \mathbb{R}$, $a \vee b := \max\{a, b\}$.

For a strictly increasing, differentiable convex function $\psi : [0, \lambda_{\max}) \to \mathbb{R}$ we let $\psi^* : [0, u_{max}) \to \mathbb{R}_{\geq 0}$ denote its convex conjugate, given by $\psi^*(u) := \sup_{\lambda\in[0,\lambda_{\max})} u\lambda - \psi(\lambda)$, where $u_{\max} := \lim_{\lambda\uparrow\lambda_{\max}} \psi'(\lambda)$. In the sequel, we will always assume $\sup_{\lambda\in[0,\lambda_{\max})} \psi'(\lambda) = \infty$, and hence will have $u_{\max} = \infty$. Some key properties of convex conjugation are that (a) ψ^* is convex, (b) $(\psi^*)^* = \psi$, and (c) $(\psi^*)' = (\psi')^{-1}$.

Let (Z, ρ) denote a metric space, and let $T \subset Z$. For $\epsilon > 0$, we say that a set $K \subset Z$ is an ϵ-covering for T if, for any $z \in T$, there exists a point $\pi(z) \in K$ satisfying $\rho(z, \pi(z)) \leq \epsilon$. We call $\pi : T \to K$ a "projection" onto the covering, which maps each point in T onto the nearest point in K (or an arbitrary one if not unique). If $K \subset T$, we call K a *proper* ϵ-covering of T. We will exclusively consider proper coverings in the sequel. We define the ϵ-covering number $N(T, \epsilon, \rho)$ of T to be the cardinality of the smallest proper ϵ-covering of T. Any proper ϵ-covering of T obtaining this minimum will be called minimal. In the special case $(Z, \rho) = (\mathbb{R}^d, \|\cdot\|)$ and $T = \mathbb{S}^{d-1}$, we denote the ϵ-covering number of T by $N_{d-1}(\epsilon)$.

Lastly, if $(S_n)_{n\geq 0}$ is some process evolving in a space $\mathcal{X}$ and $n \geq 1$, we define the nth increment of $(S_n)_{n\geq 0}$ to be $\Delta S_n := S_n - S_{n-1}$. If a filtration $(\mathcal{F}_n)_{n\geq 0}$ is understood from context, we may use the notation $\mathbb{E}_n[\cdot] = \mathbb{E}(\cdot \mid \mathcal{F}_n)$ for easing notational burden. By default, we take $\mathcal{F}_0 := \{\emptyset, \Omega\}$ and $\mathcal{F}_n = \sigma(S_1, \ldots, S_n)$.

Sub-ψ Processes: We now describe in more detail a condition that links the growth of a process $(S_n)_{n\geq 0}$ evolving in $\mathbb{R}^d$ to a corresponding "accumulated variance process" $(V_n)_{n\geq 0}$ taking values in $\mathcal{L}_+(\mathbb{R}^d)$. This linking will occur through the consideration of a family of exponential processes in which a scaled version of $(S_n)_{n\geq 0}$ along any fixed direction is compensated by $(V_n)_{n\geq 0}$ and a function ψ that measures the heaviness of the tails of ΔS_n. ψ should be thought of as acting like the cumulant generating function (or CGF) of ΔS_n — we will make this notion precise in our later discussion of CGF-like functions. These exponential processes will behave like non-negative supermartingales, and thus will allow us to apply powerful time-uniform concentration results to bound the growth of an appropriately normalized version of $(S_n)_{n\geq 0}$. Due to the central of role ψ in connecting the growth of $(S_n)_{n\geq 0}$ and $(V_n)_{n\geq 0}$, we will adopt the terminology of Howard et al. [74, 75] from the scalar case and refer to the condition as the "sub-ψ condition".

Before formally defining the sub-ψ condition, we must briefly discuss the properties of the heretofore vaguely defined function ψ. ψ will be a cumulant generating function-like (or CGF-like) function, which roughly means it behaves like the CGF of some random variable. More explicitly, by a CGF-like function, we mean a twice continuously-differentiable function $\psi : [0, \lambda_{\max}) \to \mathbb{R}_{\geq 0}$ satisfying (a) ψ is strictly convex, (b) $\psi(0) = \psi'(0) = 0$, and (c) $\psi''(0) > 0$.

Notable examples of CGF-like functions include $\psi_N(\lambda) := \frac{\lambda^2}{2}$, the CGF of a standard normal random variable;

$$\psi_{E,c}(\lambda) := \frac{-\log(1-c\lambda) - c\lambda}{c^2},$$

the CGF of a (centered) exponential random variable; $\psi_{P,c}(\lambda) := \frac{e^{c\lambda} - c\lambda - 1}{c^2}$, the CGF of a centered Poisson random variable; and

$$\psi_{G,c}(\lambda) := \frac{\lambda^2}{2(1-c\lambda)},$$

a bound on the CGF of a centered Gamma random variable. Note that, in particular, $\psi_{G,0} = \psi_N$. Basic theory regarding CGF-like functions is discussed in detail in Appendix 2.A. While we will use many nontrivial properties of CGF-like functions freely hereinafter, we will always make the proper forward reference to Appendix 2.A.

We now present the sub-ψ condition for scalar processes, and later for vector processes. First introduced in Howard et al. [74], the sub-ψ condition very heuristically states that, for each $n \geq 0$, the cumulant generating function for S_n is dominated by $V_n \cdot \psi$, where ψ is some CGF-like function per the above definition. More precisely, the sub-ψ condition for scalar-valued processes is as follows.

Definition 2.2.1. Let $\psi : [0, \lambda_{\max}) \to \mathbb{R}_{\geq 0}$ be CGF-like, let $(S_n)_{n\geq 0}$ and $(V_n)_{n\geq 0}$ be respectively $\mathbb{R}$-valued and $\mathbb{R}_{\geq 0}$-valued processes adapted to some filtration $(\mathcal{F}_n)_{n\geq 0}$. We say that $(S_n, V_n)_{n\geq 0}$ is sub-ψ (or equivalently that $(S_n)_{n\geq 0}$ is a sub-ψ process with variance proxy $(V_n)_{n\geq 0}$) if for every $\lambda \in [0, \lambda_{\max})$, the exponential process $\exp\{\lambda S_n - \psi(\lambda) V_n\}$ is (almost surely) upper bounded by some non-negative supermartingale $(L_n^\lambda)_{n\geq 0}$ with respect to $(\mathcal{F}_n)_{n\geq 0}$:

$$M_n^\lambda := \exp\{\lambda S_n - \psi(\lambda) V_n\} \leq L_n^\lambda, \qquad \text{for all } n \geq 0.$$

As an easy example, consider the case where $(X_n)_{n\geq 1}$ is a sequence of i.i.d. mean zero random variables with CGF $\psi(\lambda) = \log \mathbb{E} e^{\lambda X_1}$. Letting $S_n := \sum_{m=1}^n X_m$ and $V_n := n$, it is easy to see that M_n^λ is a non-negative martingale with respect to the natural filtration generated by the X_n's (and thus we can take $L_n^\lambda = M_n^\lambda$).

Definition 2.2.1 generalizes the above example to a setting where the random variables may have more complicated dependence structures, and "nonparametric" tail conditions, including settings where V_n can itself be adapted to $(\mathcal{F}_n)_{n\geq 0}$ (as opposed to constant or predictable variance processes), a key ingredient in self-normalized bounds. Recently, Howard et al. [74] compiled a rich selection of examples of such sub-ψ processes. For more examples, one can specialize each of the multivariate sub-ψ processes following Definition 2.2.2 below to the setting $d = 1$.

The above definition for scalar-valued processes suggests a straightforward means of generalizing the sub-ψ condition to the setting where $(S_n)_{n\geq 0}$ is $\mathbb{R}^d$-valued and $(V_n)_{n\geq 0}$ is $\mathcal{L}_+(\mathbb{R}^d)$-valued. Namely, $(S_n, V_n)_{n\geq 0}$ should be sub-ψ if, for any direction $\nu \in \mathbb{S}^{d-1}$, the scalar-valued process $(\langle \nu, S_n\rangle, \langle \nu, V_n \nu\rangle)_{n\geq 0}$ is sub-ψ. We formalize this in the following definition, which recovers Definition 2.2.1 in the case $d = 1$.

Definition 2.2.2. Let $\psi : [0, \lambda_{\max}) \to \mathbb{R}_{\geq 0}$ be CGF-like, and let $(S_n)_{n\geq 0}$ and $(V_n)_{n\geq 0}$ be respectively $\mathbb{R}^d$-valued and $\mathcal{L}_+(\mathbb{R}^d)$-valued processes adapted to some filtration $(\mathcal{F}_n)_{n\geq 0}$. We say that $(S_n, V_n)_{n\geq 0}$ is sub-ψ if, for every $\nu \in \mathbb{S}^{d-1}$, the projected process $(\langle \nu, S_n\rangle, \langle \nu, V_n \nu\rangle)_{n\geq 0}$ is sub-ψ

in the sense of Definition 2.2.1. In other words, $(S_n, V_n)_{n\geq 0}$ is sub-ψ if, for any $\nu \in \mathbb{S}^{d-1}$ and $\lambda \in [0, \lambda_{\max})$, there is a non-negative supermartingale $(L_n^{\lambda\cdot\nu})_{n\geq 0}$ with respect to $(\mathcal{F}_n)_{n\geq 0}$ such that

$$M_n^{\lambda\cdot\nu} := \exp\left\{\lambda\langle\nu, S_n\rangle - \psi(\lambda)\langle\nu, V_n\nu\rangle\right\} \leq L_n^{\lambda\cdot\nu}, \qquad \text{for all } n \geq 0.$$

It is straightforward to confirm that if $(S_n, V_n)_{n\geq 0}$ is sub-ψ, then $(S_n, V_n + \rho I_d)_{n\geq 0}$ and $(S_n, V_n \vee \rho I_d)_{n\geq 0}$ are sub-ψ as well. Furthermore, it is also straightforward to check that the rescaled process $(S_n/\sqrt{\rho}, V_n/\rho)_{n\geq 0}$ is sub-ψ_ρ, where $\psi_\rho : [0, \sqrt{\rho}\lambda_{\max}) \to \mathbb{R}_{\geq 0}$ is given by

$$\psi_\rho(\lambda) := \rho\psi(\lambda/\sqrt{\rho}).$$

These transformations are important as they will allow us to exclusively study processes satisfying $V_1 \geq 1$ in the sequel. For the sake of completeness, we prove that ψ_ρ is in fact CGF-like in Proposition 2.A.2 in Appendix 2.A. We codify the above observations into the following proposition for ease of reference.

Proposition 2.2.3. *Suppose $(S_n, V_n)_{n\geq 0}$ is sub-ψ with (inherently with respect to some filtration $(\mathcal{F}_n)_{n\geq 0}$). Then, for any fixed $\rho > 0$,*

1. *$(S_n, V_n + \rho I_d)_{n\geq 0}$ is sub-ψ with respect to $(\mathcal{F}_n)_{n\geq 0}$,*
2. *$(S_n, V_n \vee \rho I_d)_{n\geq 0}$ is sub-ψ with respect to $(\mathcal{F}_n)_{n\geq 0}$, and*
3. *$(S_n/\sqrt{\rho}, \rho^{-1}V_n)_{n\geq 0}$ is sub-ψ_ρ with respect to $(\mathcal{F}_n)_{n\geq 0}$, where $\psi_\rho(\lambda) := \rho\psi(\lambda/\sqrt{\rho})$.*

As we will see, Definition 2.2.2 will prove to be the "right" generalization of the sub-ψ condition to high-dimensional settings. In more detail, from the condition, we will derive a general, time-uniform bound on the self-normalized process $(\|V_n^{-1/2}S_n\|)_{n\geq 0}$ that will be tight up to small, multiplicative constants.

Four Examples of Sub-ψ Processes: We now provide four practically-relevant examples of multivariate sub-ψ processes — one for each of the aforementioned CGF-like functions $\psi_N, \psi_P, \psi_{E,c}$, and $\psi_{G,c}$. In each of the examples below, we assume we are studying some process $(X_n)_{n\geq 1}$ that is adapted to some filtration $(\mathcal{F}_n)_{n\geq 0}$.

1. If $X_n =_d -X_n \mid \mathcal{F}_{n-1}$ (that is, the X_n are conditionally symmetric), Lemma 3 of de la Peña et al. [41] can be used to show that $S_n := \sum_{m=1}^n X_m$ is sub-ψ_N with variance proxy $V_n := \sum_{m=1}^n X_m X_m^\top$. This provides salient example of how the sub-ψ condition can be leveraged to provide meaningful concentration for processes whose increments may even lack a well-defined mean (e.g. take the X_n to be i.i.d. Cauchy random variables).

2. If $\|X_n\| \leq c$ almost surely, a standard Bennett-style argument (see the proof of Theorem 2.9 in Boucheron et al. [18]) shows that $S_n := \sum_{m=1}^n \{X_m - \mathbb{E}_{m-1}X_m\}$ is sub-$\psi_{P,c}$ with variance proxy $V_n := \sum_{m=1}^n \mathbb{E}_{m-1}X_m X_m^\top$.

3. As will be seen in Section 2.6, if $\|X_n\| \leq 1/2$ almost surely[1], then $S_n := \sum_{m=1}^n \{X_m - \mathbb{E}_{m-1}X_m\}$ is sub-$\psi_{E,1}$ with variance proxy $V_n := \sum_{m=1}^n (X_m - \widehat{\mu}_{m-1})(X_m - \widehat{\mu}_{m-1})^\top$. In the above,

[1]Note that the the assumption $\|X_n\| \leq \frac{1}{2}$ can be replaced by any constant by appropriately changing the scale parameter of the sub-Exponential CGF.

$\widehat{\mu}_n := \frac{1}{n}\sum_{m=1}^n X_m$ is the time-average mean given the first n samples. From this condition, one can derive a multivariate, self-normalized "empirical Bernstein" inequality. This type of inequality is useful in statistical applications [161] due to the fact its tightness adapts to the observed (i.e. empirical) variance within the samples witnessed.

4. Lastly, if $\mathbb{E}_{n-1}|\langle \nu, X_n\rangle|^k \leq \frac{k!}{2}c^{k-2}\mathbb{E}_{n-1}\langle \nu, X_n\rangle^2$ for all directions $\nu \in \mathbb{S}^{d-1}$ and some constant $c > 0$, a standard application of the Bernstein condition in each direction $\nu \in \mathbb{S}^{d-1}$ (see Theorem 2.10 of Boucheron et al. [18]) yields that $S_n := \sum_{m=1}^n \{X_m - \mathbb{E}_{m-1}X_m\}$ is sub-$\psi_{G,c}$ with variance proxy $V_n := \sum_{m=1}^n \mathbb{E}_{m-1}X_m X_m^\top$.

Super-Gaussian CGFs: We draw attention to *super-Gaussian* ψ:

a CGF-like function ψ is super-Gaussian if $\frac{\psi(\lambda)}{\lambda^2}$ is an increasing function of λ.

In words, ψ is super-Gaussian if it grows at least as rapidly as ψ_N, the CGF of a $\mathcal{N}(0,1)$ random variable. Most notable examples of CGF-like functions are super-Gaussian, with particularly important examples being $\psi_N, \psi_{E,c}, \psi_{G,c}$, and $\psi_{P,c}$. Informally, one typically needs to use a super-Gaussian CGF if the underlying random process is heavier tailed than a sub-Gaussian process.

One example of a CGF that is not super-Gaussian would be $\psi_{B,p}(\lambda)$, the CGF of a centered Bernoulli random variable X with $\mathbb{P}(X = 1) = p$. We discuss equivalent definitions and properties of CGF-like functions in detail in Appendix 2.A. While our bounds will hold in the case where $(S_n, V_n)_{n\geq 0}$ is sub-ψ for arbitrary ψ, they are particularly clean when ψ is super-Gaussian, and we emphasize this case going forward.

2.3 A General Non-Asymptotic LIL for Scalar Processes

In this section, we prove a high-probability, time-uniform bound on the growth of a scalar process $(S_n)_{n\geq 0}$ normalized by some measure of accumulated variance $(V_n)_{n\geq 0}$. In particular, in Theorem 2.3.1 below, we show that if $(S_n, V_n)_{n\geq 0}$ is a sub-ψ process, then, with high probability, simultaneously for all $n \geq 0$,

$$S_n = O\left(V_n \cdot (\psi^*)^{-1}\left(\frac{1}{V_n}\log\log(V_n)\right)\right),$$

where we have omitted dependence on several user-chosen parameters and constants for the sake of exposition. Dividing both sides by $\sqrt{V_n}$ yields a result in "self-normalized" form that looks more akin to the results in subsequent sections, but we adopt the above form for consistency with existing results [74, 75]. Since $(\psi^*)^{-1}(u) \sim \sqrt{2u}$ as $u \downarrow 0$ whenever $\psi(\lambda) \sim \frac{\lambda^2}{2}$ as $\lambda \downarrow 0$ (as is the case for all CGF-like functions addressed in the previous section), for large values of V_n, the above high probability bound can be written as

$$S_n = O(\sqrt{V_n \log\log(V_n)}),$$

thus allowing our results in this section to be viewed as a non-asymptotic (i.e. finite sample) version of the law of the iterated logarithm. We further describe connections between our scalar-valued bound and the law of the iterated logarithm in Subsection 2.3.2 below.

While we construct the bounds in this section as a requisite for deriving self-normalized concentration inequalities for vector-valued processes, we believe the results are of independent interest. In particular, our results are significantly more general than those of Howard et al. [75], whose bounds serve as the current state-of-the-art for scalar-valued self-normalized concentration. Unlike the results of Howard et al. [75], which only hold for sub-$\psi_{G,c}$ (i.e. sub-Gamma) processes, our results hold for *general* sub-ψ processes. While Howard et al. [74] show that any CGF-like function ψ can be bounded pointwise by $a\psi_{G,c}$ for appropriately chosen constants $a, c > 0$, this comparison can be loose. We illustrate this in Figure 2.1 in Appendix 2.D, which shows that the time-uniform boundary presented in Theorem 2.3.1 (applied in the sub-Poisson setting $\psi = \psi_{P,c}$) can offer improved concentration over the main theorem of Howard et al. [75], which requires converting sub-$\psi_{P,c}$ process to sub-$\psi_{G,c}$ processes in order to be applied. This case is of particular interest, as straightforward calculation shows $\psi_{P,c}(\lambda) \leq \psi_{G,c}(\lambda)$ for all λ. Other examples demonstrating this disparity in bounds can be readily constructed as well. Furthermore, even in the case of sub-$\psi_{G,c}$ processes, our bounds are essentially flush with those of Howard et al. [75] in the sub-Gamma case, being multiplicatively looser by a vanishingly small factor as certain tuning parameters are appropriately selected, as illustrated in Figure 2.2, also in Appendix 2.D. We further discuss comparisons between our bounds and those of Howard et al. [75] following the proof of Theorem 2.3.1.

Before presenting the main theorem of this section, we discuss heuristically how we are able to generalize the results of Howard et al. [75]. Much like the "stitching" technique of the aforementioned authors, our argument proceeds, roughly, by breaking "intrinsic" time into geometric epochs of the form $\{\alpha^k \leq V_n < \alpha^{k+1}\}$ and then optimizing a tight linear inequality in each period. The key difference in our argument is in how we optimize this linear boundary for $(S_n)_{n\geq 0}$ in each epoch. The techniques leveraged by Howard et al. [75] yield a boundary that is defined in terms of $\psi_{G,c}^{-1}$, the inverse of the CGF-like function associated with a Gamma distribution. From our understanding of the classical Chernoff argument, we know that if a mean zero random variable X has associated CGF $\psi(\lambda) := \log \mathbb{E}e^{\lambda X}$, then we have $\mathbb{P}\left(X \geq (\psi^*)^{-1}\left(\log\left(\frac{1}{\delta}\right)\right)\right) \leq \delta$. Thus, although the Chernoff argument doesn't directly apply in this time-uniform setting, we at the very least expect to obtain a boundary defined in terms of $(\psi^*)^{-1}$. By coupling this intuition with the time-uniform line crossing inequalities of Howard et al. [74], we are able to obtain an extremely general inequality for sub-ψ processes with a surprisingly straightforward argument.

With the above discussion in mind, we present Theorem 2.3.1.

Theorem 2.3.1. *Suppose $(S_n, V_n)_{n\geq 0}$ is a real-valued sub-ψ process for some CGF-like function $\psi : [0, \lambda_{\max}) \to \mathbb{R}_{\geq 0}$ satisfying $\lim_{\lambda\uparrow\lambda_{\max}} \psi'(\lambda) = \infty$. Let $\alpha > 1, \rho > 0$, and $\delta \in (0, 1)$ be constants respectively representing the stitching epoch length, the minimum intrinsic time, and the error probability. Let $h : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be an increasing function such that $\sum_{k\in\mathbb{N}} h(k)^{-1} \leq 1$, representing how the error is spent across epochs. Define the function $\ell_\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ by*

$$\ell_\rho(v) = \log\left(h\left(\log_\alpha\left(\frac{v \vee \rho}{\rho}\right)\right)\right) + \log\left(\frac{1}{\delta}\right),$$

where we have suppressed the dependence of $\ell_\rho(v)$ on α, h for brevity. Then, we have

$$\mathbb{P}\left(\exists t \geq 0 : S_n \geq (V_n \vee \rho) \cdot (\psi^*)^{-1}\left(\frac{\alpha}{V_n \vee \rho}\ell_\rho(V_n)\right)\right) \leq \delta.$$

We provide a full proof of Theorem 2.3.1 in Section 2.7 below. Except for the unavoidable error probability δ, we briefly elaborate on the other user-specified constants that appear in the statement of the theorem:

1. $\alpha > 1$ controls the spacing of "intrinsic time" or accumulated variance of the process $(S_n)_{n\geq 0}$. Heuristically, Theorem 2.3.1 will be obtained by optimizing tight, linear boundaries on events of the form $\{\alpha^k \leq V_n < \alpha^{k+1}\}$.

2. $\rho > 0$ gives the first "intrinsic time" at which our boundaries start depending on the variance process $(V_n)_{n\geq 0}$. When $0 \leq V_n < \rho$, the boundary will only depend on ρ.

3. $h : \mathbb{R} \to \mathbb{R}_{>0}$ is a function satisfying $\sum_{k\geq 0} h(k)^{-1} \leq 1$. h defines how much of the overall probability mass associated with failure (determined by δ) to allocate to each event of the form $\{\alpha^k \leq V_n < \alpha^{k+1}\}$.

In the above, we view the parameters ρ, and h as critical, since they directly affect the shape and validity of the bound, whereas we view α as less critical, as any small variation in α will only minimally affect the tightness of the bound in terms of constants. For example, a reasonable choice of this temporal spacing parameter is $\alpha = 1.05$. Howard et al. [75] discuss reasonable choices for the function h, and we emphasize in the sequel the choice of $h(k) := (k+1)^s\zeta(s)$, where $s > 1$ is a tuning parameter and ζ is the Riemann zeta function. This choice is of particular theoretical interest as it yields non-asymptotic rates that depend on $\log\log(V_n)$ (up to constants), thus allowing our bound to be viewed as a general, non-asymptotic version of the LIL. We in particular use this choice of h in the proof of Corollary 2.3.2 in Subsection 2.3.2 below.

2.3.1 Comparison With Existing Bounds

We compare our results to those presented in Theorem 1 of Howard et al. [75], who provide time-uniform, self-normalized concentration for scalar processes in the sub-$\psi_{G,c}$ case, where we recall $\psi_{G,c}(\lambda) := \frac{\lambda^2}{2(1-c\lambda)}$ is the CGF-like function associated with a sub-Gamma random variable. We start by analyzing the special case $c = 0$, in which $\psi_{G,0} = \psi_N$ is the CGF of a $\mathcal{N}(0,1)$ random variable. In our notation, the authors show that if $(S_n, V_n)_{n\geq 0}$ is a sub-ψ_N process, then, with probability at least $1-\delta$, simultaneously for all $t \geq 0$,

$$S_n \leq \sqrt{\left(\frac{\alpha^{1/4} + \alpha^{-1/4}}{\sqrt{2}}\right)^2 (V_n \vee \rho)\ell_\rho(V_n)}.$$

Noting that $(\psi_N^*)^{-1}(u) = \sqrt{2u}$, our results yield that, with probability at least $1-\delta$, simultaneously for all $n \geq 0$,

$$S_n \leq \sqrt{2\alpha(V_n \vee \rho)\ell_\rho(V_n)}.$$

A straightforward computation yields that, for all $\alpha > 1$, $\left(\frac{\alpha^{1/4}+\alpha^{-1/4}}{\sqrt{2}}\right)^2 \leq 2\alpha$, showing that the bounds of Howard et al. [75] are (slightly) tighter than our own. However, for $\alpha < 2.06$, we have $2\alpha \leq 2\left(\frac{\alpha^{1/4}+\alpha^{-1/4}}{\sqrt{2}}\right)^2$, showing that our bounds are looser than those of Howard et al. [75] by a multiplicative factor of no more than $\sqrt{2}$ in this regime. In particular, as α is decreased towards

1, the multiplicative factor by which our bounds are suboptimal to those of Howard et al. [75] vanishes to 1.

In the general case of $c > 0$, Theorem 1 in Howard et al. [75] yields that with probability at least $1 - \delta$, we have

$$S_n \leq \sqrt{\left(\frac{\alpha^{1/4} + \alpha^{-1/4}}{\sqrt{2}}\right)^2 (V_n \vee \rho)\ell_\rho(V_n) + \left(\frac{\sqrt{\alpha}+1}{2}\right)^2 \ell_\rho(V_n)} + c\left(\frac{\sqrt{\alpha}+1}{2}\right)\ell_\rho(V_n). \tag{2.3.1}$$

Meanwhile, noting that $(\psi^*_{G,c})^{-1}(x) = \sqrt{2x} + cx$ (this can be readily checked, or see Boucheron et al. [18] e.g.), our results yield that, with probability at least $1 - \delta$, we have

$$S_n \leq \sqrt{2\alpha(V_n \vee \rho)\ell_\rho(V_n)} + c\alpha\ell_\rho(V_n).$$

In this situation, a general comparison between the resulting bounds isn't clear. Our second term is larger, but it is a lower order term. Regarding the first term, for small V_n and α, we may expect our bound to be tighter, as we don't suffer from the second additive term inside of the square root. On the other hand, for any $\alpha > 1$ and moderate to large V_n, by our analysis in the sub-Gaussian case of $\psi = \psi_N$ (which is recovered when $c = 0$), we expect the bound from Howard et al. [75] to be tighter for the same reason. We plot a detailed comparison between Theorem 2.3.1 applied in the sub-$\psi_{G,c}$ case and Equation 2.3.1 in Figure 2.2, found in Appendix 2.D.

We emphasize that the bounds of Howard et al. [75] hold *only* in the sub-Gamma case. While sub-Gamma concentration can be applied to sums of sub-Exponential and sub-Poisson random variables, this approximation is far from tight, especially in small sample sizes. Our results hold directly for *any* CGF-like function ψ, including all listed in Section 2.2.

2.3.2 Asymptotic Law of the Iterated Logarithm

In the preceding paragraphs, we derived time-uniform bounds for general scalar-valued sub-ψ processes. In particular, we argued our presented results generalized those of Howard et al. [75], who show a similar result for the case $\psi = \psi_{G,c} = \frac{\lambda^2}{2(1-c\lambda)}$ (i.e. when ψ is the CGF-like function associated with a sub-Gamma random variable). As noted above, for any fixed step size $\alpha > 1$, in the case $c = 0$ (i.e. when $\psi = \psi_N$ is the CGF of a standard Gaussian random variable), the bounds of Howard et al. [75] dominate ours, albeit by a vanishingly small multiplicative factor as $\alpha \downarrow 1$.

This begs the following question: are our bounds "optimal" in the sense that, by appropriately selecting the tuning parameters, they recover the asymptotic (upper) law of the iterated logarithm with the correct constant. In Corollary 2.3.2 below, we show that this exactly the case, and thus derive a law of the iterated logarithm for sub-ψ processes.

Corollary 2.3.2. *Let $(S_n)_{n\geq 0}$ be sub-ψ with variance proxy $(V_n)_{n\geq 0}$, and suppose that $\psi(\lambda) \sim \frac{\lambda^2}{2}$ as $\lambda \downarrow 0$ and $V_n \xrightarrow[n\to\infty]{} \infty$. Then,*

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2V_n \log\log(V_n)}} \leq 1 \quad \textit{almost surely}.$$

Corollary 2.3.2 follows as a direct consequence of Corollary 2.4.6, which provides an asymptotic law of the iterated logarithm for vector-valued sub-ψ processes, noting that the dependence of the bound on the condition number of V_n vanishes in the scalar case.

We provide some brief intuition for our proof of Corollary 2.3.2. In the proof, we consider a sequence of bounds (indexed by $n \geq 1$), with tuning parameters $(\alpha_t)_{t\geq 1}, (\delta_t)_{t\geq 1}$, and $(h_t)_{t\geq 1}$ satisfying $\alpha_t \downarrow 1$, $\delta_t \downarrow 0$, and $h_t(k) := (k+1)^{\iota_t}\zeta(\iota_t)$, where $\iota_t \downarrow 1$. We assume $\rho := 1$, as by assumption the variance $(V_n)_{n\geq 0}$ will grow towards infinity and thus the initial time at which the bound is valid will not matter. Heuristically, smaller values of the aforementioned parameters (being $\alpha_t, \delta_t, \iota_t$) imply that the ratio between our bounds and $\sqrt{2V_n \log\log(V_n)}$ will be closer to 1 for large values of V_n, but will suffer from an increased "bias" or additive penalty for small values of V_n. Since we assume V_n grows towards infinity almost surely, the effect of the additive bias becomes negligible in the large intrinsic time limit (i.e. as $V_n \to \infty$).

In more detail, for any $t \geq 1$, we show that for large values of time $n \geq N_t$ (N_t may be random), we have $V_n \cdot (\psi^*)^{-1}\left(\frac{\alpha}{V_n}\ell_1(V_n)\right) \leq (1+\eta)C_t\sqrt{2V_n \log\log(V_n)}$, where $C_t \downarrow 1$ as $n \to \infty$ and $\eta > 0$ is some pre-fixed parameter. Given we have such a bound for all $t \geq 1$, we can apply the first Borel-Cantelli lemma (which describes when certain sequences of events will happen either "infinitely often" or only "finitely often") to show that $\limsup_{n\to\infty} \frac{S_n}{\sqrt{2V_n \log\log(V_n)}} \leq (1+\eta)$. Since $\eta > 0$ was arbitrary, the desired bound follows.

2.4 Main result

We now present the main result of this paper: a time-uniform, self-normalized concentration inequality for a general class of processes evolving in $\mathbb{R}^d$. We now discuss the intuition for our argument. Our results follow by coupling our scalar-valued self-normalized inequalities, presented in the previous section, with a simple but careful geometric covering argument. At a high level, our results in the previous section could be seen as controlling the growth of the process $(S_n)_{n\geq 0}$ over various scales of "intrinsic time", determined by the accumulated variance process $(V_n)_{n\geq 0}$. Analogously, to handle the multivariate nature of results in this section, we need to carefully control how the accumulated variance process (this time a matrix-valued process), distorts the geometry of $\mathbb{R}^d$ across various scales. In this setting, the level of distortion is controlled by $\kappa(V_n) := \gamma_{\max}(V_n)/\gamma_{\min}(V_n)$, the condition number of the positive semi-definite matrix V_n (if $\gamma_{\min}(V_n) = 0$, $\kappa(V_n) = \infty$ by convention).

Theorem 2.4.1. *Suppose $(S_n)_{n\geq 0}$ is a sub-ψ process with variance proxy $(V_n)_{n\geq 0}$ taking values $\mathbb{R}^d$. Let $\alpha > 1$, $\beta > 1$, $\rho > 0$, $\epsilon \in (0,1)$, and $\delta \in (0,1)$ be constants, and let $h : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be an increasing function such that $\sum_{k\in\mathbb{N}} h(k)^{-1} \leq 1$. Define the function[2] $L_\rho : \mathcal{S}_+^d \to \mathbb{R}_{\geq 0}$ by*

$$L_\rho(V) := \log\left(h\left(\log_\alpha\left(\frac{\gamma_{\max}(V \vee \rho I_d)}{\rho}\right)\right)\right) + \log\left(\frac{1}{\delta}\frac{1}{1-\beta^{-1}}\right)$$
$$+ \log\left(\beta\sqrt{\kappa(V \vee \rho I_d)} \cdot N_{d-1}\left(\frac{\epsilon}{\beta\sqrt{\kappa(V_n \vee \rho I_d)}}\right)\right).$$

[2]Recall $N_{d-1}(\epsilon)$ was defined to be the ϵ-covering number of $\mathbb{S}^{d-1}$.

If ψ is super-Gaussian, meaning $\psi(\lambda)/(\lambda^2/2)$ is an increasing function of λ, then

$$\mathbb{P}\left(\exists n \geq 0 : \left\|(V_n \vee \rho I_d)^{-1/2} S_n\right\| \geq \frac{\sqrt{\gamma_{\min}(V_n \vee \rho I_d)}}{1-\epsilon} \cdot (\psi^*)^{-1}\left(\frac{\alpha}{\gamma_{\min}(V_n \vee \rho I_d)} L_\rho(V_n)\right)\right) \leq \delta.$$

In addition to the parameters α, ρ, and h from Theorem 2.3.1, there are two new user-specified constants that govern the geometric aspects of our bound presented in Theorem 2.4.1.

1. $\beta > 1$ controls the spacing of how the action of the sequence of matrices $(V_n)_{n\geq 0}$ distorts the geometry of $\mathbb{R}^d$. Heuristically, Theorem 2.4.1 will be obtained by optimizing self-normalized inequalities on events of the form $\{\beta^k \leq \sqrt{\kappa(V_n)} < \beta^{k+1}\}$ and carefully performing a union bound.

2. $\epsilon \in (0,1)$ controls the "mesh" or level of granularity at which we approximate the geometry of the unit sphere $\mathbb{S}^{d-1}$ in the covering argument we make.

In the vocabulary of our preceding results, we view neither β nor ϵ as being critical parameters in optimizing our boundary. In particular, for simplicity, reasonable default choices would be $\beta = 2$ and $\epsilon = \frac{1}{2}$.

Before proving Theorem 2.4.1, we comment that a result similar to the above holds even in the setting where the CGF-like function ψ is not super-Gaussian. In particular, en route to proving the above, we will show that, if $(S_n, V_n)_{n\geq 0}$ is sub-ψ for any CGF-like ψ, we have

$$\mathbb{P}\left(\exists n \geq 0 : \left\|(V_n \vee \rho I_d)^{-1/2} S_n\right\| \geq \sup_{\nu \in \mathbb{S}^{d-1}} \frac{\sqrt{\langle \nu, V_n \nu\rangle}}{1-\epsilon} \cdot (\psi^*)^{-1}\left(\frac{\alpha}{\langle \nu, V_n \nu\rangle} L_\rho(V_n)\right)\right) \leq \delta.$$

The assumption that ψ is super-Gaussian merely allows us to compute the supremum over $\nu \in \mathbb{S}^{d-1}$ in the above expression, giving the result a cleaner form. This assumption is not restrictive, as many reasonable examples of CGF-like functions are super-Gaussian (e.g. ψ_N, $\psi_{G,c}$, $\psi_{E,c}$, and $\psi_{P,c}$ to name a few encountered earlier).

To simplify the above bound further (at the cost of introducing some looseness) we can plug in upper bounds on $N_{d-1}(\epsilon)$ into Theorem 2.4.1. We prove the following lemma in Appendix 2.C, and it follows from a simple geometric argument. The following bound is not tight, but suffices for subsequent asymptotic analysis. We use the bound presented in Corollary 2.4.2 over the bound $N_{d-1}(\epsilon) \leq \left(\frac{3}{\epsilon}\right)^d$ (which follows from Lemma 5.7 of Wainwright [158]) due to slightly improved dependence on d in the exponent. If desired, one could obtain an analogue of Corollary 2.4.3 using the aforementioned bound as well, or even any bound on $N_{d-1}(\epsilon)$.

Lemma 2.4.2. *Let $\epsilon \in (0,1)$ be arbitrary and $d \geq 1$. Then,*

$$N_{d-1}(\epsilon) \leq C_d \left(\frac{3}{\epsilon}\right)^{d-1},$$

where C_d is a constant that does not depend on ϵ.

With the above bound on the covering number of $\mathbb{S}^{d-1}$, we have the following corollary.

Corollary 2.4.3. *Assume the same setup as in Theorem 2.4.1. Define $L_\rho^{cov} : \mathcal{L}_+(\mathbb{R}^d) \to \mathbb{R}_{\geq 0}$ by*

$$L_\rho^{cov}(V) := \log\left(h\left(\log_\alpha\left(\frac{\gamma_{\max}(V \vee \rho I_d)}{\rho}\right)\right)\right) + \log\left(\frac{C_d}{\delta(1-\beta^{-1})}\right) + d\log\left(\frac{3\beta\sqrt{\kappa(V \vee \rho I_d)}}{\epsilon}\right).$$

Then,

$$\mathbb{P}\left(\exists n \geq 0 : \left\|(V_n \vee \rho I_d)^{-1/2} S_n\right\| \geq \frac{\sqrt{\gamma_{\min}(V_n \vee \rho I_d)}}{1-\epsilon} \cdot (\psi^*)^{-1}\left(\frac{\alpha}{\gamma_{\min}(V_n \vee \rho I_d)} L_\rho^{cov}(V_n)\right)\right) \leq \delta.$$

Proof. The result immediately follows by applying Theorem 2.4.1 and noting the bound

$$\begin{aligned}\beta\sqrt{\kappa(V \vee \rho I_d)} \cdot N_{d-1}\left(\frac{\epsilon}{\beta\sqrt{\kappa(V \vee \rho I_d)}}\right) &\leq \frac{3\beta\sqrt{\kappa(V_n \vee \rho I_d)}}{\epsilon} \cdot \left(\frac{3\beta\sqrt{\kappa(V \vee \rho I_d)}}{\epsilon}\right)^{d-1} \\ &= \left(\frac{3\beta\sqrt{\kappa(V \vee \rho I_d)}}{\epsilon}\right)^d,\end{aligned}$$

which holds for all positive semi-definite matrices $V \in \mathcal{L}_+(\mathbb{R}^d)$. ■

Treating the tuning parameters $\alpha, \beta, \epsilon, \rho$ as constants and selecting $h : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ satisfying $h(k) = O(\log(k))$ (which, as noted by Howard et al. [75], holds when $h(k) := (k+1)^s \zeta(s)$ for any $s > 1$), Corollary 2.4.3 yields that, with high probability, simultaneously for all $n \geq 0$,

$$\left\|V_n^{-1/2} S_n\right\| = O\left(\sqrt{\gamma_{\min}(V_n)} \cdot (\psi^*)^{-1}\left(\frac{1}{\gamma_{\min}(V_n)}\left[\log\log(\gamma_{\max}(V_n)) + d\log\kappa(V_n)\right]\right)\right).$$

We now specify (in terms of big-Oh notation) our bounds to the setting of two common CGF-like functions. First, we consider the case $\psi(\lambda) = \psi_{G,c}(\lambda) = \frac{\lambda^2}{2(1-c\lambda)}$, for which we recall that $\left(\psi_{G,c}^*\right)^{-1}(x) = \sqrt{2x} + cx$, and so our bounds yield that

$$\left\|V_n^{-1/2} S_n\right\| = O\left(\sqrt{\log\log(\gamma_{\max}(V_n)) + d\log\kappa(V_n)} + \frac{c}{\sqrt{\gamma_{\min}(V_n)}}\left[\log\log(\gamma_{\max}(V_n)) + d\log\kappa(V_n)\right]\right)$$

Further specifying to the case $\psi(\lambda) = \psi_N(\lambda) = \frac{\lambda^2}{2}$ (which is equivalent to the case $\psi = \psi_{G,c}$ with $c = 0$), the above bound reduces to the form:

$$\left\|V_n^{-1/2} S_n\right\| = O\left(\sqrt{\log\log(\gamma_{\max}(V_n)) + d\log\kappa(V_n)}\right). \tag{2.4.1}$$

The bound (2.4.1), in particular, captures the asymptotic growth rate of very general classes of sub-ψ process when $\psi(\lambda) \sim \frac{\lambda^2}{2}$ as $\lambda \downarrow 0$ (in a sense that we will make fully precise soon).

2.4.1 Comparison With Existing Bounds

Now that we have presented the main result of the paper in Theorem 2.4.1 and examined the dimensional dependence of the bound in Corollary 2.4.3, we can compare our bounds to existing results in the literature. As a warm up (and sanity check), we ensure our bound presented in Theorem 2.4.1 recovers the results presented in Theorem 2.3.1, up to multiplicative or additive constants.

Comparison with Scalar Bounds: If $(S_n, V_n)_{n\geq 0}$ is sub-ψ and taking values in $\mathbb{R}$, we note that for any $\epsilon \in (0,1), N_0(\epsilon) = 2$. Thus, Theorem 2.4.1 yields that, with probability at least $1-\delta$, simultaneously for all $n \geq 0$ (assuming $V_n \geq \rho$ for simplicity),

$$\begin{aligned}\frac{S_n}{\sqrt{V_n}} &\leq \sqrt{V_n} \cdot (\psi^*)^{-1}\left(\frac{\alpha}{V_n} L_\rho(V_n)\right) \\ &= \frac{1}{1-\epsilon}\sqrt{V_n} \cdot (\psi^*)^{-1}\left(\frac{\alpha}{V_n}\left[\ell_\rho(V_n) + \log\left(\frac{1}{1-\beta^{-1}}\right) + \log(2\beta)\right]\right) \\ &= C_1\sqrt{V_n} \cdot (\psi^*)^{-1}\left(\frac{\alpha}{V_n}\left[\ell_\rho(V_n) + C_2\right]\right),\end{aligned}$$

where we have defined $C_1 := \frac{1}{1-\epsilon}$ and $C_2 := \log\left(\frac{1}{1-\beta^{-1}}\right) + \log(2\beta)$ for convenience. On the other hand, Theorem 2.3.1 yields that, with probability at least $1-\delta$, simultaneously for all $n \geq 0$ such that $V_n \geq \rho$,

$$\frac{S_n}{\sqrt{V_n}} \leq \sqrt{V_n} \cdot (\psi^*)^{-1}\left(\frac{\alpha}{V_n}\ell_\rho(V_n)\right).$$

As expected, the bound yielded by Theorem 2.4.1 is looser than the that of Theorem 2.3.1, due to the extra union bound needed in the covering argument. However, the looseness is only by an absolute multiplicative factor C_1 and an additive factor of C_2 inside $(\psi^*)^{-1}$. Since the $N_0(\epsilon) = 2$ for all $\epsilon \in (0,1)$, the multiplicative factor can be forced to be arbitrarily small by appropriately choosing the covering parameter ϵ (e.g. sending $\epsilon \downarrow 0$ yields $C_1 \downarrow 1$). Likewise, the impact of the additive factor becomes vanishingly small as $V_n \to \infty$.

Method of Mixtures Bounds: Next, we compare our multivariate, self-normalized bounds to the "method of mixtures" bounds for sub-Gaussian concentration, in particular the following bound that follows from Example 4.2 of de la Peña et al. [41] and Theorem 1 of Abbasi-Yadkori et al. [3] and has become a staple in constructing confidence sets in online learning tasks [95, 165, 48]. We rephrase their sub-Gaussian result in the setting of "sub-ψ_N" concentration to ease comparison with our results.

Fact 2.4.4 (de la Peña et al. [41], Abbasi-Yadkori et al. [3])**.** *Let $(S_n, V_n)_{n\geq 0}$ be an $\mathbb{R}^d$-valued sub-ψ_N process where $V_n = \sum_{m=1}^n \mathbb{E}_{m-1}\Delta S_m \Delta S_m^\top$. Then, for any $\delta \in (0,1)$ and any $\rho > 0$, with probability at least $1-\delta$, simultaneously for all $n \geq 0$,*

$$\left\|(V_n + \rho I_d)^{-1/2} S_n\right\| \leq \sqrt{2\log\left(\frac{1}{\delta}\sqrt{\det\left(I_d + \rho^{-1}V_n\right)}\right)}.$$

We note that the above bound holds *only* in the case where the process $(S_n)_{n\geq 0}$ has sub-Gaussian increments, and it is not obvious whether or not a similar result holds for other tails, for more CGF-like functions ψ, and adapted (not predictable) V_n. In the case $\psi = \psi_N$, as noted in (2.4.1), our bound is of the form $\left\|V_n^{-1/2}S_n\right\| = O\left(\sqrt{\log\log(\gamma_{\max}(V_n)) + d\log\kappa(V_n)}\right)$. These two bounds (those based on the determinant of the variance proxy and those based on the condition number of the variance proxy) are fundamentally incomparable. When V_n is well-conditioned, we expect our bounds to be tighter than the bound in Fact 2.4.4, as our bounds will be of order $\approx \sqrt{\log\log(\gamma_{\max}(V_n)) + d}$. If $\kappa(V_n) \approx \gamma_{\max}(V_n)$, we may expect the determinant rate bound in Fact 2.4.4 to be tighter, as the bound provided by Theorem 2.4.1 will be of order $\approx \sqrt{\log\log(\gamma_{\max}(V_n)) + d\log(\gamma_{\max}(V_n))}$, and $d\log\gamma_{\max}(V_n) \geq \log\det(V_n)$ (ignoring the shift ρ in the covariance matrix). One particularly useful feature of our bounds is that the do not require a shift in variance proxy as the bound in Fact 2.4.4 does. It is an interesting open problem to derive determinant-rate self-normalized bounds under more general tail conditions and for adapted (not predictable) V_n.

Backwards Martingale Bounds: As a last point of comparison, we relate our bounds to the recent bounds constructed by Manole and Ramdas [116] using backwards or reverse martingale techniques. We note that the bounds of Manole and Ramdas [116] hold for *any* fixed norm on $\mathbb{R}^d$ (e.g. ℓ_p norms, for instance), but we only present the result in the case of the ℓ_2 norm, as this is the setting in which our bounds are comparable. The authors leverage the following bounds in estimating an unknown, multivariate mean from i.i.d. data. In our statement below, we center all observations so that the unknown mean always takes value zero for ease of comparison.

Fact 2.4.5 (Corollary 23 of Manole and Ramdas [116])**.** *Let $S_n := \sum_{m=1}^n X_m$, where $(X_n)_{n\geq 0}$ are i.i.d. with mean 0. Let $h : \mathbb{R} \to \mathbb{R}_{\geq 0}$ satisfy $\sum_{k=0}^\infty h(k)^{-1} \leq 1$, and let $\psi : [0, \lambda_{\max}) \to \mathbb{R}_{\geq 0}$ be CGF-like. Suppose that, for any $\lambda \in [0, \lambda_{\max})$ and $n \geq 0$, $\sup_{\nu\in\mathbb{S}^{d-1}} \log \mathbb{E}e^{\lambda\langle \nu, X_n\rangle} \leq \psi(\lambda)$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, simultaneously for all $n \geq 0$,*

$$\|S_n/\sqrt{n}\| \leq \frac{\sqrt{n}}{1-\epsilon} \cdot (\psi^*)^{-1}\left(\frac{2}{n}\left[\log(h(\log_2(n))) + \log\left(\frac{1}{\delta}\right) + \log N_{d-1}(\epsilon)\right]\right).$$

It is clear that the process $(S_n)_{n\geq 0}$ is sub-ψ with variance proxy $(V_n)_{n\geq 0}$ given by $V_n := nI_d$, and so Theorem 2.4.1 (taking $\rho = 1$) applied to this setting yields that, with probability at least $1-\delta$, simultaneously for all $n \geq 1$,

$$\|S_n/\sqrt{n}\| \leq \frac{\sqrt{n}}{1-\epsilon} \cdot (\psi^*)^{-1}\left(\frac{\alpha}{n}\left[\log(h(\log_\alpha(n))) + \log\left(\frac{\beta}{\delta(1-\beta^{-1})}\right) + \log N_{d-1}(\epsilon)\right]\right).$$

In this particular setting, our bound is almost equivalent to that of Manole and Ramdas [116], being looser is a vanishingly small additive factor $\log\left(\frac{\beta}{1-\beta^{-1}}\right)$ due to the covering argument needed to control the geometric "distortions" induced by the variance proxy $(V_n)_{n\geq 0}$. However, we note that our bound is significantly more general, as it allows for arbitrary martingale dependence between observed random variables. This is in contrast to the bound of Manole and Ramdas [116], as this bound is only valid if the data are known to be i.i.d. (or, at the very least, exchangeable). The argument used by Manole and Ramdas [116] does not readily generalize

to general dependence structures because they leverage reverse martingales in the exchangeable filtration, thus requiring that the data be exchangeable.

2.4.2 Vector Laws of the Iterated Logarithm

In Corollary 2.3.2, we discussed how our scalar bounds can be used to derive a version of the law of the iterated logarithm for scalar sub-ψ processes. In particular, this bound obtained the optimal constant matching the case of i.i.d. random variables (see Durrett [50], Chapter 8 or Howard et al. [75]), showing that our bounds are unimprovable asymptotically.

In the multivariate setting, our bounds do not just depend on $\log\log(\gamma_{\max}(V_n))$, but also on $\log\kappa(V_n)$. This dependence is not simply an artefact of our analysis, as de la Peña et al. [41] show an example of a 2-dimensional process $(S_n)_{n\geq 0}$ and $(V_n)_{n\geq 0}$ satisfying $\|V_n^{-1/2}S_n\| \sim \sqrt{\log\kappa(V_n)}$ almost surely.

In this section, we aim to show that our results are asymptotically optimal in the following sense. First, we show that, under a simple set of assumptions, if $(S_n, V_n)_{n\geq 0}$ is a sub-ψ, then $\limsup_{n\to\infty} \frac{\|V_n^{-1/2}S_n\|}{\sqrt{2\log\log(\gamma_{\max}(V_n))+d\log\kappa(V_n)}} \leq 1$ almost surely. Secondly, we show that this bound is "tight" in the sense that there exists a sub-ψ process $(S_n, V_n)_{n\geq 0}$ such that $\|V_n^{-1/2}S_n\| = \Theta(\sqrt{\log\log\gamma_{\max}(V_n) + d\log\kappa(V_n)})$ almost surely.

We start by presenting the first result, which can be viewed as an "upper law of the iterated logarithm". We prove this result in Section 2.7.

Corollary 2.4.6. *Let $(S_n)_{n\geq 0}$ be an $\mathbb{R}^d$-valued sub-ψ process with variance proxy $(V_n)_{n\geq 0}$. Suppose that (a) $\psi(\lambda) \sim \frac{\lambda^2}{2}$ as $\lambda \downarrow 0$, (b) $\gamma_{\min}(V_n) \xrightarrow[n\to\infty]{} \infty$ almost surely, and (c) and $\log(\gamma_{\max}(V_n))/\gamma_{\min}(V_n)) = o(1)$ almost surely. Then,*

$$\limsup_{n\to\infty} \frac{\|V_n^{-1/2}S_n\|}{\sqrt{2\log\log\gamma_{\max}(V_n) + d\log\kappa(V_n)}} \leq 1$$

almost surely.

We can compare the above corollary to the discussion at the beginning of Section 3 of de la Peña et al. [41], where the authors show that when $(S_n)_{n\geq 0}$ and $(V_n)_{n\geq 0}$ satisfy certain assumptions based on finiteness of pth moments, one has

$$\limsup_{n\to\infty} \frac{\|(V_n+V)^{-1/2}S_n\|}{\sqrt{\log\log\gamma_{\max}(V+V_n) + \log\kappa(V_n+V)}} = O(1) \qquad \text{almost surely},$$

where the constant masked by the "Big-Oh" notation maybe be random. Our bound is more precise than their bound in that (a) we obtain an explicit constant in our asymptotic bound, (b) the bound recovers the LIL in the case $d = 1$ (see the earlier discussed Corollary 2.3.2), and (c) our bound elicits explicit dependence on the ambient dimension d.

The remaining question is if the above law of the iterated logarithm is tight. As aforementioned, de la Peña et al. [41] show the existence of a two-dimensional process satisfying $\|V_n^{-1/2}S_n\| \sim \log\kappa(V_n)$ almost surely. We first describe this example, and then show how to

extend it to higher dimensions. In particular, we will construct a process that attains the same rate as the upper bound presented in our Corollary 2.4.6, up to a small, absolute constant. We start by describing the example of de la Peña et al. [41].

Example 2.4.7. Let $(\epsilon_n)_{n\geq 1}$ be a sequence of i.i.d. $\mathcal{N}(0,1)$ random variables, and let $(\mathcal{F}_n)_{n\geq 0}$ be the natural filtration associated with $(\epsilon_n)_{n\geq 1}$. First, define the regressors $(U_n)_{n\geq 1}$ by $U_1 = 0$ and $U_{n+1} := \overline{U}_n + \overline{\epsilon}_n$, where for a sequence $(y_n)_{n\geq 1}$ we define $\overline{y}_n := \frac{1}{t}(y_1 + y_2 + \cdots + y_n)$. Then, embed these regressors into $\mathbb{R}^2$ by defining the process $(X_n)_{n\geq 1}$ as $X_n := (1, U_n)^\top$. Clearly, by construction, the process $(X_n)_{n\geq 1}$ is $(\mathcal{F}_n)_{n\geq 0}$-predictable.

With these sequentially constructed regressors, one can construct a martingale $(S_n)_{n\geq 0}$ with respect to $(\mathcal{F}_n)_{n\geq 0}$ given by $S_n := \sum_{m=1}^n \epsilon_m X_m$ and a corresponding predictable covariance process $(V_n)_{n\geq 0}$ given by $V_n = \sum_{m=1}^n X_m X_m^\top$. de la Peña et al. [41] show that the following hold almost surely:

1. $\gamma_{\max}(V_n) \sim n\,(1 + \sum_{m=1}^\infty s^{-1}\epsilon_m)$,
2. $\gamma_{\min}(V_n) \sim \frac{\log(n)}{1+\sum_{m=1}^\infty m^{-1}\epsilon_m}$, and
3. $\|V_n^{-1/2} S_n\| \sim \sqrt{\log(n)}$.

Noting that $\log \kappa(V_n) = \log(\gamma_{\max}(V_n)/\gamma_{\min}(V_n)) \sim \log(n)$, we see that we have $\|V_n^{-1/2} S_n\| \sim \sqrt{\log \kappa(V_n)}$ almost surely. Further, it is easily checked that $(S_n, V_n)_{n\geq 0}$ is sub-ψ_N, per Definition 2.2.2. Thus, this example shows that the logarithmic dependence on $\kappa(V_n)$ in Theorem 2.4.1 cannot, in general, be dropped.

While the above example demonstrates the inevitability of having $\log \kappa(V_n)$ appear in non-asymptotic, self-normalized concentration for vector-valued processes, it does not capture dependence on dimensionality. In the next example, we show that there exists sub-ψ processes $(S_n, V_n)_{n\geq 0}$ such that $\|V_n^{-1/2} S_n\| \sim \sqrt{\frac{d}{2}\log \kappa(V_n)}$, showing our upper bounds are within a multiplicative factor $\sqrt{2}$ of optimal.

Example 2.4.8. Suppose d is even. Let $(S_n^{(1)})_{n\geq 0}, \ldots, (S_n^{(d/2)})_{n\geq 0}$ be i.i.d. copies of the process constructed in Example 2.4.7, $(V_n^{(1)})_{n\geq 0}, \ldots, (V_n^{(d/2)})_{n\geq 0}$ the corresponding predictable covariance processes, and $(\mathcal{F}_n)_{n\geq 0}$ the smallest filtration for which $(\epsilon_n^{(1)})_{n\geq 1}, \ldots, (\epsilon_n^{(d/2)})_{n\geq 1}$ are adapted, i.e. the filtration given by $\mathcal{F}_n := \mathcal{F}_n^{(1)} \bigvee \cdots \bigvee \mathcal{F}_n^{(d/2)}$, where $\mathcal{F} \bigvee \mathcal{G}$ denotes the "join" of σ-algebras $\mathcal{F}, \mathcal{G}$, i.e. the smallest σ-algebra containing both.

Define the $\mathbb{R}^d$-valued process $(S_n)_{n\geq 0}$ by $S_n := \left(S_n^{(1)}, \ldots, S_n^{(d/2)}\right)$, and the corresponding covariance process $(V_n)_{n\geq 0}$ by

$$V_n := \begin{pmatrix} V_n^{(1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & V_n^{(2)} & \cdots & \mathbf{0} \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \cdots & V_n^{(d/2)} \end{pmatrix}.$$

Clearly $(S_n)_{n\geq 0}$ is $(\mathcal{F}_n)_{n\geq 0}$-adapted and $(V_n)_{n\geq 0}$ is $(\mathcal{F}_n)_{n\geq 0}$-predictable. Moreover, it can readily be checked that $(S_n, V_n)_{n\geq 0}$ is a sub-ψ_N process.

Since V_n is a block-diagonal matrix, we clearly have $\gamma_{\max}(V_n) = \max_{i\in[d/2]} \gamma_{\max}(V_n^{(i)})$ and $\gamma_{\min}(V_n) = \min_{i\in[d/2]} \gamma_{\min}(V_n^{(i)})$. Thus, using the reasoning on the almost sure behavior on $\gamma_{\max}(V_n^{(i)})$ and $\gamma_{\min}(V_n^{(i)})$ presented in Example 2.4.7, we see that $\log \kappa(V_n) \sim \log(n)$ almost surely. Further, it isn't hard to see that

$$\begin{aligned}\|V_n^{-1/2} S_n\|^2 &= S_n^\top V_n^{-1} S_n \\ &= (S_n^{(1)})^\top (V_n^{(1)})^{-1} S_n^{(1)} + \cdots + (S_n^{(d/2)})^\top (V_n^{(d/2)})^{-1} S_n^{(d/2)} \sim \frac{d}{2} \log \kappa(V_n).\end{aligned}$$

Thus, we have shown that, up to small constants, the dependence on $\log \kappa(V_n)$ and d in Theorem 2.4.1 (and thus the corresponding dependence in Corollary 2.4.6) is unimprovable.

2.5 Applications to Online Linear Regression

We now use our self-normalized bounds to construct confidence ellipsoids for slope estimation in online linear regression. In online linear regression, a statistician interacts with an environment over a sequence of rounds. At the beginning of each round, he adaptively (perhaps using observations from previous rounds) selects a point $X_n \in \mathbb{R}^d$, and then observes noisy feedback $Y_n := \langle X_n, \theta^* \rangle + \epsilon_n$, where ϵ_n represents some mean zero noise variable and θ^* is a fixed slope vector. The goal of the statistician is to produce a *confidence sequence* for the unknown slope vector — that is, a time indexed sequences of sets that all simultaneously contain the unknown parameter with high probability. We formalize the online linear regression model as follows.

Model 2.5.1 (Online Linear Regression). *Let $(\mathcal{F}_n)_{n\geq 0}$ be a filtration and $\theta^* \in \mathbb{R}^d$ a fixed (unknown) slope vector. The online linear regression model is characterized by three processes: (a) a $(\mathcal{F}_n)_{n\geq 0}$-predictable $\mathbb{R}^d$-valued sequence $(X_n)_{n\geq 1}$ representing adaptively-chosen covariates, (b) a $(\mathcal{F}_n)_{n\geq 0}$-adapted scalar-valued processes $(\epsilon_n)_{n\geq 1}$ representing noise, and (c) $(Y_n)_{n\geq 1}$ given as $Y_n = \langle X_n, \theta^* \rangle + \epsilon_n$ representing noisy responses. We assume the residual process $S_n := \sum_{m=1}^n \epsilon_m X_m$ is sub-ψ with (predictable) variance proxy $V_n := \sum_{m=1}^n X_m X_m^\top$, where ψ is a super-Gaussian CGF-like function.*

For a fixed regularization parameter $\rho > 0$, we consider the sequence of **least squares estimates with shrinkage** given by

$$\widehat{\theta}_n := (\mathbf{X}_n^\top \mathbf{X}_n \vee \rho I_d)^{-1} \mathbf{X}_n^\top \mathbf{Y}_n, \quad n \geq 1,$$

where $\mathbf{X}_n \in \mathbb{R}^{t\times d}$ has $X_1, \ldots, X_n$ as its rows and $\mathbf{Y}_n \in \mathbb{R}^d$ is a column vectors with $Y_1, \ldots, Y_n$ as its entries. Clearly $\widehat{\theta}_n$ reduces to the standard least-squares estimator when $\gamma_{\min}(\mathbf{X}_n^\top \mathbf{X}_n) \geq \rho$.

The assumption that (S_n, V_n) is sub-ψ is often mild. For example, it is satisfied (a) if $\log \mathbb{E}_{n-1} \exp\{\lambda \epsilon_n\} \leq \psi_N(\lambda)$ for all $\lambda \in \mathbb{R}$ (i.e. ϵ_n is conditionally sub-Gaussian), or (b) if $\|X_n\| \leq 1$ for all $n \geq 1$ and $\log \mathbb{E}_{n-1} \exp\{\pm\lambda \epsilon_n\} \leq \psi(\lambda)$ for some super-Gaussian ψ and all $\lambda \in [0, \lambda_{\max})$. We prove this in Proposition 2.B.1 in Appendix 2.B. The assumption that $\|X_n\| \leq 1$ for all $n \geq 1$ in the above can be replaced with the assumption that $\|X_n\| \leq R$ for any fixed $R > 0$ by appropriate rescaling. This type of boundedness assumption is regularly made in the mult-armed bandit literature [3, 30, 102], and thus has practical relevance.

2.5.1 Time-Uniform Confidence Ellipsoids

We briefly discuss how confidence ellipsoids are constructed in classical least-squares regression. In this setting, one observes a matrix of covariates $\mathbf{X} \in \mathbb{R}^{n \times d}$ and a response vector $\mathbf{Y} \in \mathbb{R}^n$ given by $\mathbf{Y} = \mathbf{X}\theta^* + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2 I_d)$. If $\mathbf{X}^\top \mathbf{X}$ is full rank, it is well-known [136, 91] that the least-squares estimate for θ^*, given by $\widehat{\theta} := (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$, satisfies

$$\sigma^{-1} \|(\mathbf{X}^\top \mathbf{X})^{1/2} (\widehat{\theta} - \theta^*)\| \sim \chi^2_q,$$

where χ^2_q denotes the Chi-squared distribution with q degrees of freedom. Letting $x_{q,\delta}$ denote its δth upper quantile[3] it follows that the set

$$\mathcal{C} := \{\theta \in \mathbb{R}^d : \sigma^{-1} \|(\mathbf{X}^\top \mathbf{X})^{1/2} (\widehat{\theta} - \theta)\| \leq x_{q,\delta}\}$$

forms an exact $1 - \delta$ confidence ellipsoid for θ^* centered at $\widehat{\theta}$.

The above confidence ellipsoid fails to be valid when $\mathbf{X}$ is no longer fixed or when the added noise variables are no longer i.i.d. Gaussian, which is the case presented in our heuristic model above. To circumvent this failure of classical statistical machinery, we can leverage our self-normalized bounds for vector-valued processes to construct confidence ellipsoids for θ^* that are valid across all time steps uniformly. We do exactly this in the following theorem.

Theorem 2.5.2. *Consider Model 2.5.1, let $\delta \in (0, 1)$ be arbitrary and set $V_n := \mathbf{X}_n^\top \mathbf{X}_n = \sum_{m=1}^n X_m X_m^\top$. Then, with probability at least $1 - \delta$, simultaneously for all $n \geq 1$, we have*

$$\|(V_n \vee \rho I_d)^{1/2} (\widehat{\theta}_n - \theta^*)\| < \frac{\sqrt{\gamma_{\min}(V_n \vee \rho I_d)}}{1 - \epsilon} \cdot (\psi^*)^{-1} \left(\frac{\alpha}{\gamma_{\min}(V_n \vee \rho I_d)} L_\rho(V_n) \right) + \sqrt{\rho} \|\theta^*\| \mathbb{1}_{\gamma_{\min}(V_n) < \rho},$$

where the parameters $\alpha, \epsilon, \beta, h$ and the function L_ρ (which partially masks parameter dependence) are as outlined in Theorem 2.4.1.

The dependence on the norm of the unknown slope vector in Theorem 2.5.2 may be a mild irritant, but note that the indicator function multiplying it ensures that the term necessarily disappears for large n. Also note that assuming a known bound on $\|\theta^*\|$ is common in many statistical tasks, in particular those related to bandit optimization [3, 30, 102]. Nonetheless, this dependence can be fully removed by sacrificing validity of the bound above bound when the minimum eigenvalue of $\mathbf{X}_n^\top \mathbf{X}_n$ is small. This is made precise as follows.

Corollary 2.5.3. *Assume the same setup as Theorem 2.5.2. Then, we have*

$$\mathbb{P}\left(\exists n \geq 0 : \|V_n^{1/2} (\widehat{\theta}_n - \theta^*)\| \geq \frac{\sqrt{\gamma_{\min}(V_n)}}{1 - \epsilon} \cdot (\psi^*)^{-1} \left(\frac{\alpha}{\gamma_{\min}(V_n)} L_\rho(V_n) \right) \text{ and } V_n \succeq \rho I_d \right) \leq \delta.$$

We can similarly derive a result for the ridge estimators of the unknown slope parameter, which are given by $\widetilde{\theta}_n := (\mathbf{X}_n^\top \mathbf{X}_n + \rho I_d)^{-1} \mathbf{X}_n^\top \mathbf{Y}_n$. We prove the following in Appendix 2.B.

[3] that is, $x_{q,\delta} > 0$ is the unique value satisfying $\mathbb{P}(X \geq x_{q,\delta}) = \delta$, where $X \sim \chi^2_q$

Corollary 2.5.4. *Assume the setup outlined in Theorem 2.5.2 above. Consider the sequence of ridge estimates $(\widetilde{\theta}_n)_{n\geq 1}$ given by*

$$\widetilde{\theta}_n := (\mathbf{X}_n^\top \mathbf{X}_n + \rho I_d)^{-1} \mathbf{X}_n^\top \mathbf{Y}_n.$$

Set $V_n := \mathbf{X}_n^\top \mathbf{X}_n$. With probability at least $1-\delta$, simultaneously for all $n \geq 1$, we have

$$\|(V_n+\rho I_d)^{1/2}(\widetilde{\theta}_n-\theta^*)\| < \frac{\sqrt{\gamma_{\min}(V_n+\rho I_d)}}{1-\epsilon}\cdot(\psi^*)^{-1}\left(\frac{\alpha}{\gamma_{\min}(V_n+\rho I_d)}L_\rho(V_n+\rho I_d)\right)+\sqrt{\rho}\|\theta^*\|,$$

where the parameters $\alpha, \epsilon, \beta, h$ and the function L_ρ (which partially masks parameter dependence) are as outlined in Theorem 2.4.1.

Comparison with Existing Bounds: Many results concerning finite-sample properties of regression estimators are based either in the setting of fixed design [158, 5] or in the the case of independent covariates [96, 97]. Moreover, these results are more often than not concerned with bounding the ℓ_2-error of the estimator, i.e. the quantity $\|\widehat{\theta}_n - \theta^*\|$, as opposed to the self-normalized quantities we study.

The main points of comparison for our results have been derived in the online learning/regression literature. We compare our results to those of Abbasi-Yadkori et al. [3]. In their work, Abbasi-Yadkori et al. [3] construct a confidence sequence for estimating an unknown slope vector θ^* by utilizing self-normalized concentration for sub-Gaussian processes (in particular, leveraging a Gaussian mixture technique that dates back to Example 4.2 in de la Peña et al. [41]). While subsequent confidence sequences have been derived in the setting of regression with variance estimation [48], semiparametric regression with bounded confounding [95], and ridge regression in reproducing kernel Hilbert spaces [165, 2], we focus just on the original contributions of Abbasi-Yadkori et al. [3] since all subsequent results exhibit the same rate and hold only in the setting of sub-Gaussian noise.

Fact 2.5.5 (Theorem 2 of Abbasi-Yadkori et al. [3])**.** *Let $(\mathcal{F}_n)_{n\geq 0}$ be a filtration, let $(X_n)_{n\geq 1}$ be an $(\mathcal{F}_n)_{n\geq 0}$-predictable sequence in $\mathbb{R}^d$, and let $(\epsilon_n)_{n\geq 1}$ be a real-valued $(\mathcal{F}_n)_{n\geq 1}$-adapted sequence such that conditional on $\mathcal{F}_{n-1}$, $\log \mathbb{E}_{n-1} \exp\{\lambda \epsilon_n\} \leq \psi_N(\lambda)$ for all $\lambda \in \mathbb{R}$. Then, for any $\rho > 0$ and $\delta \in (0,1)$,*

$$\mathbb{P}\left(\exists n \geq 0 : \|(V_n+\rho I_d)^{1/2}(\widetilde{\theta}_n - \theta^*)\| \geq \sqrt{2\log\left(\frac{1}{\delta}\sqrt{\det(I_d+\rho^{-1}V_n)}\right)} + \sqrt{\rho}\|\theta^*\|\right) \leq \delta,$$

where $\widetilde{\theta}_n$ is the ridge regression estimator outlined in Corollary 2.5.4 and $V_n := \sum_{m=1}^n X_m X_m^\top$.

We compare our results to Fact 2.5.5 in the setting $\psi = \psi_N$, as this is the only setting in which the results of Abbasi-Yadkori et al. [3] are valid. We first qualitatively compare the above confidence sequence to the one presented in Corollary 2.5.4. Both bounds suffer the same dependence on the norm of the unknown slope vectors and differ only in the first term. Namely, as noted earlier, $(\psi_N^*)^{-1}(u) = \psi_N^{-1}(u) = \sqrt{2u}$, so in this setting our bound reduces to the form

$$\|(V_n+\rho I_d)^{1/2}(\widetilde{\theta}_n - \theta^*)\| \leq \sqrt{2\alpha L_\rho(V_n+\rho I_d)} + \sqrt{\rho}\|\theta^*\|$$

$$= O\left(\sqrt{\log\log\gamma_{\max}(\rho^{-1}V_n + I_d) + d\log\kappa(V_n + \rho I_d)}\right)$$

simultaneously for all $n \geq 0$ with probability at least $1 - \delta$. Thus, the same comparison made in Subsection 2.4.1 applies in this setting — when the (shifted) covariance $V_n + \rho I_d$ is poorly conditioned, the bound presented in Fact 2.5.5 would be expected to be tighter. Likewise, when $V_n + \rho I_d$ is well-conditioned, Corollary 2.5.4 may be tighter.

A more interesting comparison is between Fact 2.5.5 and Theorem 2.5.2 (more specifically, Corollary 2.5.3 following the theorem statement). Whereas the above fact provides convergence guarantees for ridge regression estimates, Corollary 2.5.3 applies directly to the unregularized, least-squares estimates of the unknown slope vector. In particular, the bound does not depend on $\|\theta^*\|$, the norm of the unknown slope vector. This may be desirable in many statistical settings in which either advanced knowledge of such a bound is unavailable or only a loose bound on the quantity is known. Moreover, this bound is interesting in itself as no shift in covariance is required in constructing the confidence ellipsoids. The rate given by Corollary 2.5.3 is essentially the same that provided by Corollary 2.5.4 modulo the presence of a shift in the covariance matrix, i.e. the corollary yields that with high probability, uniformly in time, $\|V_n^{1/2}(\widehat{\theta}_n - \theta^*)\| = O\left(\sqrt{\log\log_{\max}(\rho^{-1}V_n) + d\log\kappa(V_n)}\right)$.

2.5.2 Applications to Vector Autoregressive Models

We now show how to apply our confidence ellipsoids from Subsection 2.5.1 in the section to a vector autoregressive model. We take inspiration from Bercu and Touati [14], who leverage self-normalized concentration results for scalar-valued processes to measure the convergence of least-squares and Yule-Walker estimates for a simple one stage autoregressive model (i.e. an $\mathrm{AR}(1)$ model). We focus solely on the least-squares estimates in the sequel. We provide a brief, high-level qualitative comparison between these results and our own. The following results may be of practical interest as autoregressive models and other time series models are frequently applied to problems in econometrics [5, 137] and finance [126, 39].

The results we provide in this section are more general than those of Bercu and Touati [14] in three ways. First, these authors assume that all noise variables are Gaussian, whereas we allow the noise to be instead conditionally sub-Gaussian. Second, we handle a vector autoregressive model, whereas Bercu and Touati [14] only handle the univariate case. Lastly, we handle the problem of general autoregression with p-stages of lag, whereas Bercu and Touati [14] only handle the case $p = 1$. Our bounds are also different than those of Bercu and Touati [14] in that they are derived in terms of the predictable covariance associated with observations, whereas those of Bercu and Touati [14] are stated in terms of total number of observations. With these comparisons in hands, we now describe the p-stage vector autoregressive model (hereinafter referred to as $\mathrm{VAR}(p)$ for short).

Model 2.5.6. *A* $\mathbf{p}$ ***stage vector-valued autoregressive model****, denoted by* $\mathrm{VAR}(p)$*, is an* $\mathbb{R}^d$*-valued process* $(Y_n)_{n\geq -p+1}$ *such that* $Y_{-p+1}, \ldots, Y_0 \in \mathbb{R}^d$ *and* $Y_n := \sum_{i=1}^p A_i Y_{n-i} + \epsilon_n$ *where (a)* $A_i \in \mathbb{R}^{d\times d}$ *are fixed matrices for all* $i \in [p]$*, and (b)* ϵ_n *satisfies* $\log \mathbb{E}_{n-1}\exp\{\lambda\langle\nu, \epsilon_n\rangle\} \leq \frac{\lambda^2}{2}$*, where* $\nu \in \mathbb{S}^{d-1}$ *and* $\lambda \in [0, \lambda_{\max})$*. In the above,* $\mathbb{E}_{n-1}[\cdot] := \mathbb{E}(\cdot \mid \mathcal{F}_{n-1})$*, where* $(\mathcal{F}_n)_{n\geq 0}$ *is the filtration given by* $\mathcal{F}_n := \sigma(Y_m : -p+1 \leq m \leq n)$*, for any* $n \geq 1$*.*

For more details on vector autoregressive models, see [68]. In words, a process $(Y_n)_{n\geq 0}$ satisfies the conditions of a p-stage autoregressive (or $\mathrm{VAR}(p)$) model if Y_n is a linear function of $Y_{n-1}, \ldots, Y_{n-p}$ plus mean zero noise. In the above, the values $Y_{-p+1}, \ldots, Y_0$ are treated as fixed nonrandom vectors, as is typical in much of the time series analysis literature. However, all results in the sequel still hold if $Y_{-p+1}, \ldots, Y_0$ are random variables that are independent of the noise sequence $(\epsilon_n)_{n\geq 1}$. Typically, the $\mathrm{VAR}(p)$ model also admits a vector mean parameter $\mu \in \mathbb{R}^d$, having the relationship $Y_n = \mu + \sum_{i=1}^p A_p Y_{n-p} + \epsilon_n$ for all $n \geq 1$, but we omit this to simplify exposition.

The goal of the statistician running an autoregressive model is twofold: (a) to estimate the unknown matrix parameters $A_1, \ldots, A_p$, and (b) to calibrate confidence in his estimates. Before discussing classical approaches to estimating these parameters, we simplify notation. We define the "stacked" transition matrix $\Pi \in \mathbb{R}^{d\times dp}$ and process vectors $(X_n)_{n\geq 1} \in \mathbb{R}^{dp}$ by

$$\Pi := (A_1, \ldots, A_p) \quad \text{and} \quad X_n := (Y_{n-1}, Y_{n-2}, \ldots, Y_{n-p})^\top.$$

For $i \in [d]$, we denote by $\pi(i) \in \mathbb{R}^{dp}$ the ith row of the stacked matrix Π. We likewise denote by $\epsilon_n(i) \in \mathbb{R}$ the ith component of the noise vector ϵ_n and $X_n(i)$ the ith component of the state vector X_n. Let $\mathbf{X}_n \in \mathbb{R}^{n\times dp}$ be the matrix with $X_1, \ldots, X_n$ as its rows, and let $\mathbf{Y}_n \in \mathbb{R}^{n\times d}$ have $Y_1, \ldots, Y_n$ as its rows. For $i \in [d]$, let $\mathbf{Y}_n(i) \in \mathbb{R}^n$ denote the ith column of $\mathbf{Y}_n$.

If $(\epsilon_n)_{n\geq 1}$ are i.i.d. $\mathcal{N}(0, \sigma^2 I_d)$ with known standard deviation σ, it is well-known (see Hamilton [68], Chapter 11) that the maximum likelihood estimate for Π at time $n \geq 1$, for now denoted $\widehat{\Pi}_n$, has rows $\widehat{\pi}_n(i)$ that are just the least-squares estimates given by

$$\widehat{\pi}_n(i) := \left(\mathbf{X}_n^\top \mathbf{X}_n\right)^{-1} \mathbf{X}_n^\top \mathbf{Y}_n(i). \tag{2.5.1}$$

It thus makes sense to study the convergence on these row-wise estimates in the remainder of this section. We focus on studying the convergence of a single row estimate, as the general case follows from union-bounding over the validity of the d row estimates. The proof of the following is a straightforward consequence of Theorem 2.5.2, and we provide the brief proof of the result in Appendix 2.B.

Corollary 2.5.7. *For a fixed coordinate $i \in [d]$, let $(\widehat{\pi}_n(i))_{n\geq 1}$ be the sequence of estimates outlined in* (2.5.1)*. Let $\rho > 0$ and $\delta \in (0,1)$ be arbitrary. Define the covariance process $(V_n)_{n\geq 1}$ by $V_n := \mathbf{X}_n^\top \mathbf{X}_n$. Then, with probability at least $1-\delta$, simultaneously for all $n \geq 1$, we have*

$$\|(V_n \vee \rho I_{dp})^{1/2}(\widehat{\pi}_n(i) - \pi(i))\| \leq \frac{1}{1-\epsilon}\sqrt{2\alpha L_\rho(V_n)} + \sqrt{\rho}\|\theta^*\|\mathbb{1}_{\gamma_{\min}(V_n)<\rho},$$

where the parameters $\alpha, \epsilon, \beta, h$ and the function L_ρ (which partially masks parameter dependence) are as outlined in Theorem 2.4.1.

We now compare Corollary 2.5.7 to traditional asymptotic analyses of equation estimation in the $\mathrm{VAR}(p)$ model. First, note that, in Model 2.5.6 and Corollary 2.5.7, we place no assumptions on the matrices $A_1, \ldots, A_p \in \mathbb{R}^{d\times d}$. This is in contrast to typical asymptotic analyses, which must assume that all solutions $z \in \mathbb{C}$ to the equation

$$\det\left(I_d + A_1 z + A_2 z^2 + \cdots + A_p z^p\right) = 0 \tag{2.5.2}$$

have modulus $|z| > 1$ (which we assume holds for validity of the following comparison). In the setting of independent Gaussian noise, as discussed above, the stacked process $(X_n)_{n\geq 1}$ is ergodic and admits some stationary distribution π over $\mathbb{R}^{dp}$. It is known is known that, for any $i \in [d]$, $\sqrt{n}V^{1/2}(\widehat{\pi}_n(i) - \pi(i)) \Rightarrow \mathcal{N}(0, \sigma^2 I_{dp})$, where $V = \mathbb{E}_\pi[X_n X_n^\top] = \lim_{n\to\infty} \frac{1}{n}\sum_{m=1}^n X_m X_m^\top$ (the final equality comes from the ergodicity of $(X_n)_{n\geq 1}$). For large n, one would thus expect that $\|V_n^{1/2}(\widehat{\pi}_n(i) - \pi(i))\| \lesssim \sqrt{dp}$ with high probability.

We compare our non-asymptotic bounds to this rate. Observe that, Corollary 2.5.7 yields that, with high-probability, simultaneously for all $n \geq 1$,

$$\|V_n^{1/2}(\widehat{\pi}_i(n) - \pi(i))\| = O\left(\sqrt{dp\log\kappa(V_n) + \log\log\left(\gamma_{\max}(V_n)\right)}\right).$$

If $\lim_{n\to\infty}\frac{1}{n}\sum_{m=1}^n X_m X_m^\top = V$ for some fixed positive-definite matrix V (as will be the case if the ϵ_n are i.i.d.) and n is a sufficiently large "target round", we can view the above as stating $\|V_n^{1/2}(\widehat{\pi}_i(t) - \pi(i))\|$ is bounded above by a term growing like $O(\sqrt{\log\log\gamma_{\max}(V_n) + dp})$ (since $\kappa(V_n) = \kappa(V) = O(1)$ for large n, almost surely). As expected in time-uniform concentration, the bounds presented in Corollary 2.5.7 are looser than those provided by the central limit theorem by a doubly logarithmic factor.

Comparison with Existing Bounds: We lastly make a brief comparison with the bounds of Bercu and Touati [14] in the univariate case. In this case, the autoregressive model is parameterized by a scalar $a \in \mathbb{R}$ instead of a sequence of matrices. We thus denote the least-squares estimator of a at time $n \geq 1$ as $\widehat{a}_n := \frac{\sum_{m=1}^n X_{n-1}X_n}{\sum_{m=1}^n X_{n-1}^2}$, departing from our notation of $\widehat{\pi}_n$, which was relevant for estimating a row in a stacked matrix. We state the bound of Bercu and Touati [14] for convenience.

Fact 2.5.8 (Corollary 5.2 of Bercu and Touati [14]). *Suppose $a \in \mathbb{R}$ is fixed. Further, suppose $(Y_n)_{n\geq 0}$ is given by $Y_0 \sim \mathcal{N}(0,1)$ and $Y_n := aY_{n-1} + \epsilon_n$, where $(\epsilon_n)_{n\geq 1}$ are a sequence of i.i.d. $\mathcal{N}(0,1)$ random variables independent of Y_0. Then, for any fixed $x > 0$ and $n \geq 1$, we have*

$$\mathbb{P}\left(|\widehat{a}_n - a| \geq x\right) \leq 2\exp\left\{\frac{-nx^2}{2(1+y_x)}\right\},$$

where y_x is the unique solution to the equation $\psi_{P,1}^*(y_x) = x^2$, where we recall $\psi_{P,1}^*(u) = (1+u)\log(1+u) - u$.

We draw several high-level comparisons between the bounds. First, the bound in Corollary 2.5.7 is self-normalized, being defined in terms of the empirical variance $V_n = \sum_{m=1}^n Y_{n-1}^2$. The bound in Fact 2.5.8, on the other hand, depends just on the number of samples used to construct the least-squares estimator, and thus is not self-normalized. Another difference between the conclusions of Fact 2.5.8 and Corollary 2.5.7 is that Fact 2.5.8 holds only for an individual, fixed sample size $n \geq 1$ whereas Corollary 2.5.7 is valid for all $n \geq 1$ *simultaneously*. To use Fact 2.5.8 to obtain a time-uniform guarantee, one would need to use a union bound argument to allocate the total failure probability over many rounds. The setting Fact 2.5.8 is also highly parametric, assuming that both the noise and initial state are i.i.d. Gaussian random variables.

Corollary 2.5.7, on the other hand, makes no such assumptions, allowing an arbitrary initial state and conditionally sub-Gaussian noise variables.

An explicit comparison of the above bounds is difficult, but we can empirically compare the bounds by simulating a simple $\mathrm{AR}(1)$ model. We provide such a comparison in Figure 2.3 in Appendix 2.D, which plots, for a fixed failure probability $\delta \in (0, 1)$ the autoregressive guarantee from Corollary 2.5.7 against the corresponding guarantee provided by Fact 2.5.8. In Subfigure 2.3a we plot the bound from Fact 2.5.8 *without* providing a union bound correction. We thus emphasize that, as plotted, the boundary is only valid point-wise, and not for all $n \geq 1$ simultaneously or for arbitrary stopping times. In Subfigure 2.3b, we make a union bound correction. Figure 2.3 indicates that Corollary 2.5.7 performs similarly to Fact 2.5.8 when specified to the scalar setting. We believe our bound may be preferable in application over that of Fact 2.5.8 as it is not only significantly more general, but it also inherently adapts to the variance of the observed autoregressive iterates.

2.6 A Self-Normalized, Multivariate Empirical Bernstein Inequality for Bounded Vectors

We construct a multivariate empirical Bernstein inequality, extending the Theorem 4 of Howard et al. [75] to higher dimensions. Empirical Bernstein-style bounds serve as a useful tool in common statistical tasks such as forming confidence sequences for estimating unknown means [161]. These bounds are of practical importance as they inherently adapt to the variance of a sequence of observations. If actual observations are tightly clustered, the resulting confidence bounds will be tighter. Likewise, if observations are well-dispersed, the resulting confidence set will be more conservative. To apply empirical Bernstein these bounds, a statistician must only know that the observations belong to a some bounded set.

To the best of our knowledge, we provide the first multivariate, self-normalized empirical Bernstein. Existing bounds either only hold in the scalar setting [161, 75], or do not normalize the quantity being estimated by the accumulated variance process (See, for instance, the work of Cutkosky [36] in the case of Hilbert space-valued variables). Providing confidence ellipsoids for mean estimation is desirable as it allows the confidence sets to reflect the "total amount of information" gathered in any given direction.

We now present the primary result of this section. In our result, we focus on the case where all observations have norm bounded above by $1/2$ for simplicity. This is mostly for theoretical convenience. While the more general setting where $(X_n)_{n\geq 1}$ belongs to some bounded, convex set is of interest, it can be readily analyzed by reducing to the case where observations lie in $\frac{1}{2}\mathbb{B}_d$[4].

[4]If $(X_n)_{n\geq 1}$ lies in some arbitrary convex, bounded set $K \subset \mathbb{R}^d$, we can first compute the outer John ellipsoid E of K, which is the minimal volume ellipsoid containing the convex set K [82]. In many settings, such as in the setting where K belongs to certain families of polytopes, there are computationally efficient algorithms that compute E [34, 149]. With E at hand, we can "recenter" our observations by defining a new sequence $(X'_n)_{n\geq 0}$ by $X'_n := X_n - p$, where $p := \int_E x dx$ is the center of mass of E. We then have the equality $E - p = \frac{1}{2}A^{1/2}\mathbb{B}_d$, where A is some positive semi-definite matrix. We thus transform our observations into a final sequence $(X''_n)_{n\geq 0}$ defined by $X''_n := A^{-1/2}(X_n - p)$, which lies almost surely in $\frac{1}{2}\mathbb{B}_d$.

For the remainder of this section, we adopt the notation ℓ_ρ^δ and L_ρ^δ instead of ℓ_ρ and L_ρ to explicitly make known the dependence on the confidence parameter δ. We make this dependence explicit as we will be union bounding in the sequel, and thus it will be useful to track the dependence.

Theorem 2.6.1. *Let $(X_n)_{n\geq 1}$ be a sequence of random vectors in $\mathbb{R}^d$ such that $\|X_n\| \leq 1/2$ almost surely, for all $n \geq 1$, and let $(\mathcal{F}_n)_{n\geq 0}$ be a filtration to which $(X_n)_{n\geq 1}$ is adapted. Then, the process $(S_n)_{n\geq 0}$ given by $S_n := \sum_{m=1}^n (X_m - \mathbb{E}_{m-1}X_m)$ is sub-$\psi_{E,1}$ with variance proxy $(V_n)_{n\geq 0}$ given by $V_n := \sum_{m=1}^n (X_m - \widehat{\mu}_{m-1})(X_m - \widehat{\mu}_{m-1})^\top$, where $\widehat{\mu}_n := n^{-1}\sum_{m=1}^n X_m$. Thus, by Theorem 2.4.1, for any fixed choice of parameters $\rho, \alpha, \delta, \beta, \epsilon, h$, we have*

$$\mathbb{P}\left(\exists n \geq 0 : \left\|(V_n \vee \rho)^{-1/2} S_n\right\| \geq \frac{\sqrt{\gamma_{\min}(V_n \vee \rho)}}{1-\epsilon} \cdot (\psi_{E,1}^*)^{-1}\left(\frac{\alpha L_\rho^\delta(V_n)}{\gamma_{\min}(V_n \vee \rho)}\right)\right) \leq \delta.$$

In particular, since a sub-$\psi_{E,1}$ process is sub-$\psi_{G,1}$, this implies that, with probability at least $1-\delta$, simultaneously for all $n \geq 0$,

$$\|(V_n \vee \rho)^{-1/2} S_n\| \leq \sqrt{2\alpha L_\rho^\delta(V_n)} + \frac{\alpha L_\rho^\delta(V_n)}{\gamma_{\min}(V_n \vee \rho)}.$$

We now compare the bound presented in Theorem 2.6.1 to existing empirical Bernstein-style results. In particular, our main point of comparison will be the following, scalar-valued bound from Howard et al. [75].

Proposition 2.6.2 ([75, Theorem 4]). *Suppose $(X_n)_{n\geq 1}$ satisfies $X_n \in [-1/2, 1/2]$ almost surely for all $n \geq 1$, and let $(S_n)_{n\geq 0}$, $(\widehat{\mu}_n)_{n\geq 0}$, and $(V_n)_{n\geq 0}$ be as in Theorem 2.6.1. For any choice of parameters α, δ, h, ρ, we have with probability at least $1-\delta$, simultaneously for all $n \geq 1$,*

$$|S_n| \leq \sqrt{k_1^2(V_n \vee \rho)\ell_\rho^{2\delta}(V_n) + k_2^2 \ell_\rho^{2\delta}(V_n)^2} + k_2 \ell_\rho^{2\delta}(V_n),$$

where $k_1 := \frac{\alpha^{1/4} + \alpha^{-1/4}}{\sqrt{2}}$, $k_2 := \frac{\sqrt{\alpha}+1}{\sqrt{2}}$, and ℓ_ρ^δ is as given in Theorem 2.3.1.

Note that $\ell_\rho^{2\delta}$ appears as opposed to ℓ_ρ^δ in Proposition 2.6.2 due to an application of a a union bound in controlling both the upper and lower tail of S_n. In the case $d = 1$, Theorem 2.6.1 yields that, with probability at least $1-\delta$, $|S_n| \leq \frac{1}{1-\epsilon}\left[\sqrt{2\alpha(V_n \vee \rho)L_\rho^\delta(V_n)} + L_\rho(V_n)\right]$. This serves as a sanity check, showing that up to small constants, the univariate bound presented in Theorem 2.6.1 is equivalent to that in Proposition 2.6.2. While one may expect the bound from Proposition 2.6.2 to be tighter for large values of V_n (as discussed in Section 2.3.1), this multiplicative gap can be made arbitrarily small by appropriately selecting tuning parameters.

2.7 Proofs of Main Results

In this section, we provide the proofs of what we view as the primary two results of this paper: Theorem 2.3.1 and Theorem 2.4.1. We additionally prove Corollary 2.4.6, which, while not a

primary contribution of this work, has a proof that is similar in spirit to the other two results derived in this section. We start with the proof of Theorem 2.3.1, as the scalar bounds derived will play an integral role in the proof of Theorem 2.4.1

Proof of Theorem 2.3.1. First, observe that it suffices to show that, in the case $(S_n, V_n)_{n\geq 0}$ is sub-ψ and $V_n \geq 1, \forall n \geq 0$, we have

$$\mathbb{P}\left(\exists n \geq 0 : S_n \geq V_n \cdot (\psi^*)^{-1}\left(\frac{\alpha}{V_n}\ell_1(V_n)\right)\right) \leq \delta, \tag{2.7.1}$$

because, in the general case, we can consider the rescaled process $(S'_n, V'_n) := (S_n/\sqrt{\rho}, (V_n \vee \rho)/\rho)_{n\geq 0}$ and apply the concentration result from the case where $\rho = 1$. In more detail, clearly by construction $V'_n \geq 1$ for all $n \geq 1$, and by Proposition 2.2.3, we know $(S'_n, V'_n)_{n\geq 0}$ is sub-ψ_ρ, where we recall $\psi_\rho(\cdot) = \rho\psi(\cdot/\sqrt{\rho})$. Thus, noting that $(\psi^*_\rho)^{-1}(x) = \sqrt{\rho}(\psi^*)^{-1}(x/\rho)$ (Proposition 2.A.2), we have

$$\begin{aligned}
\delta &\geq \mathbb{P}\left(\exists n \geq 0 : S'_n \geq V'_n \cdot (\psi^*_\rho)^{-1}\left(\frac{\alpha}{V'_n}\ell_1(V'_n)\right)\right) \\
&= \mathbb{P}\left(\exists n \geq 0 : \frac{S_n}{\sqrt{\rho}} \geq \frac{V_n \vee \rho}{\rho} \cdot \left(\psi^*_\rho\right)^{-1}\left(\frac{\alpha\rho}{V_n \vee \rho}\ell_1\left(\frac{V_n \vee \rho}{\rho}\right)\right)\right) \\
&= \mathbb{P}\left(\exists n \geq 0 : \frac{S_n}{\sqrt{\rho}} \geq \frac{V_n \vee \rho}{\rho}\sqrt{\rho} \cdot (\psi^*)^{-1}\left(\frac{\alpha}{V_n \vee \rho}\ell_\rho(V_n)\right)\right) \\
&= \mathbb{P}\left(\exists n \geq 0 : S_n \geq (V_n \vee \rho) \cdot (\psi^*)^{-1}\left(\frac{\alpha}{V_n \vee \rho}\ell_\rho(V_n)\right)\right),
\end{aligned}$$

which demonstrates the claimed bound in the theorem statement. Thus, going forward, we just prove the bound presented in (2.7.1).

For $k \in \mathbb{N}$, define the "intercept and slope" pair (x_k, m_k) by

$$x_k := \alpha^k(\psi^*)^{-1}\left(\frac{\log(h(k)/\delta)}{\alpha^k}\right), \qquad m_k := \alpha^k,$$

and define $g_k : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ by

$$g_k(v) := x_k + \mathfrak{s}\left(\frac{x_k}{m_k}\right)(v - m_k),$$

where $\mathfrak{s}$ is the "slope transform" outlined in Appendix 2.A. Since we have assumed $\lim_{\lambda\uparrow\lambda_{\max}} \psi'(\lambda) = \infty$, we can apply Lemma 2.A.5 to obtain

$$\mathbb{P}\left(\exists n \geq 0 : S_n \geq g_k(V_n)\right) \leq \exp\left\{-m_k\psi^*\left(\frac{x_k}{m_k}\right)\right\} = \frac{\delta}{h(k)}.$$

Now, since $\mathfrak{s}(u) \leq u$ (Proposition 2.A.6), observe that for $\alpha^k \leq v < \alpha^{k+1}$, we have

$$\min_{j\in\mathbb{N}} g_j(v) \leq g_k(v) = x_k + \mathfrak{s}\left(\frac{x_k}{m_k}\right)(v - m_k)$$

$$
\begin{aligned}
&\leq x_k + \frac{x_k}{m_k}(v - m_k) = v\frac{x_k}{m_k} \\
&= v \cdot (\psi^*)^{-1}\left(\frac{\log(h(k)/\delta)}{\alpha^k}\right) \\
&\leq v \cdot (\psi^*)^{-1}\left(\frac{\alpha}{v}\log\left(\frac{h(\log_\alpha(v))}{\delta}\right)\right) \\
&= v \cdot (\psi^*)^{-1}\left(\frac{\alpha}{v}\ell_1(v)\right),
\end{aligned} \tag{2.7.2}
$$

where the third inequality comes from the fact $k \leq \log_\alpha(v)$, h is increasing, and $v \leq \alpha^{k+1}$. Now, observe that we have, by a union bound

$$
\begin{aligned}
\mathbb{P}\left(\exists n \geq 0 : S_n \geq V_n \cdot (\psi^*)^{-1}\left(\frac{\alpha\ell_1(V_n)}{V_n}\right)\right) &\leq \mathbb{P}\left(\exists n \geq 0 : S_n \geq \min_{k\in\mathbb{N}} g_k(V_n)\right) \\
&= \mathbb{P}\left(\bigcup_{k\in\mathbb{N}} \{\exists n \geq 0 : S_n \geq g_k(V_n)\}\right) \\
&\leq \sum_{k\in\mathbb{N}} \mathbb{P}(\exists n \geq 0 : S_n \geq g_k(V_n)) \\
&\leq \delta \sum_{k\in\mathbb{N}} h(k)^{-1} \leq \delta,
\end{aligned}
$$

completing the proof. ■

We now go about proving Theorem 2.4.1. Before proving the theorem, we state a simple geometric lemma that will be needed in proving our result. In short, the following lemma states that a certain change of variables on $\mathbb{S}^{d-1}$ does not increase the distance between points of a covering to a significant degree. We prove the following in Appendix 2.C.

Lemma 2.7.1. *Let K be a proper ϵ-cover of $\mathbb{S}^{d-1}$, and let $\pi : \mathbb{S}^{d-1} \to K$ be a projection mapping onto the cover K. Let T be a positive-definite matrix, and let $\kappa := \frac{\gamma_{\max}(T)}{\gamma_{\min}(T)}$ denote its condition number. Let $\pi_T : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ be defined as $\pi_T(\nu) := \frac{T^{1/2}\pi(\omega)}{\|T^{1/2}\pi(\omega)\|}$, where $\omega \in \mathbb{S}^{d-1}$ is the unique element satisfying*

$$
\nu = \frac{T^{1/2}\omega}{\|T^{1/2}\omega\|}. \tag{2.7.3}
$$

Then, for any $\nu \in \mathbb{S}^{d-1}$, we have

$$
\|\nu - \pi_T(\nu)\| \leq \sqrt{\kappa}\epsilon.
$$

With the above lemma we can now prove the main result of the paper.

Proof of Theorem 2.4.1. Observe that if $(S_n, V_n)_{n\geq 0}$ is a sub-ψ process (in the sense of Definition 2.2.2), then so is $(S_n, V_n \vee \rho I_d)$, so it suffices to assume $V_n \succeq \rho I_d$ going forward.

For $j \in \mathbb{N}$, let K_j be a fixed, minimal proper $\frac{\epsilon}{\beta^j}$-cover of the unit sphere $\mathbb{S}^{d-1}$, and let $N_j := N(\mathbb{S}^{d-1}, \epsilon/\beta^j, \|\cdot\|)$. Let $\ell_\rho^{(j)} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be the function ℓ_ρ defined in Theorem 2.3.1 with δ set to the value δ_j defined by

$$\delta_j := \left(\frac{1-\beta^{-1}}{\beta^j}\right)\delta/N_j.$$

That is, $\ell_\rho^{(j)}$ is the function given by

$$\begin{aligned}\ell_\rho^{(j)}(v) &:= \log\left(h\left(\log_\alpha\left(\frac{v\vee\rho}{\rho}\right)\right)\right) + \log\left(\frac{1}{\delta_j}\right) \\ &= \log\left(h\left(\log_\alpha\left(\frac{v\vee\rho}{\rho}\right)\right)\right) + \log\left(\frac{\beta^j}{\delta(1-\beta^{-1})}N_j\right).\end{aligned}$$

Now, since $(S_n, V_n)_{n\geq 0}$ is an $\mathbb{R}^d$-valued sub-ψ process, by Definition 2.2.2, we know that, for any fixed $\nu \in \mathbb{S}^{d-1}$, $(\langle\nu, S_n\rangle, \langle\nu, V_n\nu\rangle)_{n\geq 0}$ is sub-ψ in the scalar sense of Definition 2.2.1. Hence, by applying Theorem 2.3.1, for any fixed $\nu \in \mathbb{S}^{d-1}$, we have

$$\mathbb{P}\left(\exists n \geq 0 : \langle\nu, S_n\rangle \geq \langle\nu, V_n\nu\rangle \cdot (\psi^*)^{-1}\left(\frac{\alpha}{\langle\nu, V_n\nu\rangle}\ell_\rho^{(j)}(\langle\nu, V_n\nu\rangle)\right)\right) \leq \left(\frac{1-\beta^{-1}}{\beta^j}\right)\delta/N_j. \tag{2.7.4}$$

Noting that $\langle\nu, V_n\nu\rangle \leq \gamma_{\max}(V_n)$ for all $\nu \in \mathbb{S}^{d-1}$ and that $(\psi^*)^{-1}$ is an increasing function of its argument, we see that (2.7.4) still holds with $\ell_\rho^{(j)}(\langle\nu, V_n\nu\rangle)$ replaced by $\ell_\rho^{(j)}(\gamma_{\max}(V_n))$.

Now, for each $j \in \mathbb{N}$, define the "bad" event B_j as

$$B_j := \left\{\exists n \geq 0, \nu \in K_j : \langle\nu, S_n\rangle \geq \langle\nu, V_n\nu\rangle \cdot (\psi^*)^{-1}\left(\frac{\alpha}{\langle\nu, V_n\nu\rangle}\ell_\rho^{(j)}(\gamma_{\max}(V_n))\right)\right\}.$$

A straightforward union bound over the N_j elements of the cover K_j alongside (2.7.4) yields that $\mathbb{P}(B_j) \leq \frac{1-\beta^{-1}}{\beta^j}\delta$. Defining now the global "bad" event B as

$$B := \left\{\exists j \in \mathbb{N}, \exists \nu \in K_j, \exists n \geq 0 : \langle\nu, S_n\rangle \geq \langle\nu, V_n\nu\rangle \cdot (\psi^*)^{-1}\left(\frac{\alpha}{\langle\nu, V_n\nu\rangle}\ell_\rho^{(j)}(\gamma_{\max}(V_n))\rangle\right)\right\} = \bigcup_{j\in\mathbb{N}} B_j.$$

An additional straightforward union bound over indices $j \in \mathbb{N}$ yields

$$\mathbb{P}(B) = \mathbb{P}\left(\bigcup_{j\in\mathbb{N}} B_j\right) \leq \sum_{j\in\mathbb{N}} \mathbb{P}(B_j) \leq (1-\beta^{-1})\delta\sum_{j\in\mathbb{N}}\beta^{-j} = \delta.$$

Now, for $j \in \mathbb{N}$ and $n \geq 0$, let $\pi_n^{(j)} := \pi_{V_n}$ be the projection mapping from $\mathbb{S}^{d-1}$ onto the finite set $K_j(n) := \left\{V_n^{1/2}\nu/\|V_n^{1/2}\nu\| : \nu \in K_j\right\} \subset \mathbb{S}^{d-1}$, as in Lemma 2.7.1. Note that while $K_j(n)$ is a random subset of the unit sphere (through its dependence on the "accumulated variance" operator V_n at time n), the underlying $\frac{\epsilon}{\beta^j}$-cover K_j of $\mathbb{S}^{d-1}$ is fixed. Further, for $j \in \mathbb{N}$ and $n \geq 0$, define the event $E_j(n)$ by

$$E_j(n) := \left\{\beta^j \leq \sqrt{\kappa(V_n)} < \beta^{j+1}\right\}.$$

On the event $E_j(n)$, for any $j \in \mathbb{N}$ and $n \geq 0$, we have

$$\begin{aligned}
\left\|V_n^{-1/2} S_n\right\| &= \sup_{\omega \in \mathbb{S}^{d-1}} \left\langle \omega, V_n^{-1/2} S_n \right\rangle = \sup_{\omega \in \mathbb{S}^{d-1}} \left\{ \left\langle \omega - \pi_n^{(j+1)}(\omega), V_n^{-1/2} S_n \right\rangle + \left\langle \pi_n^{(j+1)}(\omega), V_n^{-1/2} S_n \right\rangle \right\} \\
&\leq \sup_{\omega \in \mathbb{S}^{d-1}} \left\| \omega - \pi_n^{(j+1)}(\omega) \right\| \cdot \left\| V_n^{-1/2} S_n \right\| + \sup_{\omega \in K_{j+1}(n)} \left\langle \omega, V_n^{-1/2} S_n \right\rangle \\
&\leq \frac{\epsilon}{\beta^{j+1}} \sqrt{\kappa(V_n)} \left\| V_n^{-1/2} S_n \right\| + \sup_{\nu \in K_{j+1}} \left\langle \frac{V_n^{1/2} \nu}{\|V_n^{1/2} \nu\|}, V_n^{-1/2} S_n \right\rangle \\
&\leq \epsilon \left\| V_n^{-1/2} S_n \right\| + \sup_{\nu \in K_{j+1}} \frac{\langle \nu, S_n \rangle}{\sqrt{\langle \nu, V_n \nu \rangle}}.
\end{aligned}$$

In the above, the first equality comes from the variational representation of the norm $\|\cdot\|$ and the second equality comes from adding and subtracting $\left\langle \pi_n^{(j+1)}(\omega), V_n^{-1/2} S_n \right\rangle$. Further, the first inequality comes from splitting the supremum and applying Cauchy-Schwarz to the first term, the second inequality comes from applying Lemma 2.7.1 to $\left\| \omega - \pi_n^{(j+1)}(\omega) \right\|$ and applying the definition of $K_{j+1}(n)$, and the final inequality comes from simplifying the second term and from observing that, on the event $E_j(n)$, $\sqrt{\kappa(V_n)} < \beta^{j+1}$.

Further, observe that, on the event $E_j(n)$, we have the inequality

$$\begin{aligned}
\ell_\rho^{(j+1)}(\gamma_{\max}(V_n)) &= \log\left(h\left(\log_\alpha\left(\frac{\gamma_{\max}(V_n) \vee \rho}{\rho} \right) \right) \right) + \log\left(\frac{1}{\delta(1-\beta^{-1})} \right) + \log\left(N_j \beta^{j+1} \right) \\
&\leq \log\left(h\left(\log_\alpha\left(\frac{\gamma_{\max}(V_n) \vee \rho}{\rho} \right) \right) \right) + \log\left(\frac{1}{\delta(1-\beta^{-1})} \right) \\
&\quad + \log\left(\beta \sqrt{\kappa(V_n)} N_{d-1}\left(\frac{\epsilon}{\beta \sqrt{\kappa(V_n)}} \right) \right) \\
&= L_\rho(V_n).
\end{aligned}$$

In the above, the inequality follows from observing that $\beta^j \leq \sqrt{\kappa(V_n)}$. From this, rearranging, we see that, for any $j \in \mathbb{N}$ and $n \geq 0$, on the event $E_j(n) \cap B^c$ we have

$$\begin{aligned}
\left\| V_n^{-1/2} S_n \right\| &\leq \frac{1}{1-\epsilon} \sup_{\nu \in K_{j+1}} \frac{\langle \nu, S_n \rangle}{\sqrt{\langle \nu, V_n \nu \rangle}} \\
&\leq \frac{1}{1-\epsilon} \sup_{\nu \in K_{j+1}} \sqrt{\langle \nu, V_n \nu \rangle} \cdot (\psi^*)^{-1}\left(\frac{\alpha}{\langle \nu, V_n \nu \rangle} \ell_\rho(\gamma_{\max}(V_n)) \right) \\
&\leq \frac{1}{1-\epsilon} \sup_{\nu \in K_{j+1}} \sqrt{\langle \nu, V_n \nu \rangle} \cdot (\psi^*)^{-1}\left(\frac{\alpha}{\langle \nu, V_n \nu \rangle} L_\rho(V_n) \right) \\
&\leq \frac{1}{1-\epsilon} \sup_{\nu \in \mathbb{S}^{d-1}} \sqrt{\langle \nu, V_n \nu \rangle} \cdot (\psi^*)^{-1}\left(\frac{\alpha}{\langle \nu, V_n \nu \rangle} L_\rho(V_n) \right).
\end{aligned}$$

If $(j_n)_{n\in\mathbb{N}}$ is any sequence of natural numbers, and we define $G^{(j_n)} := \bigcap_{n\geq 0}\{E_{j_n}(n)\cap B^c\}$, it is clear that, on the event $G^{(j_n)}$, the inequality

$$\left\|V_n^{-1/2}S_n\right\| \leq \frac{1}{1-\epsilon}\sup_{\nu\in\mathbb{S}^{d-1}}\sqrt{\langle\nu, V_n\nu\rangle}\cdot(\psi^*)^{-1}\left(\frac{\alpha}{\langle\nu, V_n\nu\rangle}L_\rho(V_n)\right) \tag{2.7.5}$$

holds simultaneously for all $n \geq 0$. Noting that we have the identity $B^c = \biguplus_{(j_n)_{n\in\mathbb{N}}} G^{(j_n)}$ yields that (2.7.5) actually holds simultaneously for all $n \geq 0$ on the event B^c. What we have done in the above is break the "good" event B^c into geometric buckets based on the condition number at each time, and then noted that the regardless of the realized sequence of condition numbers $(\kappa(V_n))_{n\geq 0}$, the target inequality holds.

This proves the claim for arbitrary CGF-like functions ψ, which is presented following Theorem 2.4.1. Now, if we further assume ψ is super-Gaussian, on the event B^c defined above, we have

$$\begin{aligned}\left\|V_n^{-1/2}S_n\right\| &\leq \frac{1}{1-\epsilon}\sup_{\nu\in\mathbb{S}^{d-1}}\sqrt{\langle\nu, V_n\nu\rangle}\cdot(\psi^*)^{-1}\left(\frac{\alpha}{\langle\nu, V_n\nu\rangle}L_\rho(V_n)\right)\\ &= \frac{1}{1-\epsilon}\sup_{x\in[\gamma_{\min}(V_n),\gamma_{\max}(V_n)]}\sqrt{x}\cdot(\psi^*)^{-1}\left(\frac{\alpha}{x}L_\rho(V_n)\right).\end{aligned}$$

Now, by Lemma 2.A.3, we know the assumption that ψ is a super-Gaussian CGF-like function implies that ψ^* is a sub-Gaussian CGF-like function. Moreover by the same proposition, we see that $\psi^*(C\cdot)$ is a sub-Gaussian CGF-like function for any positive $C > 0$. Consequently, by Proposition 2.A.3, we see that $(\psi^*)^{-1}(Cu)/\sqrt{u}$ is an increasing function of u, and thus by making the change of variable $x := \frac{1}{u}$, that $\sqrt{x}(\psi^*)^{-1}\left(\frac{C}{x}\right)$ a decreasing function of x. Thus, we have that, on the event B^c (which, we recall, occurs with probability at least $1-\delta$)

$$\begin{aligned}\left\|V_n^{-1/2}S_n\right\| &\leq \frac{1}{1-\epsilon}\sup_{x\in[\gamma_{\min}(V_n),\gamma_{\max}(V_n)]}\sqrt{x}\cdot(\psi^*)^{-1}\left(\frac{\alpha}{x}L_\rho(V_n)\right)\\ &\leq \frac{\sqrt{\gamma_{\min}(V_n)}}{1-\epsilon}\cdot(\psi^*)^{-1}\left(\frac{\alpha}{\gamma_{\min}(V_n)}L_\rho(V_n)\right)\end{aligned}$$

simultaneously for all $n \geq 0$, proving the desired result. A symmetric argument holds in the case that the CGF-like function ψ is instead sub-Gaussian, with $\gamma_{\max}(V_n)$ replacing $\gamma_{\min}(V_n)$ in the final inequality.

■

Lastly, we prove Corollary 2.4.6, which in turn can be used to derive Corollary 2.3.2. While we do not consider this corollary a primary contribution of our work, we include the proof in this section due to its closeness (in spirit) to the previous two proofs.

Proof of Corollary 2.4.6. Recalling that $(\psi_N^*)^{-1}(u) = \sqrt{2u}$, the assumption that $\psi(\lambda) \sim \frac{\lambda^2}{2}$ implies there, for any $\eta > 0$, there exists an $\overline{u} \in \mathbb{R}_{>0}$ such that

$$(\psi^*)^{-1}(u) \leq (1+\eta)(\psi_N^*)^{-1}(u) = (1+\eta)\sqrt{2u}$$

for all $u \in [0, \overline{u}]$. Let $(\alpha_t)_{t\geq 1}, (\iota_t)_{t\geq 1}, (\epsilon_t)_{t\geq 1}$, and $(\beta_t)_{t\geq 1}$ be such that $\alpha_t, \iota_t, \beta_t \downarrow 1$ and $\epsilon_t \downarrow 0$ monotonically and let $(\delta_t)_{t\geq 1}$ be such that (a) $\delta_t \downarrow 0$ monotonically and (b) $\sum_{t=1}^{\infty} \delta_t < \infty$. Define the sequence of functions $h_t : \mathbb{N} \to \mathbb{R}_{\geq 0}$ by $h_t(s) := \zeta(\iota_t)(1+s)^{\iota_t}$. Let N_t be the (almost surely finite) random time given by

$$N_t := \inf \Big\{ n \geq 0 : \frac{\alpha}{\gamma_{\min}(V_{n'})} L_1(V_{n'}) \leq \overline{u} \ \ \forall n' \geq n, \ \text{ and } \\ \log\left(\frac{C_d \zeta(\iota_t)}{\delta_t (\log(\alpha_t))^{\iota_t}(1-\beta_t^{-1})}\right) + d \log\left(\frac{3\beta_t}{\epsilon_t}\right) \leq \frac{\iota_t}{t} \log\log(\gamma_{\max}(V_n)) \Big\}.$$

Theorem 2.4.1 instantiated with the covering number bound in Lemma 2.4.2 implies that, with probability at least $1 - \delta_t$, simultaneously for all $n \geq N_t$, we have

$$\begin{aligned} \left\| V_n^{-1/2} S_n \right\| &\leq \frac{\sqrt{\gamma_{\min}(V_n)}}{1-\epsilon_t} \cdot (\psi^*)^{-1}\left(\frac{\alpha_t}{\gamma_{\min}(V_n)} L_1(V_n)\right) \\ &\leq \frac{1+\eta}{1-\epsilon_t} \sqrt{2\alpha_t \left[\iota_t \log\log(V_n) + \log\left(\frac{C_d \zeta(\iota_t)}{\delta_t (\log(\alpha_t))^{\iota_t}(1-\beta_t^{-1})}\right) + d \log\left(\frac{3\beta_t \sqrt{\kappa(V_n)}}{\epsilon_t}\right)\right]} \\ &\leq \frac{1+\eta}{1-\epsilon_t} \sqrt{2\alpha_t \iota_t \left(1 + \frac{1}{t}\right) \log\log(\gamma_{\max}(V_n)) + \alpha_t d \log \kappa(V_n)}. \end{aligned}$$

Thus, for $t \geq 1$, define the event A_n by

$$A_t = \left\{ \exists n \geq N_t : \left\| V_n^{-1/2} S_n \right\| \geq \frac{1+\eta}{1-\epsilon_t} \sqrt{2\alpha_t \iota_t \left(1 + \frac{1}{t}\right) \log\log(V_n) + \alpha_t d \log \kappa(V_n)} \right\},$$

and observe that by the above argument $\mathbb{P}(A_t) \leq \delta_t$. Note that, for arbitrary $\gamma > 1$, we have

$$A_\gamma := \left\{ \|V_n^{-1/2} S_n\| > (1+\eta)\sqrt{\gamma \left[2\log\log(V_n) + d \log \kappa(V_n)\right]} \ \text{ i.o.} \right\} \subset \limsup_{t\to\infty} A_t := \bigcap_{t\geq 1} \bigcup_{k \geq t} A_k,$$

where i.o. denotes an event occurring infinitely often. By the first Borel-Cantelli lemma (see Durrett [50], Chapter 2) we have

$$\mathbb{P}(A_\gamma) \leq \mathbb{P}\left(\bigcap_{t\geq 1} \bigcup_{k\geq t} A_k\right) = 0,$$

since $\sum_{t=1}^{\infty} \mathbb{P}(A_t) \leq \sum_{t=1}^{\infty} \delta_t < \infty$. Thus, with probability 1, we have

$$\limsup_{n\to\infty} \frac{S_n}{(1+\eta)\sqrt{\gamma \left[2\log\log(V_n) + d \log \kappa(V_n)\right]}} \leq 1,$$

but since $\eta > 0$ and $\gamma > 1$ where arbitrary, the result follows.

■

2.8 Conclusion and Discussion

Self-normalized quantities arise naturally in a variety of high-dimensional statistical tasks, with online learning [4, 2, 165, 30, 33], time series analysis [14, 137], and hypothesis testing [142, 143, 130, 161] being several notable examples. Despite their crucial role in common statistical tasks, very little has been explored in terms of self-normalized concentration outside of the sub-Gaussian setting. In this paper, we present a time-uniform, self-normalized concentration for sub-ψ processes, i.e. processes whose increments, roughly, have cumulant generating function bounded by ψ. Our results are closed form, have small constants, and have parameters that can be fine-tuned for a statistician's desired application. Moreover, with our bounds, we can establish an asymptotic law of the iterated logarithm for vector-valued processes that recovers the law of iterated logarithm for scalar sub-ψ processes first established by Howard et al. [75].

Along with our primary result on the self-normalized concentration of vector-valued processes, we make variety of additional contributions. En route to proving Theorem 2.4.1, we prove a non-asymptotic law of the iterated for sub-ψ processes, generalizing the results of Howard et al. [75] beyond just the sub-Gamma setting. Likewise, we demonstrate how to leverage our self-normalized inequalities in several practical statistical settings. In particular, we derive non-asymptotically valid confidence ellipsoids for online linear regression, describe how to construct confidence sets for vector autoregressive models, and prove a multivariate empirical Bernstein inequality. There are undoubtedly many more settings in which our bounds can be applied, and we leave the exploration of these applications for interesting future work.

While the results presented in this paper are quite general, there are still many interesting questions about self-normalized concentration to be answered. As a first example, existing results on the self-normalized concentration of sub-Gaussian random vectors yield a bound that is proportional to $O\left(\sqrt{\log\det(V_n)}\right)$ [41, 42, 3]. This is in contrast to the results discussed in this work, which provide bounds of the form $O\left(\sqrt{\log\log\gamma_{\max}(V_n) + d\log\kappa(V_n)}\right)$. As discussed in Section 2.4, neither form of bound uniformly dominates the other. In particular, when V_n is well-conditioned, our concentration results may be preferable, but for poorly-conditioned V_n, determinant rate bounds may be desirable. A major open question is whether determinant rate bounds can be obtained for general sub-ψ processes, or if the determinant rate is just attainable in the sub-Gaussian setting. The techniques discussed in this paper do not seem directly applicable to this setting, and so we thus leave obtaining determinant rate bounds as compelling future work.

This work demonstrates that simple, closed-form bounds on self-normalized processes can be established under very general distributional assumptions. While existing works consider a setting in which the increments of processes are sub-Gaussian, concentration of measure should not be viewed as a "one size fits all" phenomenon. For instance, the noise observed in taking real-world may not be sub-Gaussian, but rather perhaps sub-Exponential, sub-Gamma, or even heavy tailed. Overall, our bounds provide a means by which the statistician can properly calibrate confidence in these more delicate settings.

2.A Properties of CGF-like Functions

The *cumulant generating function* (or *CGF*) of a random variable plays an integral role in understanding concentration of measure phenomena, such as through the classical Chernoff style of argument [74, 18]. Suppose X is a random variable such that $\mathbb{E}X = 0$, $\mathbb{E}e^{\lambda X} < \infty$ for all $\lambda \in [0, \lambda_{\max})$, and $\lim_{\lambda \uparrow \lambda_{\max}} \mathbb{E}e^{\lambda X} = \infty$. The cumulant generating function of X, which can be thought of as "compressing" all of the moments of X into a single function, is the mapping $\psi : [0, \lambda_{\max}) \to \mathbb{R}_{\geq 0}$ given by $\psi(\lambda) := \log \mathbb{E}e^{\lambda X}$.

In this appendix, we study properties of *cumulant generating function-like* (or *CGF-like*) functions, which are functions that may not be the CGF of any random variable, but display similar analytic properties to CGFs. If $\psi : [0, \lambda_{\max}) \to \mathbb{R}_{\geq 0}$ is the CGF of a random variable, straightforward calculation yields that $\psi(0) = \psi'(0) = 0$, $\psi''(\lambda) > 0$, and ψ is strictly convex. As such, we say a twice continuously differentiable function $\psi : [0, \lambda_{\max}) \to \mathbb{R}_{\geq 0}$ is CGF-like if it obeys these aforementioned properties. We study various properties of CGF-like functions in the sequel, as these properties form the foundation of our results studying the self-normalized concentration of sub-ψ processes.

Proposition 2.A.1. *Suppose $\psi : [0, \lambda_{\max}) \to \mathbb{R}_{\geq 0}$ is CGF-like. Then convex conjugate $\psi^* : [0, u_{\max}) \to \mathbb{R}_{\geq 0}$ defined by $\psi^*(u) := \sup_{\lambda \in [0, \lambda_{\max})} \lambda u - \psi(u)$, is also CGF-like, where $u_{\max} := \sup_{\lambda \in [0, \lambda_{\max})} \psi'(\lambda)$.*

Proof. Clearly ψ^* is convex and twice continuously-differentiable. Next, observe that

$$\psi^*(0) = \sup_{\lambda \in [0, \lambda_{\max})} \{-\psi(\lambda)\} = \psi(0) = 0.$$

Further, using the fact that $(\psi^*)' = (\psi')^{-1}$, we have that

$$(\psi^*)'(0) = (\psi')^{-1}(0) = 0.$$

Lastly, we have that

$$(\psi^*)''(0) = ((\psi')^{-1})'(0) = \frac{1}{\psi''((\psi')^{-1}(0))} = \frac{1}{\psi''(0)} > 0.$$

Thus, ψ^* is also CGF-like. ■

If $\psi : [0, \lambda_{\max}) \to \mathbb{R}_{\geq 0}$ is CGF-like, then for any $\rho > 0$, the "rescaled" function $\psi_\rho : [0, \sqrt{\rho}\lambda_{\max}) \to \mathbb{R}_{\geq 0}$ given by $\psi_\rho(\lambda) := \rho\psi(\lambda/\sqrt{\rho})$ is also CGF-like. These rescaled CGF-like functions arise naturally in studying processes that have been re-normalized to have $V_n \succeq \mathrm{id}_H$ for all $n \geq 0$. These rescaled functions ψ_ρ exhibit the following properties.

Proposition 2.A.2. *Let $\psi : [0, \lambda_{\max}) \to \mathbb{R}_{\geq 0}$ be CGF-like, and let $\psi_\rho : [0, \sqrt{\rho}\lambda_{\max}) \to \mathbb{R}_{\geq 0}$ be as above. The following hold.*

1. *ψ_ρ is a CGF-like function.*
2. *$\psi_\rho^*(u) = \rho\psi^*(u/\sqrt{\rho})$.*
3. *$(\psi_\rho^*)^{-1}(x) = \sqrt{\rho}(\psi^*)^{-1}\left(\frac{x}{\rho}\right)$.*

Proof. The validity of the first claim follows immediately by the definition of a CGF-like function.

To see the validity of the second claim, note that

$$\psi_\rho^*(u) = \sup_{\lambda \in [0, \sqrt{\rho}\lambda_{\max})} \left\{ u\lambda - \rho\psi\left(\frac{\lambda}{\sqrt{\rho}}\right) \right\}.$$

Differentiating the inner expression on the right-hand side and setting equal to zero furnishes that the supremum is obtained at $\lambda = \sqrt{\rho}(\psi')^{-1}(u/\sqrt{\rho})$. Plugging this back into the above expression yields

$$\begin{aligned}
\psi_\rho^*(u) &= \sqrt{\rho}u(\psi')^{-1}\left(\frac{u}{\sqrt{\rho}}\right) - \rho\psi\left((\psi')^{-1}\left(\frac{u}{\sqrt{\rho}}\right)\right) \\
&= \rho\left[\frac{u}{\sqrt{\rho}}(\psi')^{-1}\left(\frac{u}{\sqrt{\rho}}\right) - \psi\left((\psi')^{-1}\left(\frac{u}{\sqrt{\rho}}\right)\right)\right] \\
&= \rho\psi^*\left(\frac{u}{\sqrt{\rho}}\right),
\end{aligned}$$

which proves the second item.

Laslty, the third item can be readily checked as

$$\psi_\rho^*\left(\sqrt{\rho}(\psi^*)^{-1}\left(\frac{x}{\rho}\right)\right) = \rho(\psi^*)\left(\frac{\sqrt{\rho}}{\sqrt{\rho}}(\psi^*)^{-1}\left(\frac{x}{\rho}\right)\right) = \rho\frac{x}{\rho} = x.$$

Applying $(\psi_\rho^*)^{-1}$ to both sides thus yields the desired result.

■

Throughout our work, we are especially interested in studying sub-ψ processes whose increments exhibit tail behavior that is either "heavier" or "lighter" than that of a Gaussian random-variable. More concretely, we study processes where ψ is a *super-Gaussian* (respectively *sub-Gaussian*) CGF-like function, i.e. a CGF-like function where $\frac{\psi(\lambda)}{\lambda^2}$ is a non-decreasing (respectively non-increasing) function of λ. In words, a CGF-like function ψ is super-Gaussian (or sub-Gaussian) if it increases more rapidly (less rapdily) than the CGF of a standard normal random variable. We focus on super-Gaussian CGF-like functions in the sequel, but exactly analogous results hold for sub-Gaussian CGF-like functions. Super-Gaussian CGF-like functions enjoy a number of convenient properties and equivalent definitions, which we enumerate below.

Proposition 2.A.3. *Suppose $\psi : [0, \lambda_{\max}) \to \mathbb{R}_{\geq 0}$ is a CGF-like function. The following hold.*

1. *ψ is super-Gaussian if and only if $\psi'(\lambda) \geq \frac{2\psi(\lambda)}{\lambda}$.*
2. *If ψ is super-Gaussian, then so is $\varphi := a\psi(b\cdot) : [0, \lambda_{\max}/b) \to \mathbb{R}_{\geq 0}$ for any $a, b > 0$.*
3. *If ψ is super-Gaussian, then $\frac{\psi^{-1}(x)}{\sqrt{x}}$ is a decreasing function of $x \in [0, \infty)$.*
4. *ψ is super-Gaussian if and only if its convex conjugate ψ^* is sub-Gaussian.*

Proof. 1. Differentiating via the product rule yields

$$\left(\frac{\psi(\lambda)}{\lambda^2}\right)' = \frac{\psi'(\lambda)}{\lambda^2} - \frac{2\psi(\lambda)}{\lambda^3}.$$

Consequently, we have

$$\left(\frac{\psi(\lambda)}{\lambda^2}\right)' \geq 0 \quad \Leftrightarrow \quad \psi'(\lambda) \geq \frac{2\psi(\lambda)}{\lambda},$$

proving the desired result.

2. This result follows from the equivalent condition presented in the first part of the proposition. In particular, observe that we have

$$\varphi'(\lambda) = ab\psi'(b\lambda) \geq 2ab\frac{\psi(b\lambda)}{b\lambda} = \frac{2\varphi(\lambda)}{\lambda},$$

proving the desired result.

3. Straightforward calculus yields

$$\left(\frac{\psi^{-1}(x)}{\sqrt{x}}\right)' = \frac{(\psi^{-1})'(x)}{\sqrt{x}} - \frac{1}{2}\frac{\psi^{-1}(x)}{x^{3/2}} = \frac{1}{\sqrt{x}\psi'(\psi^{-1}(x))} - \frac{1}{2}\frac{\psi^{-1}(x)}{x^{3/2}}.$$

Next, the assumption of ψ being super-Gaussian yields

$$\psi'(\psi^{-1}(x)) \geq \frac{2\psi(\psi^{-1}(x))}{\psi^{-1}(x)} = \frac{2x}{\psi^{-1}(x)}.$$

Combining these two panels furnishes

$$\left(\frac{\psi^{-1}(x)}{\sqrt{x}}\right)' = \frac{1}{\sqrt{x}\psi'(\psi^{-1}(x))} - \frac{1}{2}\frac{\psi^{-1}(x)}{x^{3/2}} \leq \frac{1}{2}\frac{\psi^{-1}(x)}{x^{3/2}} - \frac{1}{2}\frac{\psi^{-1}(x)}{x^{3/2}} = 0,$$

which is what we wanted.

4. We prove the forward direction as the proof of the reverse direction is exactly analogous. Recall that the super-Gaussianity of ψ implies that for all $\lambda \in [0, \lambda_{\max})$, we have $\psi'(\lambda) \geq \frac{2\psi(\lambda)}{\lambda}$. In particular, taking $\lambda = (\psi^*)'(u)$ for $u \in [0, u_{\max})$ for $u_{\max} := \sup_\lambda \psi'(\lambda)$ yields:

$$u = \psi'((\psi')^{-1}(u)) = \psi'((\psi^*)'(u)) \geq \frac{2\psi((\psi^*)'(u))}{(\psi^*)'(u)}.$$

Rearranging and noting that $\psi = (\psi^*)^*$ yields

$$\begin{aligned}(\psi^*)'(u) &\geq \frac{2\psi((\psi^*)'(u))}{u} = \frac{2\sup_{w\in[0,u_{\max})}\{w(\psi^*)'(u) - \psi^*(w)\}}{u} \\ &\geq \frac{2\{u(\psi^*)'(u) - (\psi^*)(u)\}}{u} = 2(\psi^*)'(u) - \frac{2\psi^*(u)}{u}.\end{aligned}$$

Now, subtracting $2(\psi^*)'(u)$ from both sides yields

$$-(\psi^*)'(u) \geq -\frac{2\psi^*(u)}{u}.$$

Multiplying both sides by -1 furnishes the desired result.

■

We conclude this section by discussing the slope transform, a recently proposed transform of a CGF-like function that can be used to construct time-uniform, line-crossing inequalities for martingales [74].

Definition 2.A.4. Suppose $\psi : [0, \lambda_{\max}) \to \mathbb{R}_{\geq 0}$ is a CGF-like function. The **slope transform** associcated with ψ is the mapping $\mathfrak{s} : [0, u_{\max}) \to \mathbb{R}_{\geq 0}$ given by

$$\mathfrak{s}(u) := \frac{\psi\left((\psi^*)'(u)\right)}{(\psi^*)'(u)}.$$

The slope transform, while abstract and perhaps a bit unintuitive in nature, is of great utility in optimizing our time-uniform, scalar-valued inequalities in the main body of this paper. In particular, we will leverage the following inequality in the proof of Theorem 2.3.1. In the following, recall that for a fixed CGF-like function $\psi : [0, \lambda_{\max}) \to \mathbb{R}_{\geq 0}$, we defined the quantity $u_{\max}$ as $u_{\max} := \sup_\lambda \psi'(\lambda)$. For most examples considered in this paper (in particular in the case of super-Gaussian ψ), $u_{\max} = \infty$.

Lemma 2.A.5 (Howard et al. [74])**.** *Suppose $\psi : [0, \lambda_{\max}) \to \mathbb{R}_{\geq 0}$ is CGF-like, and suppose $(S_n, V_n)_{n\geq 0}$ is a sub-ψ process, per Definition 2.2.1. Then, for any $m > 0$, $\delta \in (0, 1)$, and any $x \in (0, mu_{\max})$, we have*

$$\mathbb{P}\left(\exists n \geq 0 : S_n \geq x + \mathfrak{s}\left(\frac{x}{m}\right)(V_n - m)\right) \leq \exp\left\{-m\psi^*\left(\frac{x}{m}\right)\right\}.$$

While the slope transform $\mathfrak{s}(u)$ may be a generally complicated function, the following upper bound allows us to greatly simplify our analysis. It is proven in Howard et al. [74].

Proposition 2.A.6. *Suppose $\psi : [0, \lambda_{\max}) \to \mathbb{R}_{\geq 0}$ is CGF-like. Let $\mathfrak{s} : [0, u_{\max}) \to \mathbb{R}_{\geq 0}$ be the associated slope transform. Then, for any $u \in [0, u_{\max})$, $\mathfrak{s}(u) \leq u$.*

2.B Proofs of Results from Sections 2.5 and 2.6

In this appendix, we provide proofs for all results related to applications of Theorem 2.4.1. We start by proving the regression-based results from Section 2.5, and then move on to proving our empirical Bernstein bound, as discussed in Section 2.6. We begin by providing practically-relevant examples of when the residual process $S_n = \sum_{m=1}^n \epsilon_m X_m$ defined in Model 2.5.6 is sub-ψ with variance proxy $V_n = \sum_{m=1}^n X_m X_m^\top$.

Proposition 2.B.1. *Suppose $(X_n)_{n\geq 1}$, $(\epsilon_n)_{n\geq 1}$, and $(\mathcal{F}_n)_{n\geq 0}$ are as outlined in Model 2.5.1. Let us define the residual process $(S_n)_{n\geq 0}$ by $S_n := \sum_{m=1}^n \epsilon_m X_m$ and the covariance process $(V_n)_{n\geq 0}$ by $V_n := \sum_{m=1}^n X_m X_m^\top$. Then, $(S_n)_{n\geq 0}$ is sub-ψ with variance proxy $(V_n)_{n\geq 0}$ if either of the following conditions is satisfied.*

1. $(\epsilon_n)_{n\geq 1}$ *satisfies* $\log \mathbb{E}_{n-1} \exp\{\lambda \epsilon_n\} \leq \psi_N(\lambda)$ *for all* $n \geq 1$ *and* $\lambda \geq 0$.
2. $\|X_n\| \leq 1$ *almost surely for all* $n \geq 1$ *and* $(\epsilon_n)_{n\geq 1}$ *satisfies* $\log \mathbb{E}_{n-1} \exp\{\lambda \epsilon_n\} \leq \psi(\lambda)$ *for all* $n \geq 1$ *and* $\lambda \in [0, \lambda_{\max})$*, where* $\psi : [0, \lambda_{\max}) \to \mathbb{R}_{\geq 0}$ *is some super-Gaussian CGF-like function.*

Proof. The proof of 1 is straightforward, so we just prove 2. Observe that, from the assumption that $\psi : [0, \lambda_{\max}) \to \mathbb{R}_{\geq 0}$ is a super-Gaussian CGF-like function, we have that, for any $\lambda_1 < \lambda_2 \in [0, \lambda_{\max})$,

$$\frac{\psi(\lambda_1)}{\lambda_1^2} \leq \frac{\psi(\lambda_2)}{\lambda_2^2}.$$

Consequently, for any direction $\nu \in \mathbb{S}^{d-1}$, $\lambda \in [0, \lambda_{\max})$, and $\|x\| \leq 1$, we have

$$\frac{\psi(\lambda \langle \nu, x \rangle)}{\lambda^2 \langle \nu, x \rangle^2} \leq \frac{\psi(\lambda)}{\lambda^2}.$$

Combining this with with the CGF bound on the noise variable ϵ_n presented in Proposition 2.B.1 (along with the assumption that $\|X_n\| \leq 1$), we have

$$\log \mathbb{E}\left(e^{\lambda \langle \nu, X_n \rangle \epsilon_n} \mid \mathcal{F}_{n-1}\right) \leq \psi(\lambda \langle \nu, X_n \rangle) \leq \langle \nu, X_n \rangle^2 \psi(\lambda),$$

where in the above we have used the fact that X_n is $\mathcal{F}_{n-1}$-measurable. This immediately yields that, for any $\lambda \in [0, \lambda_{\max})$ and $\nu \in \mathbb{S}^{d-1}$, the process $(M_n^{\lambda,\nu})_{n\geq 0}$ given by

$$M_n^{\lambda,\nu} := \exp\left\{\lambda \sum_{m \leq n} \epsilon_m \langle \nu, X_m \rangle - \psi(\lambda) \sum_{m \leq n} \langle \nu, X_m \rangle^2\right\} = \exp\left\{\lambda \langle \nu, S_n \rangle - \psi(\lambda) \langle \nu, V_n \nu \rangle\right\}$$

is a non-negative supermartingale. Consequently, the scalar-valued process $(\langle \nu, S_n \rangle, \langle \nu, V_n \nu \rangle)_{n\geq 0}$ is sub-ψ for any $\nu \in \mathbb{S}^{d-1}$. Thus, by definition, the vector process $(S_n, V_n)_{n\geq 0}$ is sub-ψ in the vector-valued sense provided in Definition 2.2.2.

■

We now prove Theorem 2.5.2.

Proof of Theorem 2.5.2. For a Hermitian matrix $A \in \mathbb{R}^{d\times d}$ let $A \wedge \rho I_d$ be defined equivalently to $A \vee \rho I_d$ except with the eigenvalue being set to $\gamma_i(A) \wedge \rho$ versus $\gamma_i(A) \vee \rho$. Observe that we have the identity

$$A = A \vee \rho I_d + A \wedge \rho I_d - \rho I_d. \tag{2.B.1}$$

Note that we can write the difference between our estimate and the true slope parameter as

$$\begin{aligned}
\widehat{\theta}_n - \theta^* &= (V_n \vee \rho I_d)^{-1} \mathbf{X}_n^\top (\mathbf{X}_n \theta^* + \epsilon_{1:t}) - \theta^* \\
&= (V_n \vee \rho I_d)^{-1} (\mathbf{X}_n^\top \mathbf{X}_n \vee \rho I_d + \mathbf{X}_n^\top \mathbf{X}_n \wedge \rho I_d - \rho I_d) \theta^* + (V_n \vee \rho I_d)^{-1} S_n - \theta^* \\
&= (V_n \vee \rho I_d)^{-1} \left(\mathbf{X}_n^\top \mathbf{X}_n \wedge \rho I_d - \rho I_d\right) \theta^* + (V_n \vee \rho I_d)^{-1} S_n,
\end{aligned}$$

where in the above we have defined the "residual process" $(S_n)_{n\geq 0}$ as $S_n := \sum_{m=1}^n \epsilon_m X_m \in \mathbb{R}^d$. In the above, the second line follows from the first by applying the equality outlined in

Equation (2.B.1), the third follows from the second by recalling $V_n = \mathbf{X}_n^\top \mathbf{X}_n$ and noting a cancellation between the first and last term.

Thus, applying the triangle inequality gives us

$$\begin{aligned}\|(V_n \vee \rho I_d)^{1/2}(\widehat{\theta}_n - \theta^*)\| &\leq \|(V_n \vee \rho I_d)^{-1/2}(\mathbf{X}_n^\top \mathbf{X}_n \wedge \rho I_d - \rho I_d)\theta^*\| + \|(V_n \vee \rho I_d)^{-1/2} S_n\| \\ &\leq \sqrt{\rho}\|\theta^*\| \mathbb{1}_{\gamma_{\min}(\mathbf{X}_n^\top \mathbf{X}_n) < \rho} + \|(V_n \vee \rho I_d)^{-1/2} S_n\|,\end{aligned}$$

where the second line follows from the first via straightforward algebraic manipulation and bounding. What remains is to bound $\|(V_n \vee \rho I_d)^{-1/2} S_n\|$. But since we have assumed $(S_n)_{n \geq 0}$ is sub-ψ with variance proxy $(V_n)_{n \geq 0}$, Theorem 2.4.1 implies that, with probability at least $1 - \delta$, simultaneously for all $n \geq 1$, we have

$$\|(V_n \vee \rho I_d)^{-1/2} S_n\| \leq \frac{\sqrt{\gamma_{\min}(V_n \vee \rho I_d)}}{1 - \epsilon} \cdot (\psi^*)^{-1}\left(\frac{\alpha}{\gamma_{\min}(V_n \vee \rho I_d)} L_\rho(V_n)\right),$$

which finishes the proof. ■

We now prove Corollary 2.5.4. The proof below is almost identical to the proof of Theorem 2.5.2, modulo slight modifications, so omit many details.

Proof of Corollary 2.5.4. Using a similar line of reasoning, we see that we have the (deterministic) inequality

$$\begin{aligned}\|(V_n + \rho I_d)^{1/2}(\widetilde{\theta}_n - \theta^*)\| &\leq \rho\|(V_n + \rho I_d)^{-1/2}\theta^*\| + \|(V_n + \rho I_d)^{-1/2} S_n\| \\ &\leq \sqrt{\rho}\|\theta^*\| + \|(V_n + \rho I_d)^{-1/2} S_n\|,\end{aligned}$$

where $(S_n)_{n \geq 0}$ is the residual process outlined in the proof of Theorem 2.5.2. The result now follows by noting that $(S_n, V_n)_{n \geq 0}$ is sub-ψ in the vector-sense of Definition 2.2.2. ■

What remains is to prove Corollary 2.5.7, which concerns the estimation of model parameters in the $\mathrm{VAR}(p)$ model. The proof of the corollary just involves casting the estimation of model parameters in terms of the online linear regression model, i.e. Model 2.5.1. By the assumption that $\psi = \psi_N$, per the discussion following the statement of Theorem 2.5.2, it is not necessary to assume $\|X_n\| \leq 1$ for all $n \geq 1$.

Proof of Corollary 2.5.7. Let $(\mathcal{F}_n)_{n \geq 0}$ be the filtration outlined in Model 2.5.6, i.e. $\mathcal{F}_n := \sigma(Y_m : -p + 1 \leq m \leq n)$. Note that the $\mathbb{R}^k$-valued sequence $(X_n)_{n \geq 1}$ is $(\mathcal{F}_n)_{n \geq 1}$-predictable and the $\mathbb{R}^d$-valued noise sequence $(\epsilon_n)_{n \geq 1}$ is $(\mathcal{F}_n)_{n \geq 0}$-adapted. Further noting the identity

$$Y_n(i) = \langle \pi(i), X_n \rangle + \epsilon_n,$$

we see that we are exactly in the setting of Model 2.5.1. Thus, applying Theorem 2.5.2 yields the desired result. ■

Lastly, we prove Theorem 2.6.1, which provides a self-normalized, time-uniform empirical Bernstein inequality for multivariate processes. In the proof of Theorem 2.6.1, we will need the following lemma, which can be extracted from the proof of Theorem 4 in Howard et al. [75], which in turn generalizes a result by Fan et al. [59].

Lemma 2.B.2 (Theorem 4 of Howard et al. [75]). *Let $(X_n)_{n\geq 0}$ be a real-valued sequence of random variables adapted to some filtration $(\mathcal{F}_n)_{n\geq 0}$. Suppose that $|X_n| \leq 1/2$ almost surely for all $n \geq 1$. Then, for any $\lambda \in [0, \lambda)$, the process*

$$L_n^\lambda := \exp\left\{\lambda \sum_{m\leq n}(X_m - \mathbb{E}_{m-1}X_m) - \psi_{E,1}(\lambda)\sum_{m\leq n}(X_m - \widehat{\mu}_{m-1})^2\right\}$$

is a non-negative supermartingale with respect to $(\mathcal{F}_n)_{n\geq 0}$. Consequently, $\left(\sum_{m=1}^n (X_m - \mathbb{E}_{m-1}X_m)\right)_{n\geq 0}$ is sub-$\psi_{E,1}$ with variance proxy $\left(\sum_{m=1}^n (X_m - \widehat{\mu}_{m-1})^2\right)_{n\geq 0}$.

We now prove Theorem 2.6.1. All we need to do in the proof is check that the process $(S_n, V_n)_{n\geq 0}$ is sub-ψ_E in the sense of Definition 2.2.2. This boils down to checking that the projection of $(S_n, V_n)_{n\geq 0}$ onto any direction vector is sub-ψ_E in the scalar sense. With Lemma 2.B.2 in hand, checking this condition becomes trivial.

Proof. To prove the result, it suffices to check that (S_n, V_n) is sub-$\psi_{E,1}$, per Definition 2.2.2. Thus, we show that, for any $\nu \in \mathbb{S}^{d-1}$, $(\langle \nu, S_n\rangle, \langle \nu, V_n\nu\rangle)_{n\geq 0}$ is sub-$\psi_{E,1}$ in the sense of Definition 2.2.1. Clearly, $\langle \nu, S_n\rangle \in [-1/2, 1/2]$ almost surely. Further, we have

$$\begin{aligned}
\langle \nu, V_n\nu\rangle &= \sum_{m=1}^n \langle \nu, (X_m - \widehat{\mu}_{m-1})(X_m - \widehat{\mu}_{m-1})^\top \nu\rangle \\
&= \sum_{m=1}^n \langle \nu, X_m - \widehat{\mu}_{m-1}\rangle^2 \\
&= \sum_{m=1}^n (\langle \nu, X_m\rangle - \langle \nu, \widehat{\mu}_{m-1}\rangle)^2.
\end{aligned}$$

Thus, Lemma 2.B.2 implies that $(\langle \nu, S_n\rangle, \langle \nu, V_n\nu\rangle)_{n\geq 0}$ is sub-$\psi_{E,1}$, so the first claim follows. The second claim follows from Theorem 2.4.1. Finally, for any $\lambda \in [0, 1)$,

$$\psi_{E,1}(\lambda) = -\log(1-\lambda) - \lambda \leq \frac{\lambda^2}{2(1-\lambda)} =: \psi_{G,1}(c),$$

so $(S_n, V_n)_{n\geq 0}$ is sub-$\psi_{G,1}$ also. Noting that

$$(\psi_{G,1}^*)^{-1}(u) = \sqrt{2u} + u$$

yields the final claim. Proofs of these two facts surrounding $\psi_{E,1}$ and $\psi_{G,1}$ can be found in Boucheron et al. [18]. ■

2.C Proofs of Technical Lemmas

In this section, we provides proofs for the technical lemmas used in proving the main results of this paper. We start by proving Lemma 2.7.1, which is used in the proof of Theorem 2.4.1.

Proof of Lemma 2.7.1. Let $\nu \in \mathbb{S}^{d-1}$ be arbitrary and let $\omega \in \mathbb{S}^{d-1}$ be the unique vector satisfying (2.7.3). By the definition of ω and π_T, we have

$$\begin{aligned}\|\nu - \pi_T(\nu)\| &= \left\| \frac{T^{1/2}\omega}{\|T^{1/2}\omega\|} - \frac{T^{1/2}\pi(\omega)}{\|T^{1/2}\pi(\omega)\|} \right\| \\ &\leq \max\left\{ \left\| \frac{T^{1/2}\omega}{\|T^{1/2}\omega\|} - \frac{T^{1/2}\pi(\omega)}{\|T^{1/2}\omega\|} \right\|, \left\| \frac{T^{1/2}\omega}{\|T^{1/2}\pi(\omega)\|} - \frac{T^{1/2}\pi(\omega)}{\|T^{1/2}\pi(\omega)\|} \right\| \right\} \\ &\leq \frac{\gamma_{\max}(T^{1/2})}{\|T^{1/2}\omega\| \wedge \|T^{1/2}\pi(\omega)\|} \|\omega - \pi(\omega)\| \leq \sqrt{\kappa}\epsilon.\end{aligned}$$

Above, the second inequality follows from pulling out the denominator in each term of the maximum and bounding $\|T^{1/2}(\omega - \pi(\omega))\| \leq \gamma_{\max}(T^{1/2})\|\omega - \pi(\omega)\|$ and the last inequality follows as $\|T^{1/2}\omega\| \wedge \|T^{1/2}\pi(\omega)\| \geq \gamma_{\min}(T^{1/2})$, and $\|\omega - \pi(\omega)\| \leq \epsilon$ by definition of projection onto a cover. The first inequality follows from a simple calculation. To elaborate, assume $\|T^{1/2}\omega\| \neq \|T^{1/2}\pi(\omega)\|$, as in the case of equality there is nothing to prove in the inequality. Notice that if $\|T^{1/2}\omega\| < \|T^{1/2}\pi(\omega)\|$, then $q := T^{1/2}\omega/\|T^{1/2}\omega\|$ lies on the surface of the unit ball, and $p := T^{1/2}\pi(\omega)/\|T^{1/2}\omega\|$ lies outside of the unit ball (i.e. has norm greater than 1). The projection of p onto the unit ball is exactly $T^{1/2}\pi(\omega)/\|T^{1/2}\pi(\omega)\|$, which is closer to q than p since projections onto convex sets decrease Euclidean distance to all points. The maximum above comes from handling the case $\|T^{1/2}\omega\| > \|T^{1/2}\pi(\omega)\|$, which is analogous. This shows the desired result. ■

We now prove Lemma 2.4.2, which is leveraged in the proof of Corollary 2.4.6 and Corollary 2.4.3.

Proof of Lemma 2.4.2. We start by providing an upper bound on the proper ϵ-covering number for $\mathbb{S}_\infty^{d-1} := \{x \in \mathbb{R}^d : \|x\|_\infty := \max_{j\in[d]} |x_j| = 1\}$. Note that we can write

$$\mathbb{S}_\infty^{d-1} = \bigcup_{i=1}^{d} F_i^+ \cup F_i^-,$$

where $F_i^+ := \{x \in \mathbb{R}^d : x_i = 1, \|x_{-i}\|_\infty \leq 1\}$ and $F_i^- := \{x \in \mathbb{R}^d : x_i = -1, \|x_{-i}\|_\infty \leq 1\}$, where $x_{-i} \in R^{d-1}$ is the vector x with the ith component omitted. For any $i \in [d]$, $s \in \{+, -\}$, there is a natural isometry between F_i^s and the $(d-1)$-dimensional ℓ_∞ ball defined as $\mathbb{B}_{d-1}^\infty := \{x \in \mathbb{R}^{d-1} : \|x\|_\infty \leq 1\}$ given by $x \in \mathbb{R}^d \mapsto x_{-i} \in \mathbb{R}^{d-1}$. In particular, this implies the proper ϵ-covering number of F_i^s under the ℓ_2-norm is bounded as

$$N\left(F_i^s, \epsilon, \|\cdot\|\right) = N\left(\mathbb{B}_{d-1}^\infty, \epsilon, \|\cdot\|\right) \leq \frac{\mathrm{Vol}_{d-1}(\mathbb{B}_{d-1}^\infty)}{\mathrm{Vol}_{d-1}(\mathbb{B}_{d-1})} \left(\frac{3}{\epsilon}\right)^{d-1},$$

where the last inequality follows from Lemma 5.7 of Wainwright [158]. From this, we see that we have the bound

$$N\left(\mathbb{S}_\infty^{d-1}, \epsilon, \|\cdot\|\right) \leq 2d \frac{\mathrm{Vol}_{d-1}(\mathbb{B}_{d-1}^\infty)}{\mathrm{Vol}_{d-1}(\mathbb{B}_{d-1})} \left(\frac{3}{\epsilon}\right)^{d-1} = C_d \left(\frac{3}{\epsilon}\right)^{d-1},$$

where we have summed over the $2d$ different $(d-1)$-dimensional faces $F_1^+, F_1^-, \ldots, F_d^+, F_d^-$ and defined the constant $C_d := 2d\frac{\mathrm{Vol}_{d-1}(\mathbb{B}_{d-1}^\infty)}{\mathrm{Vol}_{d-1}(\mathbb{B}_{d-1})}$.

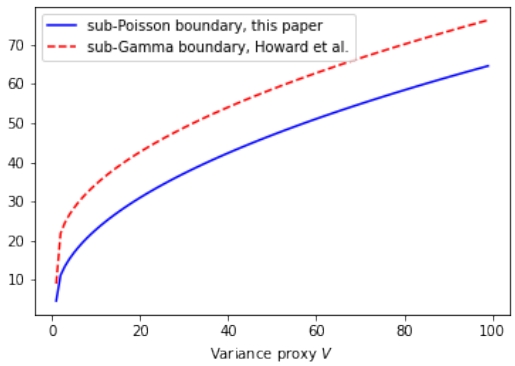
Let K now denote a minimal proper ϵ-covering of $\mathbb{S}_\infty^{d-1}$, and let $\pi : \mathbb{S}_\infty^{d-1} \to K$ denote the projection onto the covering. Further, let $p : \mathbb{R}^d \to \mathbb{B}_d$ denote the ℓ_2 projection onto the unit ball. We claim that the set $K' := \{p(x) : x \in K\}$ is a proper ϵ-covering of $\mathbb{S}^{d-1}$ under the ℓ_2 norm. The fact that $p(x) \in \mathbb{S}^{d-1}$ is immediate. Next, note that for any $y \in \mathbb{S}^{d-1}$, there is some $x \in \mathbb{S}_\infty^{d-1}$ such that $p(x) = y$. Observe that $z := p(\pi(x)) \in K'$. Since p is an ℓ_2 projection, we know that

$$\|y - z\| = \|p(x) - p(\pi(x))\| \leq \|x - \pi(x)\| \leq \epsilon,$$

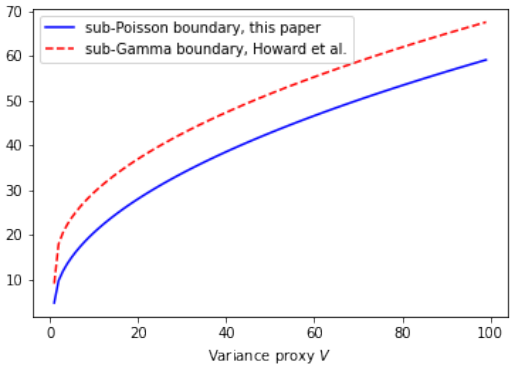
so we have shown that K' is a proper ϵ-covering.

■

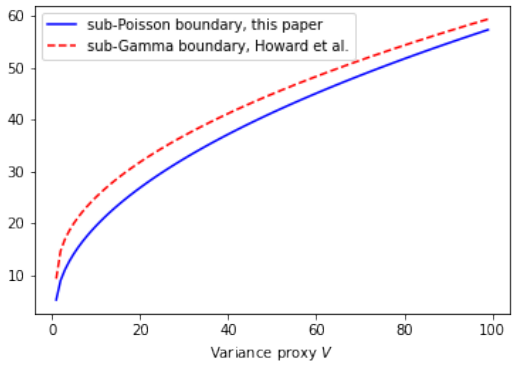
2.D Figures



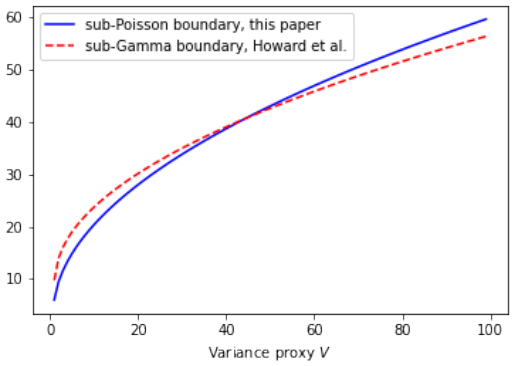
(a) $\alpha = 1.01$



(b) $\alpha = 1.05$

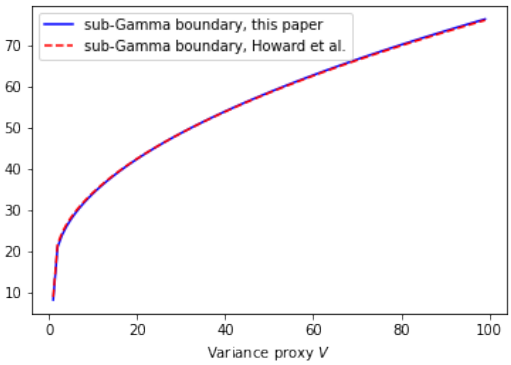


(c) $\alpha = 1.25$

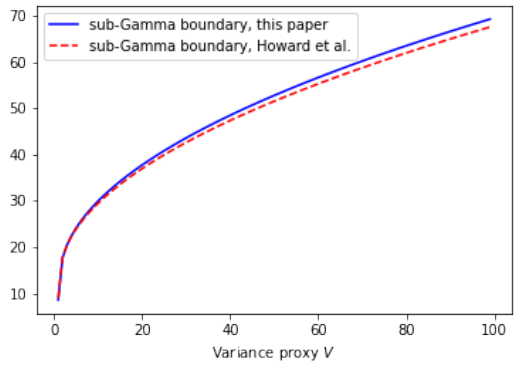


(d) $\alpha = 1.5$

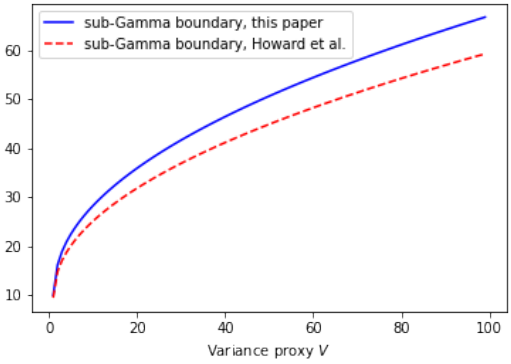
Figure 2.1: Comparing the boundary of Theorem 2.3.1 in the case $\psi = \psi_{P,c}$ with the boundary of Theorem 1 in Howard et al. [75], recapped in (2.3.1). Note that to apply the boundary of Howard et al. [75], we need to leverage the fact that a sub-$\psi_{P,c}$ process $(S_n)_{n\geq 0}$ is also sub-$\psi_{G,c}$ with the same variance proxy $(V_n)_{n\geq 0}$. We have made the parameter selection $c = 1$, $\delta = 0.01$, $\rho = 1$, and $h(k) = (1+k)^2\zeta(2)$, and have correspondingly varied α over several values. We see that for reasonably small choices of intrinsic time spacing $\alpha > 1$, our boundary is tighter than that of Howard et al. [75]. Thus, we see that although a sub-$\psi_{P,c}$ process can be viewed as a sub-$\psi_{G,c}$ process, this conversion introduces looseness, making our time-uniform concentration result generally preferable in this setting.



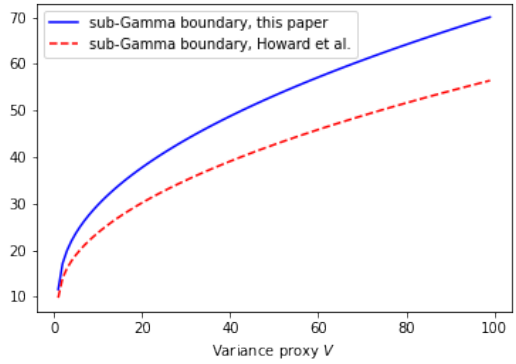
(a) $\alpha = 1.01$



(b) $\alpha = 1.05$

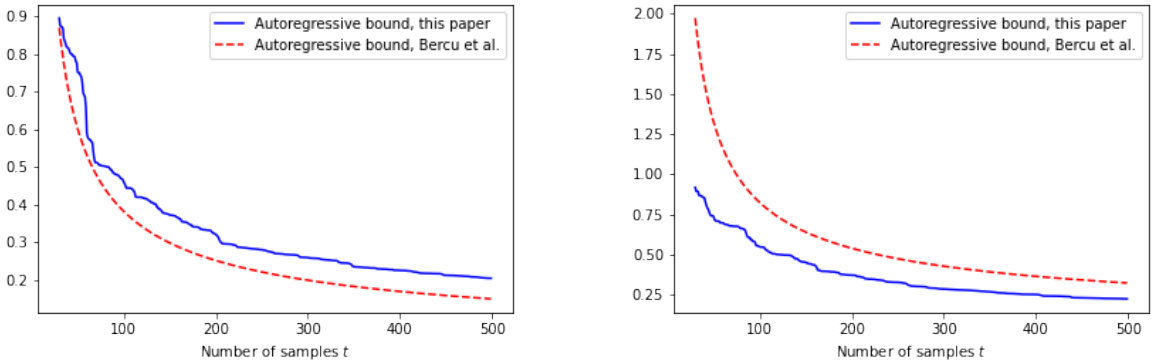


(c) $\alpha = 1.25$



(d) $\alpha = 1.5$

Figure 2.2: Comparing the boundary of Theorem 2.3.1 in the case $\psi = \psi_{G,c}$ with the boundary of Howard et al. [75] (presented in Equation 2.3.1). We have made the parameter selection $c = 1$, $\delta = 0.01$, $\rho = 1$, and $h(k) = (1+k)^2\zeta(2)$, and have correspondingly varied α over several values. As expected from our discussion, our boundary is looser than that of Howard et al. [75] for all values of α, with the gap between the boundaries vanishing as the geometric spacing α of variance/intrinsic time is decreased towards 1. Since $\alpha = 1.01$ or $\alpha = 1.05$ are reasonable choices for applying our concentration inequalities, our bounds are just as applicable as those of Howard et al. [75] even in the sub-Gamma setting.



(a) Fact 2.5.8 vs. Corollary 2.5.7 *without* union bound (b) Fact 2.5.8 vs. Corollary 2.5.7 *with* union bound

Figure 2.3: A comparison of the bounds on $|\widehat{a}_n - a|$ provided by Fact 2.5.8 and Corollary 2.5.7. In plotting both bounds, we have fixed the failure probability as $\delta = 0.01$. We have numerically solved for x such that the right hand side of Fact 2.5.8 is equal to the target failure probability. When applying Corollary 2.5.7, we have set $\alpha = 1.5, h(k) = (1+k)^2\zeta(2), \rho = 1$, and note that dependence on ϵ and β can be removed in the univariate case. In Subfigure 2.3a, we plot Fact 2.5.8 point-wise (i.e. we set the failure probability to be δ for each sample size n), and in Subfigure 2.3b, we take a union bound over samples, setting the failure probability to be $\frac{6\delta}{n^2\pi^2}$ for each t.

Chapter 3

Mean Estimation in Banach Spaces Under Infinite Variance and Martingale Dependence

We consider estimating the shared mean of a sequence of heavy-tailed random variables taking values in a Banach space. We revisit and extend a simple truncation-based mean estimator by Catoni and Giulini. While existing truncation-based approaches require a bound on the raw (non-central) second moment of observations, our results hold under a bound on either *the central or non-central pth moment for some $p > 1$. In particular, our results hold for distributions with infinite variance. The main contributions of the paper follow from exploiting connections between truncation-based mean estimation and the concentration of martingales in 2-smooth Banach spaces. We prove two types of time-uniform bounds on the distance between the estimator and unknown mean: line-crossing inequalities, which can be optimized for a fixed sample size n, and non-asymptotic law of the iterated logarithm type inequalities, which match the tightness of line-crossing inequalities at all points in time up to a doubly logarithmic factor in n. Our results do not depend on the dimension of the Banach space, hold under martingale dependence, and all constants in the inequalities are known and small.*

3.1 Introduction

Mean estimation is perhaps the most important primitive in the statistician's toolkit. When the data is light-tailed (perhaps sub-Gaussian, sub-Exponential, or sub-Gamma), the sample mean is the natural estimator of this unknown population mean. However, when the data fails to have finite moments, the naive plug-in mean estimate is known to be sub-optimal.

The failure of the plug-in mean has led to a rich literature focused on *heavy-tailed* mean estimation. In the univariate setting, statistics such as the thresholded/truncated mean estimator [153, 78], trimmed mean estimator [125, 113], median-of-means estimator [124, 81, 9], and the Catoni M-estimator [22, 159] have all been shown to exhibit favorable convergence guarantees. When a bound on the variance of the observations is known, many of these estimates enjoy

sub-Gaussian rates of performance [110], and this rate gracefully decays when only a bound on the pth central moment is known for some $p > 1$ [19].

In the more challenging setting of *multivariate* heavy-tailed data, modern methods include the geometric median-of-means estimator [121], the median-of-means tournament estimator [112], and the truncated mean estimator [23]. We provide a more detailed account in Section 3.1.2.

Of the aforementioned statistics, the truncated mean estimator is by far the simplest. This estimator, which involves truncating observations to lie within an appropriately-chosen ball centered at the origin, is extremely computationally efficient and can be updated online, very desirable for applied statistical tasks. However, this estimator also possesses a number of undesirable properties. First, it is not translation invariant, with bounds that depend on the *raw* moments of the random variables. Second, it requires a known bound on the pth moment of observations for some $p \geq 2$, thus requiring that the observations have finite variance. Third, bounds are only known in the setting of finite-dimensional Euclidean spaces — convergence is not understood in the setting of infinite-dimensional Hilbert spaces or Banach spaces.

The question we consider here is simple: are the aforementioned deficiencies fundamental to truncation-based estimators, or can they be resolved with an improved analysis? The goal of this work is to show that the latter is true, demonstrating how a truncation-based estimator can be extended to handle fewer than two central moments in general classes of Banach spaces.

3.1.1 Our Contributions

In this work, we revisit and extend a simple truncation-based mean estimator due to Catoni and Giulini [23]. Our estimator works by first using a small number of samples to produce a naive mean estimate, say through a sample mean. Then, the remaining sequence of observations is truncated to lie in an appropriately-sized ball centered at this initial mean estimate. These truncated samples are then averaged to provide a more robust estimate of a heavy-tailed mean.

While existing works study truncation-based estimators via PAC-Bayesian analyses [23, 32, 100], we find it more fruitful to study these estimators using tools from the theory of Banach space-valued martingales. In particular, by proving a novel extension of classical results on the time-uniform concentration of bounded martingales due to Pinelis [128, 129], we are able to greatly improve the applicability of truncation-based estimators. In particular, our estimator and analysis improves over that in [23] in the following ways:

1. The analysis holds in arbitrary 2-smooth Banach spaces instead of just finite-dimensional Euclidean space. This not only includes Hilbert spaces but also the commonly-studied L^α and ℓ^α spaces for $2 \leq \alpha < \infty$.

2. Our results require only a known upper bound on the conditional central pth moment of observations for some $p > 1$, and are therefore applicable to data lacking finite variance. Existing bounds for truncation estimators, on the other hand, require a bound on the non-central second moment.

3. Our bounds are time-uniform and hold for data with a martingale dependence structure. We prove two types of inequalities: *line-crossing inequalities*, which can be optimized for a

target sample size, and *non-asymptotic law of the iterated logarithm (LIL) type inequalities*, which match the tightness of the boundary-crossing inequalities at all times simultaneously up to a doubly logarithmic factor in the sample size.

4. We show that our estimator exhibits strong practical performance, and that our derived bounds are tighter than existing results in terms of constants. We run simulations which demonstrate that, for appropriate truncation diameters, the distance between our estimator and the unknown mean is tightly concentrated around zero.

Informally, if we assume that the central pth moments of all observations are conditionally bounded by v, and we let $\widehat{\mu}_n$ denote our estimate after n samples, then we show that

$$\|\widehat{\mu}_n - \mu\| = O\left(v^{1/p}(\log(1/\delta)/n)^{\frac{p-1}{p}}\right) \qquad \text{with probability } \geq 1-\delta.$$

As far as we are aware, the only other estimator to obtain the same guarantee in a similar setting is Minsker's geometric median-of-means [121] (while he doesn't state this result explicitly, it is easily derivable from his main bound). Minsker also works in a Banach space, but assumes that it is separable and reflexive, whereas we will assume that it is separable and smooth. While we obtain the same rates, we feel that our truncation-style estimator has several benefits over geometric median-of-means. First, it is computationally lightweight and easy to compute exactly. Second, our line-crossing inequalities do not require as many tuning parameters to instantiate. Third, we handle martingale dependence while Minsker does not. Finally, our analysis is significantly different from Minsker's—and from existing analyses of other estimators under heavy-tails—and may be of independent interest.

3.1.2 Related Work

Section 3.1.1 discussed the relationship between this paper and the two most closely related works of Catoni and Giulini [23] and Minsker [121]. We now discuss how our work is related to the broader literature, none of which addresses our problem directly, but tackles simpler special cases of our problem (e.g., assuming more moments or boundedness, or with observations in Hilbert spaces or Euclidean spaces).

Heavy-tailed mean estimation under independent observations. Truncation-based (also called threshold-based) estimators have a rich history in the robust statistics literature, dating back to works from Tukey, Huber, and others [78, 153]. These estimators have either been applied in the univariate setting or in $\mathbb{R}^d$ as in Catoni and Giulini [23]. A related estimator is the so-called trimmed-mean estimator, which removes extreme observations and takes the empirical mean of the remaining points [125, 113]. For real-valued observations with finite variance, the trimmed-mean has sub-Gaussian performance [125].

Separately, Catoni and Giulini [22] introduce an approach for mean estimation in $\mathbb{R}^d$ based on M-estimators with a family of appropriate influence functions. This has come to be called "Catoni's M-estimator." It requires at least two moments and fails to obtain sub-Gaussian rates. It faces the the additional burden of being less computationally efficient. A series of followup

works have improved this estimator in various ways: Chen et al. [26] extend it to handle a p-th moment for $p \in (1, 2)$ for real-valued observations, Gupta et al. [67] refine and sharpen the constants, and Mathieu [120] studies the optimality of general M-estimators for mean estimation.

Another important line of work on heavy-tailed mean estimation is based on median-of-means estimators [124, 81, 9]. These estimators generally break a dataset into several folds, compute a mean estimate on each fold, and then compute some measure of central tendency amongst these estimates. For real-valued observations, Bubeck et al. [19] study a median-of-means estimator that holds under infinite variance. Their estimator obtains the same rate as ours and Minsker's. Most relevant for our work is the result on *geometric median-of-means* due to Minsker [121], which can be used to aggregate several independent mean estimates in general separable Banach spaces. In Hilbert spaces, when instantiated with the empirical mean under a finite variance assumption, geometric median-of-means is nearly sub-Gaussian (see discussion in Section 3.1.1). We compare our threshold-based estimator extensively to geometric median-of-means in the sequel and demonstrate that we obtain the same rate of convergence.

Another important result is the multivariate tournament median-of-means estimator due to Lugosi and Mendelson [112]. For i.i.d. observations in $(\mathbb{R}^d, \|\cdot\|_2)$ with shared covariance matrix (operator) Σ, then Lugosi and Mendelson [112] show this estimator can obtain the optimal sub-Gaussian rate of $O(\sqrt{\mathrm{Tr}(\Sigma)/n} + \sqrt{\|\Sigma\|_{\mathrm{op}} \log(1/\delta)/n})$. However, this result requires the existence of a covariance matrix and does not extend to a bound on the p-th moment for $p \in (1, 2)$, which is the main focus of this work.

While the original form of the tournament median-of-means estimator was computationally inefficient (with computation hypothesized to be NP-Hard in a survey by Lugosi and Mendelson [110]), a computationally efficient approximation was developed by Hopkins [72], with followup work improving the running time [28]. Tournament median-of-means was extended to general norms in $\mathbb{R}^d$ [111], though the authors note that this approach is still not computationally feasible. Median-of-means style approaches have also been extended to general metric spaces [77, 29]. Of the above methods, only the geometric median-of-means estimator can handle observations that lack finite variance.

Sequential concentration under martingale dependence. Time-uniform concentration bounds, or concentration inequalities that are valid at data-dependent stopping times, have been the focus of significant recent attention [74, 75, 166]. Such results are often obtained by identifying an underlying nonnegative supermartingale and then applying Ville's inequality [157], a strategy that allows for martingale dependence quite naturally. This approach is also used here. Wang and Ramdas [159] extend Catoni's M-estimator to handle both infinite variance and martingale dependence in $\mathbb{R}$, while Chugg et al. [32] give a sequential version of the truncation estimator in $\mathbb{R}^d$, though they require a central moment assumption and finite variance. The analyses of both Catoni and Giulini [23] and Chugg et al. [32] rely on so-called "PAC-Bayes" arguments [21, 33]. Intriguingly, while we analyze a similar estimator, our analysis avoids such techniques and is much closer in spirit to Pinelis-style arguments [128, 129].

Howard et al. [74, 75] provide a general collection of results on time-uniform concentration for scalar processes, which in particular imply time-uniform concentration results for some heavy-tailed settings (e.g. symmetric observations). Likewise, Whitehouse et al. [166] provide

a similar set of results in $\mathbb{R}^d$. While interesting, we note that these results differ from our own in that they are *self-normalized*, or control the growth of a process appropriately normalized by some variance proxy (here a mixture of adapted and predictable covariance). The results also don't apply when only a bound on the pth moment is known, and the latter set of results have explicit dependence on the ambient dimension d.

Concentration in Hilbert and Banach Spaces. There are several results related to concentration in infinite-dimensional spaces. A series of works has developed self-normalized, sub-Gaussian concentration bounds in Hilbert spaces [164, 2, 30] based on the famed method of mixtures [40, 41]. These results have not been extended to more general tail conditions. Significant progress has been made on the concentration of bounded random variables in smooth and separable Banach spaces. Pinelis [128, 129] presented a martingale construction for bounded observations, thus enabling dimension-free Hoeffding and Bernstein inequalities. Dimension-dependence is replaced by the smoothness parameter of the Banach space, which for most practical applications (in Hilbert spaces, say) equals one. These results were strengthened slightly by Howard et al. [74]. Recently, Martinez-Taboada and Ramdas [118] gave an *empirical*-Bernstein inequality in Banach spaces, also using a Pinelis-like construction. Our work adds to this growing literature by extending Pinelis' tools to the heavy-tailed setting.

3.1.3 Preliminaries

We introduce some of the background and notation required to state our results. We are interested in estimating the shared, conditional mean μ of a sequence of random variables $(X_n)_{n\geq 1}$ living in some separable Banach space $(\mathbb{B}, \|\cdot\|)$. Recall that a Banach space is a complete normed vector space; examples include Hilbert spaces, ℓ^α sequence spaces, and L^α spaces of functions. We make the following central assumption.

Assumption 1. *We assume $(X_n)_{n\geq 1}$ are a sequence of $\mathbb{B}$-valued random variables adapted to a filtration $\mathcal{F} \equiv (\mathcal{F}_n)_{n\geq 0}$ such that*

(1) $\mathbb{E}(X_n \mid \mathcal{F}_{n-1}) = \mu$*, for all $n \geq 1$ and some unknown $\mu \in \mathbb{B}$, and*

(2) $\sup_{n\geq 1} \mathbb{E}\left(\|X_n - \mu\|^p \mid \mathcal{F}_{n-1}\right) \leq v < \infty$ *for some known constants $p \in (1, 2]$ and $v > 0$.*

The martingale dependence in condition (1) above is weaker than the traditional i.i.d. assumption, requiring only a constant *conditional* mean. This is useful in applications such as multi-armed bandits, where we cannot assume that the next observation is independent of the past. Meanwhile, condition (2) allows for infinite variance, a weaker moment assumption than past works studying concentration of measure in Banach spaces (e.g., [121, 128, 129]). In Appendix 3.A we replace condition (2) with a bound on the raw moment (that is, $\mathbb{E}(\|X_n\|^p|\mathcal{F}_{n-1})$) for easier comparison with previous work. We note that other works studying truncation-based estimators have exclusively considered the $p \geq 2$ setting where observations admit covariance matrices [32, 23, 110]. We focus on $p \in (1, 2]$ in this work, but it is likely our techniques could be naturally extended to the $p \geq 2$ setting. We leave this as interesting future work.

In order obtain concentration bounds, we must assume the Banach space is reasonably well-behaved. This involves assuming that is it both separable and *smooth*. A space is separable if it

contains a countable, dense subset, and a real-valued function $f : \mathbb{B} \to \mathbb{R}$ is $(2, \beta)$-smooth if, for all $x, y \in \mathbb{B}$, $f(0) = 0$, $|f(x + y) - f(x)| \leq \|y\|$, and

$$f^2(x + y) + f^2(x - y) \leq 2f^2(x) + 2\beta^2\|y\|^2. \tag{3.1.1}$$

We assume that the norm is smooth in the above sense.

Assumption 2. *We assume that the Banach space $(\mathbb{B}, \|\cdot\|)$ is both separable and $(2, \beta)$-smooth, meaning that the norm satisfies* (3.1.1)*.*

Assumption 2 is common when studying Banach spaces [128, 129, 74, 118]. We emphasize that β is not akin to the dimension of the space. For instance, infinite-dimensional Hilbert spaces have $\beta = 1$ and L^α and ℓ^α spaces have $\beta = \sqrt{\alpha - 1}$ for $\alpha \geq 2$. Thus, bounds which depend on β are still dimension-free.

Notation and background. For notational simplicity, we define the conditional expectation operator $\mathbb{E}_{n-1}[\cdot]$ to be $\mathbb{E}_{n-1}[X] := \mathbb{E}(X \mid \mathcal{F}_{n-1})$ for any $n \geq 1$. If $S \equiv (S_n)_{n\geq 0}$ is some stochastic process, we denote the n-th increment as $\Delta S_n := S_n - S_{n-1}$ for any $n \geq 1$. For any process or sequence $a \equiv (a_n)_{n\geq 1}$, denote by a^n the first n values: $a^n = (a_1, \ldots, a_n)$. We say the process S is *predictable* with respect to filtration $\mathcal{F}$, if S_n is $\mathcal{F}_{n-1}$-measurable for all $n \geq 1$. Our analysis will make use of both the Fréchet and Gateaux derivatives of functions in a Banach space. We do not define these notions here but instead refer to Ledoux and Talagrand [106].

Outline. Section 3.2 provides statements of the main results. Our main result, Theorem 3.2.1, is a general template for obtaining bounds (time-uniform boundary-crossing inequalities in particular) on truncation-style estimators. Corollary 3.2.2 then instantiates the template with particular parameters to obtain tightness for a fixed sample size. Section 3.3 is dedicated to the proof of Theorem 3.2.1. Section 3.4 then uses a technique known as "stitching" to extend our line-crossing inequalities to bounds which shrink to zero over time at an iterated logarithm rate. Finally, Section 3.5 provides several numerical experiments demonstrating the efficacy of our proposed estimator in practice.

3.2 Main Result

Define the mapping

$$\texttt{Trunc} : \mathbb{B} \to [0, 1] \text{ by } x \mapsto \frac{1 \wedge \|x\|}{\|x\|}. \tag{3.2.1}$$

Clearly, $\texttt{Trunc}(x)x$ is just the projection of x onto the unit ball in B. Likewise, $\texttt{Trunc}(\lambda x)x$ is the projection of x onto the ball of radius λ^{-1} in B. We note that the truncated observations $\texttt{Trunc}(X_n)X_n$ are themselves random variables, which are adapted to the underlying filtration $\mathcal{F}$.

As we discussed in Section 3.1.1, previous analyses of truncation-style estimators have relied on a bound on the raw second moment. To handle a central moment assumption, we will center

our estimator around a naive mean estimate which has worse guarantees but whose effects wash out over time.

To formalize the above, our estimate of μ at time n will be

$$\widehat{\mu}_n(k) \equiv \widehat{\mu}_n(k, \lambda, \widehat{Z}_k) := \frac{1}{n} \sum_{k<m\leq n} \left\{ \texttt{Trunc}(\lambda(X_m - \widehat{Z}_k))(X_m - \widehat{Z}_k) + \widehat{Z}_k \right\}, \tag{3.2.2}$$

where $\widehat{Z}_k$ is a naive mean estimate formed using the first k samples and $\lambda > 0$ is some fixed hyperparameter. Defining $\widehat{Z}_0 = 0$ when $k = 0$, we observe that $\widehat{\mu}_n(0)$ is the usual truncation estimator, analyzed by Catoni and Giulini [23] in the fixed-time setting and Chugg et al. [32] in the sequential setting. To state our result, define

$$K_p := \frac{1}{p/q+1} \left(\frac{p/q}{p/q+1} \right)^{p/q} \quad \text{where } \frac{1}{p} + \frac{1}{q} = 1, \tag{3.2.3}$$

which depends on the Holder conjugate q of p. Note that $K_p < 1$ for all $p > 1$. In fact, $\lim_{p\to 1} K_p = 1$, $\lim_{p\to\infty} K_p = 0$, and K_p is decreasing in p. We also define the constant

$$\mathfrak{C}_p(\mathbb{B}) = \begin{cases} 2^{p-1}(\frac{e^2-3}{4}), & \text{if } (\mathbb{B}, \|\cdot\|) \text{ is a Hilbert space,} \\ \beta^2 2^{p+1}(\frac{e^2-3}{4}), & \text{otherwise,} \end{cases} \tag{3.2.4}$$

which depends on the geometry and smoothness β of the Banach space $(\mathbb{B}, \|\cdot\|)$. In a Hilbert space (for which $\beta = 1$), the variance of our supermartingale increments can be more easily bounded. If the norm is not induced by an inner product, then $\mathfrak{C}_p(\mathbb{B})$ suffers an extra factor of four. Note that $\frac{e^2-3}{4} < 1.1$.

Our main result is the following template for bounding the deviations of $\widehat{\mu}_n$ assuming some sort of concentration of $\widehat{Z}_k$ around μ.

Theorem 3.2.1 (Main result)**.** *Let $X_1, X_2, \ldots$ be random variables satisfying Assumption 1 which lie in some Banach space $(\mathbb{B}, \|\cdot\|)$ satisfying Assumption 2. Suppose we use the first k samples to construct $\widehat{Z}_k$ which satisfies, for any $\delta \in (0, 1]$,*

$$\mathbb{P}(\|\mu - \widehat{Z}_k\| \geq r(\delta, k)) \leq \delta, \tag{3.2.5}$$

for some function $r : (0, 1] \times \mathbb{N} \to \mathbb{R}_{\geq 0}$. Fix any $\delta \in (0, 1]$. Decompose δ as $\delta = \delta_1 + \delta_2$ where $\delta_1, \delta_2 > 0$. Then, for any $\lambda > 0$, with probability $1 - \delta$, simultaneously for all $n \geq k$, we have:

$$\left\| \widehat{\mu}_n(k, \lambda, \widehat{Z}_k) - \mu \right\| \leq \lambda^{p-1}(\mathfrak{C}_p(\mathbb{B}) + K_p 2^{p-1})(v + r(\delta_2, k)^p) + \frac{\log(2/\delta_1)}{\lambda(n-k)}. \tag{3.2.6}$$

The guarantee provided by Theorem 3.2.1 is a line-crossing inequality in the spirit of [74]. That is, if we multiply both sides by $n-k$, it provides a time-uniform guarantee on the probability that the left hand side deviation between $\widehat{\mu}_n$ and μ ever crosses the line parameterized by the right hand side of (3.2.6). If we optimize the value of λ for a particular sample size n^*, the bound will remain valid for all sample sizes, but will be tightest at and around $n = n^*$. To obtain bounds that are tight for all n simultaneously, one must pay an additional iterated logarithmic price in

n. To accomplish this, Section 3.4 will deploy a carefully designed union bound over geometric epochs—a technique known as "stitching" [75]. However, for practical applications where the sample size is known in advance, we recommend Theorem 3.2.1 and its corollaries.

Next we provide a guideline on choosing λ in Theorem 3.2.1. The proof is straightforward.

Corollary 3.2.2. *In Theorem 3.2.1, consider taking*

$$\lambda = \left(\frac{\log(2/\delta_1)}{(\mathfrak{C}_p(\mathbb{B}) + K_p 2^{p-1})(n-k)(v + r(\delta_2, k)^p)} \right)^{1/p}. \tag{3.2.7}$$

Then, with probability at least $1 - \delta_1 - \delta_2$, we have

$$\|\widehat{\mu}_n(k) - \mu\| \leq \left((\mathfrak{C}_p(\mathbb{B}) + K_p 2^{p-1})(v + r(\delta_2, k)^p) \right)^{1/p} \left(\frac{\log(2/\delta_1)}{n-k} \right)^{(p-1)/p}.$$

In particular, as long as $k = o(n)$, $r(\delta, k) = o(1)$ and $\delta_1, \delta_2 = \Theta(\delta)$, we have

$$\|\widehat{\mu}_n(k) - \mu\| = O\left(v^{1/p} \left(\frac{\log(1/\delta)}{n} \right)^{(p-1)/p} \right). \tag{3.2.8}$$

This is the desired rate per the discussion in Section 3.1.1, matching the rate of other estimators which hold under infinite variance. In particular, it matches the rates of Bubeck et al. [19] in scalar settings and Minsker [121] in Banach spaces.

Now let us instantiate Theorem 3.2.1 when we take $\widehat{Z}_k$ to be either the sample mean or Minsker's geometric median-of-means. The latter provides a better dependence on δ_2 but at an additional computational cost. As we'll see in Section 3.5, this benefit is apparent for small sample sizes, but washes out as n grows.

Corollary 3.2.3. *Let $(\mathbb{B}, \|\cdot\|)$ satisfy Assumption 2 and $X_1, \ldots, X_n$ satisfy Assumption 1. For some $k < n$, let $\widehat{Z}_k$ be the empirical mean of the first k observations. Given $\delta > 0$, decompose it as $\delta = \delta_1 + \delta_2$ for any $\delta_1, \delta_2 > 0$. Then, with probability $1 - \delta$,*

$$\|\widehat{\mu}_n(k) - \mu\| \leq 2v^{1/p} C_p^{1/p} \left(\frac{\log(2/\delta_1)}{(n-k)^{1/p}} \right)^{(p-1)/p} \left(1 + O\left(\frac{1}{\delta_2 k^{p-1}} \right) \right), \tag{3.2.9}$$

where $C_p = \mathfrak{C}_p(\mathbb{B}) + K_p 2^{p-1}$. If, on the other hand, $\widehat{Z}_k$ is the geometric median-of-means estimator with appropriate tuning parameters, then with probability $1 - \delta$,

$$\|\widehat{\mu}_n(k) - \mu\| \leq 2v^{1/p} C_p^{1/p} \left(\frac{\log(2/\delta_1)}{(n-k)^{1/p}} \right)^{p-1} \left(1 + O\left(\frac{\log(1/\delta_2)^{p-1}}{k^{(p-1)/p}} \right) \right). \tag{3.2.10}$$

When $\widehat{Z}_k$ is the empirical mean, if $k = k(n) = \lfloor \log_2(n) \rfloor$ (say), we have $n - k(n) \geq n/2$ for $n \geq 2$, so (3.2.9) recovers the rate in (3.2.8), since the additional factor of $O(\frac{1}{\delta_2 k(n)^{p-1}}) = O(\frac{1}{\delta_2 \log(n)^{p-1}})$ is $o(1)$ and vanishes. When $\widehat{Z}_k$ is the geometric median-of-means, this error term vanishes even faster.

3.3 Proof of Theorem 3.2.1

We will prove a slightly more general result that reduces to Theorem 3.2.1 in a special case. Throughout this section, fix two $\mathcal{F}$-predictable sequences, $(\widehat{Z}_n) \in \mathbb{B}^{\mathbb{N}}$ and $(\lambda_n) \in \mathbb{R}_+^{\mathbb{N}}$. Define

$$\widehat{\xi}_n \equiv \widehat{\xi}_n(\lambda^n, Z^n) := \sum_{m \le n} \lambda_m \big\{ \texttt{Trunc}(\lambda_m Y_m) Y_m + \widehat{Z}_m \big\} \text{ where } Y_m := X_m - \widehat{Z}_m, \tag{3.3.1}$$

If we take λ_n to be constant and $\widehat{Z}_m = \widehat{Z}$ to be $\mathcal{F}_0$-measurable, then $\widehat{\xi}_n = n\lambda\widehat{\mu}_n$. We will make such a substitution at the end of this analysis to prove the desired result. However, working with the more general process (3.2.2) has advantages. In particular, it allows us to consider sequences of predictable mean-estimates, if desired.

Our preliminary goal is to find a process $(V_n)_{n \ge 0}$ such that the process

$$M_n(\lambda^n) = \exp\left\{ \left\| \widehat{\xi}_n - \mu \sum_{m \le n} \lambda_m \right\| - V_n(\lambda^n) \right\}, \tag{3.3.2}$$

is upper bounded by a nonnegative supermartingale; in other words; in recent parlance, it is an e-process [131]. Applying Ville's inequality will then give us a time-uniform bound on the deviation of the process $\|(\sum_{m \le n} \lambda_m)^{-1} \widehat{\xi}_n - \mu\|$ in terms of $V_n(\lambda)$. We will let

$$\begin{aligned} & V_n(\lambda) = (\mathfrak{C}_p(\mathbb{B}) + K_p 2^{p-1}) G_n, \\ \text{where} \quad & G_n \equiv G_n(\lambda^n, \widehat{Z}^n) := \sum_{m \le n} \lambda_m^p (v + \|\mu - \widehat{Z}_m\|^p), \end{aligned} \tag{3.3.3}$$

is a weighted measure of the deviation of the naive estimates $\widehat{Z}_1, \dots, \widehat{Z}_n$ from μ. Since it is difficult to reason about the difference between $\widehat{\xi}_n$ and $\mu \sum_{m \le n} \lambda_m$ directly, we introduce the proxy

$$\widetilde{\mu}_n(\lambda) := \mathbb{E}_{n-1}[\texttt{Trunc}(\lambda Y_n) Y_n] + \widehat{Z}_n, \tag{3.3.4}$$

and argue about $\|\widehat{\xi}_n - \sum_{m \le n} \lambda_m \widetilde{\mu}_m(\lambda_m)\|$ and $\lambda_m \|\widetilde{\mu}_m(\lambda_m) - \mu\|$. We then bound the difference $\|\widehat{\xi}_n - \mu \sum_{m \le n} \lambda_m\|$ using the triangle inequality.

3.3.1 Step 1: Bounding $\|\widetilde{\mu}_n(\lambda) - \mu\|$

We need the following analytical property of Trunc, which will be useful in bounding the truncation error with fewer than two moments. We note that the following lemma was used by Catoni and Giulini [23] for $k \ge 1$. We prove the result for $k > 0$.

Lemma 3.3.1. *For any $k > 0$ and $x \in \mathbb{B}$, we have that*

$$1 - \texttt{Trunc}(x) \le \frac{\|x\|^k}{k+1} \left(\frac{k}{k+1} \right)^k.$$

Proof. Fix $k > 0$. It suffices to show that $f(t) := 1 - \frac{1 \wedge t}{t} \leq \frac{t}{k+1}\left(\frac{k}{k+1}\right)^k =: g_k(t)$ for all $t \geq 0$. For $t \in [0,1]$, the result is obvious. For $t \geq 1$, we need to do a bit of work. First, note that $g_k(1) > f(1) = 0$, and that both g_k and f are continuous. Further, we only have $g_k(t) = f(t)$ precisely when $t = \frac{k+1}{k}$. Let this value of t be t^*. This immediately implies that $g_k(t) \geq f(t)$ for $t \in [1, t^*]$. To check the inequality for all $t \geq t^*$, it suffices to check that $f'(t) < g_k'(t)$. We verify this by direct computation. First, $f'(t) = \frac{1}{t^2}$. Likewise, we have that $g_k'(t) = t^{k-1}\left(\frac{k}{k+1}\right)^{k+1}$. Taking ratios, we see that

$$\frac{g_k'(t)}{f'(t)} = t^{k+1}\left(\frac{k}{k+1}\right)^{k+1} \geq \left(\frac{k+1}{k}\right)^{k+1}\left(\frac{k}{k+1}\right)^{k+1} = 1,$$

proving the desired result. ∎

We can now proceed to bounding $\|\widetilde{\mu}_n(\lambda) - \mu\|$.

Lemma 3.3.2. *Let X be a $\mathbb{B}$-valued random variable and suppose $\mathbb{E}_{n-1}\|X - \mu\|^p \leq v < \infty$. Let $\widehat{Z}_n$ be $\mathcal{F}_{n-1}$-predictable and $\widetilde{\mu}_n$ be as in* (3.3.4)*. Then:*

$$\|\mu - \widetilde{\mu}_n(\lambda)\| \leq K_p 2^{p-1}\lambda^{p-1}(v + \|\widehat{Z}_n - \mu\|^p).$$

Proof. Since $\widehat{Z}_n$ is predictable, we may treat it as some constant z when conditioning on $\mathcal{F}_{n-1}$. Using Holder's inequality, write

$$\begin{aligned}
\|\mu - \widetilde{\mu}_n(\lambda)\| &= \|\mathbb{E}_{n-1}[X_n] - \mathbb{E}_{n-1}[\texttt{Trunc}(\lambda(X_n - z))(X_n - z) + z]\| \\
&= \|\mathbb{E}_{n-1}[\{1 - \texttt{Trunc}(\lambda(X_n - z))\}(X_n - z)\| \\
&\leq \mathbb{E}[|1 - \texttt{Trunc}(\lambda(X_n - z))|^q]^{1/q}\mathbb{E}[\|X_n - z\|^p]^{1/p},
\end{aligned}$$

where $1/p + 1/q = 1$. The second expectation on the right hand side can be bounded using Minkowski's inequality and the fact that $\|\cdot\|^p$ is convex for $p \geq 1$:

$$\begin{aligned}
\mathbb{E}_{n-1}[\|X_n - z\|^p] &= \mathbb{E}_{n-1}[\|X_n - \mu + \mu - z\|^p] \\
&\leq 2^{p-1}\left(\mathbb{E}_{n-1}[\|X_n - \mu\|^p] + \|z - \mu\|^p\right) \\
&\leq 2^{p-1}(v + \|z - \mu\|^p). \qquad (3.3.5)
\end{aligned}$$

Next, by Lemma 3.3.1, we have for any $k > 0$.

$$\mathbb{E}_{n-1}[|1 - \texttt{Trunc}(\lambda(X_n - z))|^q] \leq \mathbb{E}_{n-1}\left[\left(\frac{\lambda^k\|X_n - z\|^k}{k+1}\left(\frac{k}{k+1}\right)^k\right)^q\right].$$

In particular, selecting $k = \frac{p}{q}$, we have

$$\mathbb{E}_{n-1}[|1 - \texttt{Trunc}(\lambda(X_n - z))|^q] \leq K_p^q\mathbb{E}_{n-1}\left[\lambda^p\|X_n - z\|^p\right] \leq K_p^q\lambda^p 2^{p-1}(v + \|z - \mu\|^p),$$

where K_p as defined in (3.2.3). Piecing everything together, we have that

$$\mathbb{E}_{n-1}[|1 - \texttt{Trunc}(\lambda(X_n - z))|^q]^{1/q} \leq K_p\lambda^{p/q}2^{(p-1)/q}(v + \|z - \mu\|^p)^{1/q}.$$

Therefore, recalling that $p/q = p - 1$, we have $\|\mu - \widetilde{\mu}_n(\lambda)\| \leq K_p\lambda^{p-1}2^{p-1}(v + \|z - \mu\|^p)$, which is the desired result. ∎

3.3.2 Step 2: Bounding $\|\widehat{\xi}_n(\lambda^n) - \sum_{m\le n}\lambda_m\widetilde{\mu}_m(\lambda_m)\|$

We can now proceed to bounding $\|\widehat{\xi}_n(\lambda^n) - \sum_{m\le n}\lambda_m\widetilde{\mu}_m(\lambda_m)\| = \|S_n(\lambda^n, \widehat{Z}^n)\|$ where

$$S_n \equiv S_n(\lambda^n, \widehat{Z}^n) := \sum_{m=1}^{n} \lambda_m \left\{ \texttt{Trunc}(\lambda_m Y_m) Y_m + \widehat{Z}_m - \widetilde{\mu}_m(\lambda_m) \right\}. \tag{3.3.6}$$

and $\widetilde{\mu}$ is as in (3.3.4). Note that S is a martingale with respect to $\mathcal{F}$. The following proposition is the most technical result in the paper. It follows from a modification of the proof of Theorem 3.2 in Pinelis [129], combined with a Bennett-type inequality for 2-smooth separable Banach spaces presented in Pinelis [129, Theorem 3.4]. We present the full result here, even those parts found in Pinelis' earlier work, for the sake of completeness.

Proposition 3.3.3. *Let $(X_t)_{t\ge1}$ be a process satisfying Assumption 1 and lying in a Banach space $(\mathbb{B}, \|\cdot\|)$ satisfying Assumption 2. Then, the exponential process*

$$U_n(\lambda^n, \widehat{Z}^n) := \exp\left\{ \left\| S_n(\lambda^n, \widehat{Z}^n) \right\| - \mathfrak{C}_p(\mathbb{B}) G_n \right\},$$

is bounded above by a nonnegative supermartingale with initial value 2, where G_n is defined by (3.3.3)*.*

Proof. Fix some $n \ge 1$ and let $U_n = U_n(\lambda^n, \widehat{Z}_n)$. We first observe that

$$\begin{aligned} \|\Delta S_n\| &= \lambda_n \|\texttt{Trunc}(\lambda_n Y_n) Y_n + \widehat{Z}_n - \widetilde{\mu}_n(\lambda_n)\| \\ &\le \lambda_n \|\texttt{Trunc}(\lambda_n Y_n) Y_n\| + \lambda_n \|\widehat{Z}_n - \widetilde{\mu}_n(\lambda_n)\| \le 2, \end{aligned}$$

by definition of Trunc. Let $T_n = \texttt{Trunc}(\lambda_n Y_n) Y_n$. If $(\mathbb{B}, \|\cdot\|)$ is a Hilbert space with inner product $\langle\cdot,\cdot\rangle$ (which induces $\|\cdot\|$), then

$$\mathbb{E}_{n-1}\|\Delta S_n\|^2 = \lambda_n^2 \mathbb{E}_{n-1}\langle T_n - \mathbb{E}_{n-1}T_n, T_n - \mathbb{E}_{n-1}T_n\rangle \le \lambda_n^2 \mathbb{E}_{n-1}\|T_n\|^2.$$

Otherwise, we have

$$\begin{aligned} \mathbb{E}_{n-1}\|\Delta S_n\|^2 &\le \lambda_n^2 \{\mathbb{E}_{n-1}\left(\|T_n\| + \|\mathbb{E}_{n-1}T_n\|\right)^2\} \\ &\le 2\lambda_n^2 \{\mathbb{E}_{n-1}\|T_n\|^2 + \|\mathbb{E}_{n-1}T_n\|^2\} \le 4\lambda_n^2 \mathbb{E}_{n-1}\|T_n\|^2, \end{aligned}$$

where the penultimate inequality uses that $(a+b)^2 \le 2a^2 + 2b^2$ and the final inequality follows from Jensen's inequality. Therefore, we can write

$$\mathbb{E}_{n-1}\|\Delta S_n\|^2 \le C\lambda_n^2 \mathbb{E}_{n-1}\|T_n\|^2, \tag{3.3.7}$$

where $C = 1$ if $\|\cdot\|$ is induced by an inner product, and $C = 4$ otherwise. We note that this extra factor of 4 is responsible for the two cases in the definition of $\mathfrak{C}_p(\mathbb{B})$ in (3.2.4). Carrying on with the calculation, write

$$\mathbb{E}_{n-1}\|\Delta S_n\|^2 \le C\lambda_n^2 \mathbb{E}_{n-1}\left[\|T_n\|^p \|T_n\|^{2-p}\right]$$

$$
\begin{aligned}
&\leq C\lambda_n^p \mathbb{E}_{n-1} \|T_n\|^p \\
&\leq C\lambda_n^p \mathbb{E}_{n-1} \|Y_n\|^p \\
&\leq C\lambda_n^p 2^{p-1}(v + \|\mu - \widehat{Z}_n\|^p), \qquad (3.3.8)
\end{aligned}
$$

where the final inequality follows by the same argument used to prove (3.3.5) in Lemma 3.3.2. We have shown that the random variable $\|\Delta S_n\|$ is bounded and its second moment (conditioned on the past) can be controlled, which opens the door to Pinelis-style arguments (see Pinelis [129, Theorem 3.4] in particular). Define the function $\varphi : [0,1] \to \mathbb{R}_{\geq 0}$ by

$$
\varphi(\theta) := \mathbb{E}_{n-1} \cosh\left(\|S_{n-1} + \theta \Delta S_n\|\right).
$$

In principle, the norm function need not be smooth, and so the same applies to φ. However, Pinelis [129] proved that one may assume smoothness of the norm without loss of generality (see Pinelis [129, Remark 2.4]). Thus, a second order Taylor expansion yields

$$
\mathbb{E}_{n-1} \cosh(\|S_n\|) = \varphi(1) = \varphi(0) + \varphi'(0) + \int_0^1 (1-\theta)\varphi''(\theta) d\theta.
$$

Observe that

$$
\begin{aligned}
\varphi''(\theta) &\leq \beta^2 \mathbb{E}_{n-1}\left[\|\Delta S_n\|^2 \cosh(\|S_{n-1}\|) e^{\theta\|\Delta S_n\|}\right] \\
&\leq \beta^2 \cosh(\|S_{n-1}\|) \mathbb{E}_{n-1}\left[\|\Delta S_n\|^2\right] e^{2\theta},
\end{aligned}
$$

where the first inequality follows from the proof of Theorem 3.2 in Pinelis [129] and Theorem 3 in Pinelis [128], and the second inequality is obtained in view of $\|\Delta S_n\| \leq 2$.

Next, by the chain rule, we have

$$
\begin{aligned}
\varphi'(0) &= \frac{d}{dt}\left(\mathbb{E}_{n-1} \cosh(\|S_{n-1} + t\Delta S_n\|)\right)\Big|_{t=0} \\
&= \mathbb{E}_{n-1}\left[\frac{d}{dt} \cosh(\|S_{n-1} + t\Delta S_n\|)\Big|_{t=0}\right] \\
&= \mathbb{E}_{n-1}\left[\langle D_f \|f\|\big|_{f=S_{n-1}}, \Delta S_n\rangle \cdot \frac{d}{dx}\cosh(x)\big|_{x=\|S_{n-1}\|}\right] \\
&= \left\langle D_f \|f\|\big|_{f=S_{n-1}}, \mathbb{E}_{n-1}\Delta S_n \right\rangle \cdot \frac{d}{dx}\cosh(x)\big|_{x=\|S_{n-1}\|} \\
&= 0,
\end{aligned}
$$

where $\langle D_f \varphi(f)\,|_{f=g}, y - x\rangle$ denotes the Gateaux derivative of φ with respect to f at g in the directon of $y - x$. The final equality follows from the fact that $(S_n)_{n\geq 0}$ is itself a martingale with respect to $(\mathcal{F}_n)_{n\geq 1}$. Thus, leveraging that $\varphi'(0) = 0$, we have

$$
\begin{aligned}
\mathbb{E}_{n-1} \cosh(\|S_n\|) &= \varphi(0) + \varphi'(0) + \int_0^1 (1-\theta)\varphi''(\theta) d\theta \\
&\leq \cosh(\|S_{n-1}\|)\left(1 + \beta^2 \mathbb{E}_{n-1}\left[\|\Delta S_n\|^2\right] \int_0^1 (1-\theta) e^{2\theta} d\theta\right)
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(i)}{\leq} \cosh(\|S_{n-1}\|)\left(1+\beta^2\left(\frac{e^2-3}{4}\right)\mathbb{E}_{n-1}\left[\|\Delta S_n\|^2\right]\right)\\
&\overset{(ii)}{\leq} \cosh(\|S_{n-1}\|)\left(1+\mathfrak{C}_p(\mathbb{B})\lambda_n^p(v+\|\mu-\widehat{Z}_n\|^p)\right)\\
&\overset{(iii)}{\leq} \cosh(\|S_{n-1}\|)\exp\left\{\mathfrak{C}_p(\mathbb{B})\lambda_n^p(v+\|\mu-\widehat{Z}_n\|^p)\right\},
\end{aligned}
$$

where (i) is obtained in view of $\int_0^1(1-\theta)e^{a\theta}d\theta = \frac{e^a-a-1}{a^2}$, (ii) is obtained from (3.3.8) (and also using that $\beta = 1$ in a Hilbert space), and (iii) follows from $1+u \leq e^u$ for all $u \in \mathbb{R}$. Since $n \geq 1$ was arbitrary, rearranging yields that the process defined by $\cosh(\|S_n\|)\exp\{-\mathfrak{C}_p(\mathbb{B})G_n\}$ is a nonnegative supermartingale. Noting that $\frac{1}{2}\exp(\|S_n\|) \leq \cosh(\|S_n\|)$ yields the claimed result. ■

3.3.3 Step 3: Bounding $M_n(\lambda^n)$

We now combine Lemma 3.3.2 and Proposition 3.3.3 to write down an explicit form for the supermartingale $M_n(\lambda^n)$ in (3.3.2).

Lemma 3.3.4. *Let $(X_n)_{n\geq 1}$ and $(\mathbb{B}, \|\cdot\|)$ be as in Proposition 3.3.3. Then, the process $(M_n(\lambda^n))$ defined by*

$$
M_n(\lambda^n) := \exp\left\{\left\|\widehat{\xi}_n - \mu\sum_{m\leq n}\lambda_m\right\| - (\mathfrak{C}_p(\mathbb{B}) + K_p 2^{p-1})G_n\right\},
$$

is bounded above by a nonnegative supermartingale with initial value 2.

Proof. Recall that $\widetilde{\mu}_n(\lambda) = \mathbb{E}_{n-1}[\texttt{Trunc}(\lambda Y_n)Y_n] + \widehat{Z}_n$. Applying the triangle inequality twice and Lemma 3.3.2 once, we obtain

$$
\begin{aligned}
\left\|\widehat{\xi}_n - \mu\sum_{m\leq n}\lambda_m\right\| &\leq \left\|\widehat{\xi}_n - \sum_{m\leq n}\lambda_m\widetilde{\mu}_m(\lambda_m)\right\| + \sum_{m\leq n}\lambda_m\|\widetilde{\mu}(\lambda_m) - \mu\|\\
&\leq \|S_n\| + K_p 2^{p-1}\sum_{m\leq n}\lambda_m^p(v+\|\mu-\widehat{Z}_m\|^p) = \|S_n\| + K_p 2^{p-1}G_n.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
M_n(\lambda^n) &= \exp\left\{\left\|\widehat{\xi}_n - \mu\sum_{m\leq n}\lambda_m\right\| - (\mathfrak{C}_p(\mathbb{B}) + K_p 2^{p-1})G_n\right\}\\
&\leq \exp\left\{\|S_n\| + K_p 2^{p-1}G_n - (\mathfrak{C}_p(\mathbb{B}) + K_p 2^{p-1})G_n\right\}\\
&= \exp\left\{\|S_n\| - \mathfrak{C}_p(\mathbb{B})G_n\right\},
\end{aligned}
$$

which is itself upper bounded by a nonnegative supermartingale with initial value 2 by Proposition 3.3.3. ■

We are finally ready to prove Theorem 3.2.1, which follows as a consequence of the following result.

Proposition 3.3.5. *Let $(\mathbb{B}, \|\cdot\|)$ satisfy Assumption 2 and $(X'_n)_{n\geq 1}$ satisfy Assumption 1 with respect to some filtration $(\mathcal{G}_n)_{n\geq 0}$. Suppose $\widehat{Z}$ is $\mathcal{G}_0$-measurable and there exists some function $r : (0,1) \to \mathbb{R}_{\geq 0}$ such that, for any $\delta \in (0,1]$,*

$$\mathbb{P}(\|\mu - \widehat{Z}\| \geq r(\delta)) \leq \delta. \tag{3.3.9}$$

Fix any $\delta \in (0,1]$. Decompose δ as $\delta = \delta_1 + \delta_2$ where $\delta_1, \delta_2 > 0$. Then, for any $\lambda > 0$, with probability $1-\delta$, simultaneously for all $n \geq 1$, we have:

$$\|\widehat{\mu}_n - \mu\| \leq \lambda^{p-1}(\mathfrak{C}_p(\mathbb{B}) + K_p 2^{p-1})(v + r(\delta_2)^p) + \frac{\log(2/\delta_1)}{\lambda n}, \tag{3.3.10}$$

where $\widehat{\mu}_n = \frac{1}{n}\sum_{m\leq n}\{\texttt{Trunc}(\lambda(X'_m - \widehat{Z}))(X'_m - \widehat{Z}) + \widehat{Z}\}$.

Proof. Let $B_1 = \{\exists n : M_n(\lambda^n) \geq 2/\delta_1\}$ where (M_n) is as in Lemma 3.3.4. By Ville's inequality (Section 3.1.3), $\mathbb{P}(B_1) \leq \delta_1$. Let $B_2 = \{\|\mu - \widehat{Z}\| \geq r(\delta_2)\}$. By assumption, $\mathbb{P}(B_2) \leq \delta_2$. Set $B = B_1 \cup B_2$ so that $\mathbb{P}(B) \leq \delta$. We take the sequence of predictable values (λ_n) in Lemma 3.3.4 to be constant and set $\lambda_n = \lambda > 0$ for all n. On the event B^c we have $\log(M_n(\lambda^n)) \leq \log(2/\delta_1)$ for all $n \geq 1$. That is, with probability $1-\delta$,

$$\|\widehat{\xi}_n - \lambda n \mu\| \leq (\mathfrak{C}_p(\mathbb{B}) + K_p 2^{p-1})G_n + \log(2/\delta_1), \tag{3.3.11}$$

and

$$G_n = n\lambda^p(v + \|\mu - \widehat{Z}\|^p) \leq n\lambda^p(v + r(\delta_2)^p). \tag{3.3.12}$$

Substituting (3.3.12) into (3.3.11) and dividing both sides by $n\lambda$ gives the desired result. ■

Proof of Theorem 3.2.1. Given $(X_n)_{n\geq 1}$ as in the statement of Theorem 3.2.1 apply Proposition 3.3.5 with $X'_n = X_{n+k}$ and $\mathcal{G}_n = \mathcal{F}_{n+k}$ for all $n \geq 0$. ■

3.4 Law of the Iterated Logarithm Rates

In the previous section, we derived a time-uniform, line-crossing inequality that controlled (with high probability) the deviation between a truncated mean estimator and the unknown mean. This inequality was parameterized by a scalar/truncation level λ, which, when optimized appropriately, could guarantee a width of $O(v^{1/p}(\log(1/\delta)/n)^{(p-1)/p})$ with probability at least $1-\delta$ for a preselected sample size n. However, in many settings, one may not know a target sample size in advance and may wish to observe the data sequentially and stop adaptively at a data-dependent stopping time.

To generalize our bound to an anytime-valid setting (i.e., one where the sample size is not known in advance and may be data-dependent), we use a technique known as *stitching* [75]. This involves deploying Theorem 3.2.1 once per (geometrically spaced) epoch, and then using a carefully constructed union bound to obtain coverage simultaneously for all sample sizes.

The idea is to apply Theorem 3.2.1 once per geometrically spaced epoch with different parameters k and λ in each epoch. We then take a union bound over epochs. Due to the time-uniformity of Theorem 3.2.1, the resulting estimator can be updated within the epoch,

not only at their boundaries. The bound depends on a "stitching function" h which satisfies $\sum_{j\geq 1} 1/h(j) = 1$ and a parameter η which determines the geometric spacing of the epochs, which are the intervals $[\eta^j, \eta^{j+1})$. For simplicity we take $\eta = 2$.

Theorem 3.4.1 (Stitching)**.** *Let $(\mathbb{B}, \|\cdot\|)$ satisfy Assumption 2 and $(X_n)_{n\geq 1}$ satisfy Assumption 1. Suppose that for each k, $\widehat{Z}_k$ is a $\mathcal{F}_{k-1}$-predictable estimate such that*

$$\mathbb{P}(\|\mu - \widehat{Z}_k\| \geq r(\delta, k)) \leq \delta, \tag{3.4.1}$$

for some $r : (0,1) \times \mathbb{N} \to \mathbb{R}_{\geq 0}$. Given any n, let $j_n = \lfloor \log_2(n) \rfloor$. Let $h : \mathbb{N} \to \mathbb{R}_{>0}$ satisfy $\sum_{j\geq 1} 1/h(j) \leq 1$. Fix any $\delta \in (0,1]$ and let $\widehat{\mu}_n$ be as in (3.2.2)*. Then there exist constants $(\lambda_j)_{j\geq 1}$ such that with probability $1-\delta$, simultaneously for all $n \geq 2$, we have:*

$$\|\widehat{\mu}_n(j_n, \lambda_{j_n}, \widehat{Z}_{j_n}) - \mu\| = O\left((v + r(\delta/2, j_n)^p)^{1/p} \left(\frac{\log(h(j_n)/\delta)}{n} \right)^{(p-1)/p} \right). \tag{3.4.2}$$

A few words are in order before we prove Theorem 3.4.1. As we discussed above, the idea is to apply a different estimator $\widehat{\mu}_n(j_n)$ in each epoch $[2^j, 2^{j+1})$. That is, the number of observations we set aside for the naive estimate in epoch $[2^j, 2^{j+1})$ is j. (One could replace j by any $k(j)$ where $k(j)$ grows slower than 2^j.) The bound holds for all $n \geq 2$ so that we avoid various trivialities about defining the naive estimate $\widehat{Z}_1$. Finally, note that to get iterated logarithm rates, h can be any polynomial which satisfies $\sum_{j\geq 1} h^{-1}(j) \leq 1$ (e.g., $h(j) = j(j+1)$).

Proof of Theorem 3.4.1. We will apply Theorem 3.2.1 once in every epoch $[2^j, 2^{j+1})$ for $j \geq 1$. In epoch $[2^j, 2^{j+1})$ we apply the estimator $\widehat{\mu}_n(j) = \widehat{\mu}_n(j, \lambda_j, \widehat{Z}_j)$, where $\lambda_j > 0$ is fixed. For any $\delta_j > 0$, Theorem 3.2.1 provides the guarantee that

$$\mathbb{P}(\exists n \in [2^j, 2^{j+1}) : \|\widehat{\mu}_n(j) - \mu\| \geq W(n, j)) \leq \delta_j,$$

where

$$W(n,j) = \lambda_j^{p-1}(\mathfrak{C}_p(\mathbb{B}) + K_p 2^{p-1})(v + r(\delta_j/2, j)^p) + \frac{\log(4/\delta_j)}{\lambda_j(n-j)}. \tag{3.4.3}$$

(Here the two terms above have split δ_j into $\delta_j/2 + \delta_j/2$). Let $\delta_j = \delta/h(j)$ so that $\sum_j \delta_j \leq \delta$. Note that j_n corresponds to the epoch in which n belongs, i.e., $n \in [2^{j_n}, 2^{j_n+1})$. Therefore,

$$\begin{aligned}
&\mathbb{P}\left(\exists n \geq n_0 : \|\widehat{\mu}_n(j_n) - \mu\| \geq W(n, j_n)\right) \\
&\leq \sum_{j\geq 1} \mathbb{P}(\exists n \in [2^j, 2^{j+1}) : \|\widehat{\mu}_n(j_n) - \mu\| \geq W(n, j_n)) \\
&\leq \sum_{j\geq 1} \delta_j \leq \delta.
\end{aligned}$$

It remains to select λ_j so that $W(n, j_n)$ decreases at the desired rate. Choose

$$\lambda_j = \left(\frac{\log(4/\delta_j)}{D(v + r_j^p)} \cdot \frac{\ell_j}{2^j} \right)^{1/p},$$

where $D = \mathfrak{C}_p(\mathbb{B}) + K_p 2^{p-1}$, $r_j = r(\delta_j/2, j)$, and $\ell_j = \log(1/\delta_j)$. With this choice, (3.4.3) becomes

$$W(n, j_n) = (D(v + r_{j_n}^p))^{1/p} (\log(4/\delta_{j_n}))^{1-1/p} \left(\left(\frac{\ell_{j_n}}{2^{j_n}} \right)^{1-1/p} + \left(\frac{2^{j_n}}{\ell_{j_n}} \right)^{1/p} \cdot \frac{1}{n - j_n} \right)$$

Now, since $n - j_n = n - \lfloor \log_2(n) \rfloor \geq n/2$ for $n \geq 2$, $2^{j_n} \leq n$, and $\ell_{j_n} \geq 1$, we have

$$\left(\frac{2^{j_n}}{\ell_{j_n}} \right)^{1/p} \cdot \frac{1}{n - j_n} \leq \frac{2}{n^{1-1/p}} = o(1).$$

Further, $2^{j_n} \geq n/2$ and $\log_2(n) \leq j_n + 1$, so

$$\left(\frac{\ell_{j_n}}{2^{j_n}} \right)^{1-1/p} \leq \left(\frac{2 \log(h(j_n)/\delta)}{n} \right)^{1-1/p} = O\left(\frac{\log(h(\lfloor \log_2(n) \rfloor)/\delta)}{n} \right)^{1-1/p}.$$

Noticing that $\log(4/\delta_{j_n}) = O(\log(h(\lfloor \log_2(n) \rfloor)/\delta))$ by the same reasoning, we have

$$W(n, j_n) = O\left((v + r_{j_n}^p)^{1/p} \left(\frac{\log h(\lfloor \log_2(n) \rfloor) + \log(1/\delta)}{n} \right)^{1-1/p} \right),$$

as claimed. ■

As was done with Theorem 3.2.1, one can instantiate Theorem 3.4.1 with particular estimators to achieve specific rates. For instance, if $\widehat{Z}_k$ is the plug-in mean estimate, then we can take $r(\delta, k)^p = O(\frac{v}{\delta k^{p-1}})$, so $r(\delta/2, j_n)^p = O(\frac{v}{\delta \log(n)^{p-1}}) = o(1)$. If, in addition, we take say $h(j) = j(j+1)$ for $j \geq 1$, we achieve a final rate of

$$O\left(v^{1/p} \left(\frac{\log \log n + \log(1/\delta)}{n} \right)^{1-1/p} \right), \tag{3.4.4}$$

which loses only an iterated logarithm factor compared to the line-crossing inequality presented in Section 3.2. For $p = 2$, this asymptotic width is optimal by the law of the iterated logarithm [69, 132]. For $1 < p < 2$, such a law does not necessarily exist—it depends on whether the distribution is in the domain of partial attraction of a Gaussian [92, 115]. Thus, while we cannot claim asymptotic optimality in this case, we note that our result extends and compliments recent efforts to obtain confidence sequences with iterated logarithm rates to the case of infinite variance (e.g., [75, 159, 32]).

For the purposes of constructing time-uniform bounds in practice, it's worth tracking the constants throughout the proof of Theorem 3.4.1. Doing so, we obtain a width of

$$W(n, j_n) = (D(v + r(\delta/2, j_n)^p))^{1/p} \left(\left(\frac{2 \log(h(j_n)/\delta)}{n} \right)^{1-1/p} + \frac{2}{n^{1-1/p}} \right), \tag{3.4.5}$$

where $D = \mathfrak{C}_p(B) + K_p 2^{p-1}$ and $j_n = \lfloor \log_2(n) \rfloor$.

3.5 Bound Comparison and Simulations

In the above sections, we argued that the truncated mean estimator, when appropriately optimized, could obtain a distance from the true mean of $O\big(v^{1/p}\left(\log(1/\delta)/n\right)^{(p-1)/p}\big)$ with high probability. In particular, this rate matched that of the geometric median-of-means estimator due to Minsker [121]. In this section, we study the empirical instead of theoretical performance of our bounds and estimator.

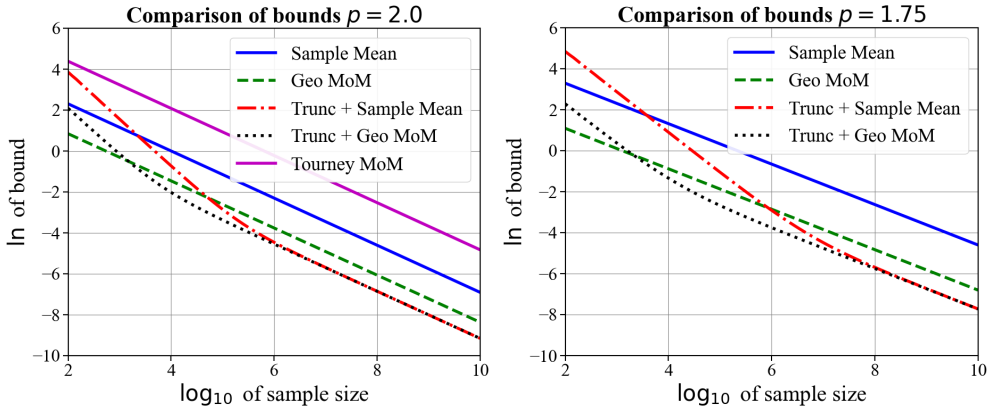
Comparing Tightness of Bounds In Figure 3.1, we compare the confidence bounds obtained for our truncation-based estimators optimized for a fix sample size (Corollary 3.2.3) against other bounds in the literature. Namely, we compare against geometric median-of-means [121], the sample mean, and (in the case a shared covariance matrix exists for observations) the tournament median-of-means estimator [112]. We plot the natural logarithm of the bounds against the logarithm base ten of the sample sizes n for $n \in [10^2, 10^{10}]$ and for $p \in \{1.25, 1.5, 1.75, 2.0\}$. We assume $\delta = 10^{-4}$ and $v = 1$. For truncation-based estimates, we assume $k = \lfloor n/10 \rfloor$ samples are used to produce the initial mean estimate and the remaining $n - k$ are used for the final mean estimate. We plot the resulting bounds for when the initial mean estimate is either computed using the sample mean or geometric median-of-means. For the tournament median-of-means estimate, we assume observations take their values in $\mathbb{R}^d$ for $d = 100$, and that the corresponding covariance matrix is the identity $\Sigma = I_d/d$.

As expected, all bounds have a slope of $-(p-1)/p$ when n is large, indicating equivalent dependence on the sample size. For all values of p, the truncation-based estimator using geometric median-of-means as an initial estimate obtains the tightest rate once moderate sample sizes are reached ($n = 10^4$ or $n = 10^5$). When $p \in \{1.25, 1.5\}$, much larger sample sizes are needed for truncation-based estimates with a sample mean initial estimate to outperform geometric median means (needing $\geq 10^{10}$ samples for $p = 1.25$). For $p = 2.0$ (i.e., finite variance) the tournament median-of-means estimate, despite achieving optimal sub-Gaussian dependence on $\lambda_{\max}(\Sigma)$ and $\mathrm{Tr}(\Sigma) = v$, performs worse than even the naive mean estimate. This is due to prohibitively large constants. These plots suggest that the truncation-based estimate is a practical and computationally efficient alternative to approaches based on median-of-means.

Performance of Estimators on Simulated Data In Figure 3.2, we examine the performance of the various mean estimators by plotting the distance between the estimates and the true mean. To do this, we sample $n = 100,000$ i.i.d. samples $X_1, \ldots, X_n \in \mathbb{R}^d$ for $d = 10$ in the following way. First, we sample i.i.d. directions $U_1, \ldots, U_n \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ from the unit sphere. Then, we sample i.i.d. magnitudes $Y_1, \ldots, Y_n \sim \mathrm{Pareto}(a)$ from the Pareto II (or Lomax) distribution with $a = 1.75$.[1] The learner then observes $X_1 = Y_1 \cdot U_1, \ldots, X_n = Y_n \cdot U_n$, and constructs either a geometric median estimate, a sample mean estimate, or a truncated mean estimate.

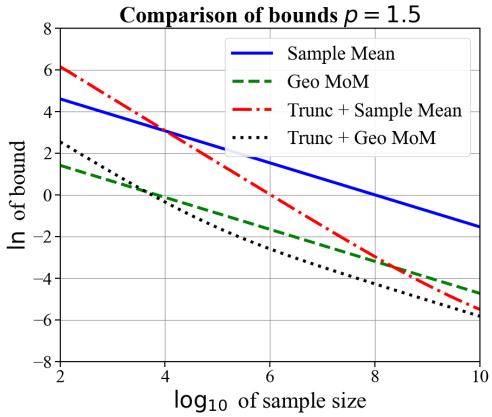
To compute the number of folds for geometric median-of-means, we follow the parameter settings outlined in Minsker [121] and assume a failure probability of $\delta = 10^{-4}$ (although we are not constructing confidence intervals, the failure probability guides how to optimize the estimator). Once again, we consider the truncated mean estimator centered at both the sample mean

[1] If $Y \sim \mathrm{Pareto}(a)$, the Y has inverse polynomial density $\propto (1+x)^{-a}$.

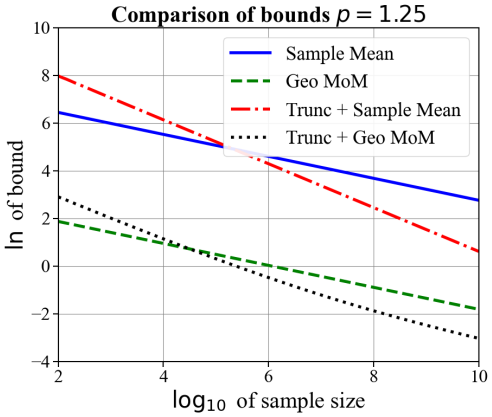


(a) $p = 2.0$

(b) $p = 1.75$

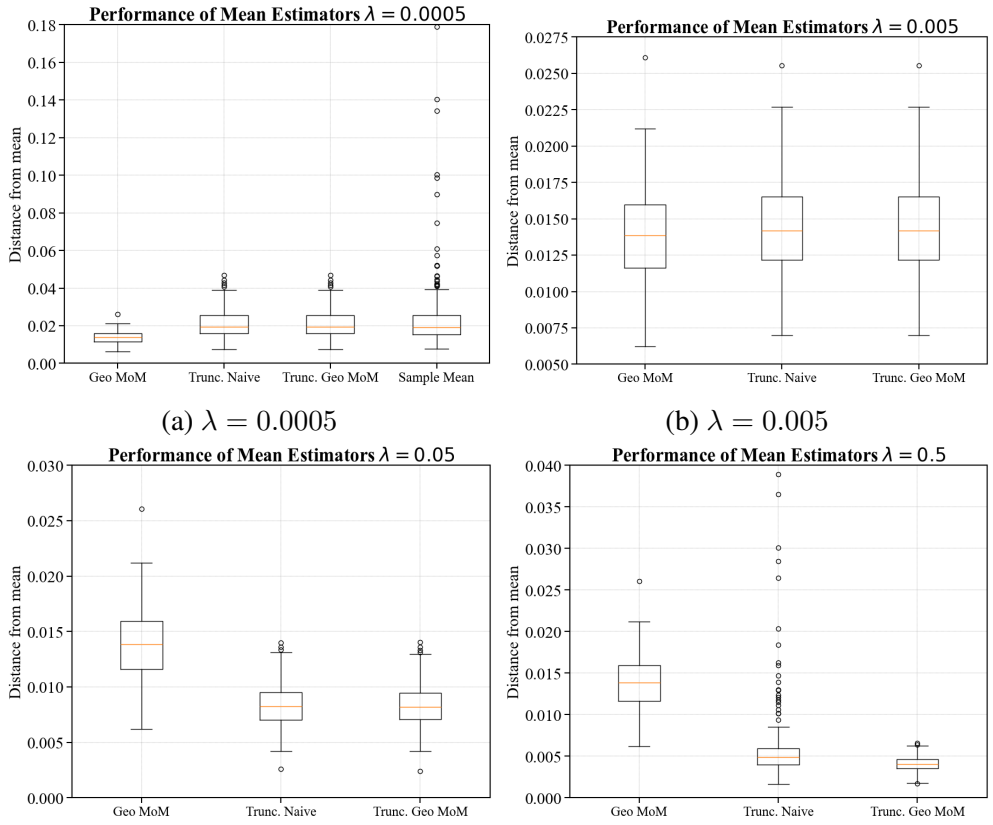


(c) $p = 1.5$



(d) $p = 1.25$

Figure 3.1: For $p \in \{1.25, 1.5, 1.75\}$, we plot the tightness of optimized bounds associated with the sample mean, geometric median-of-means (Geo-MoM), truncation with initial sample mean estimate, and truncation with initial Geo-MoM estimate. We assume $n \in [10^2, 10^{10}]$, $v = 1.0$, $\delta = 10^{-4}$, and $k = n/10$. In the case $p = 2.0$, we assume a shared covariance matrix Σ exists so we can plot the tournament median-of-means bounds assuming $\lambda_{\max}(\Sigma) = v/d$ and $d = 100$.



(a) $\lambda = 0.0005$ (b) $\lambda = 0.005$

(c) $\lambda = 0.05$ (d) $\lambda = 0.5$

Figure 3.2: We compare the empirical distributions of distance between the mean estimate and the true mean for a variety of estimators. We generate $n = 10^6$ i.i.d. samples in $\mathbb{R}^{10}$ as outlined above, and use $k = \lfloor \sqrt{10^6} \rfloor$ samples to construct initial mean estimates. We compute these estimates of 250 runs. For truncation-based estimates, we consider $\lambda \in [0.0005, 0.005, 0.05, 0.5]$. We only include the sample mean in the first plot for readability.

and a geometric median-of-means estimate. We always use $k = \lfloor \sqrt{n} \rfloor$ samples to construct the initial estimate, and produce a plot for hyperparameter $\lambda \in [0.0005, 0.005, 0.05, 0.5]$.

We construct these estimators over 250 independent runs and then construct box and whisker plots summarizing the empirical distance between the estimators and the true mean. The boxes have as a lower bound the first quartile $Q1$, in the middle the sample median M, and at the top the third quartile $Q3$. The whiskers of the plot are given by the largest and smallest point falling within $M \pm 1.5 \times (Q3 - Q2)$, respectively. All other points are displayed as outliers. We only include the sample mean in the first plot as to not compress the empirical distributions associated with other estimates.

As expected, the sample mean suffers heavily from outliers. For $\lambda \in \{0.0005, 0.005\}$ (corresponding to truncation at large radii), the geometric median-of-means estimate is roughly two times closer to the mean than either truncation-based estimate. In the setting of aggressive truncation ($\lambda \in \{0.05, 0.5\}$), the truncated mean estimator centered at the geometric median-of-means initial estimate offers a significantly smaller distance to the true mean than just geometric

median-of-means alone. The truncated estimate centered at the sample mean performs similarly for $\lambda = 0.05$, but suffers heavily from outliers when $\lambda = 0.5$. Interestingly, the recommended truncation level for optimizing tightness at $n = 100,000$ samples is $\lambda \approx 0.0004$ per Corollary 3.2.2. Our experiments reflect that one may want to truncate more aggressively than is recommended in the corollary. In practice, one could likely choose an appropriate truncation level through cross-validation.

3.6 Summary

In this work, we presented a novel analysis of a simple truncation/threshold-based estimator of a heavy-tailed mean in smooth Banach spaces, strengthening the guarantees on such estimators that currently exist in the literature. In particular, we allow for martingale dependence between observations, replace the assumption of finite variance with a finite p-th moment for $1 < p \leq 2$, and let the centered p-th moment be bounded instead of the raw p-th moment (thus making the estimator translation invariant). Our bounds are also time-uniform, meaning they hold simultaneously for all sample sizes. We provide both a line-crossing inequality that can be optimized for a particular sample size (but remains valid at all times), and a bound whose width shrinks to zero at an iterated logarithm rate. Experimentally, our estimator performs quite well compared to more computationally intensive methods such as geometric median-of-means, making it an appealing choice for practical problems.

3.A Noncentral moment bounds

For completeness, we state our bound when we assume only a bound on the raw (uncentered) p-th moment of the observations. This was the setting studied by Catoni and Giulini [23]. We replace assumption 1 with the following:

Assumption 3. *We assume $(X_n)_{n\geq 1}$ are a sequence of $\mathbb{B}$-valued random variables adapted to a filtration $(\mathcal{F}_n)_{n\geq 0}$ such that*

(1) $\mathbb{E}(X_n \mid \mathcal{F}_{n-1}) = \mu$*, for all $n \geq 1$ and some unknown $\mu \in \mathbb{B}$, and*

(2) $\sup_{n\geq 1} \mathbb{E}\left(\|X_n\|^p \mid \mathcal{F}_{n-1}\right) \leq v < \infty$ *for some known $p \in (1, 2]$ and some known constant $v > 0$.*

With only the raw moment assumption, we do not try and center our estimator. Instead we deploy $\widehat{\mu}_n(0, \lambda, 0) = \frac{1}{n}\sum_{m\leq n} \texttt{Trunc}(\lambda X_m) X_m$. With this estimator we obtain the following result, which achieves the same rate as Catoni and Giulini [23] and Chugg et al. [32].

Theorem 3.A.1. *Let $X_1, X_2, \ldots$ be random variables satisfying Assumption 3 which live in some Banach space $(\mathbb{B}, \|\cdot\|)$ satisfying Assumption 2. Fix any $\delta \in (0, 1]$. Then, for any $\lambda > 0$, with probability $1 - \delta$, simultaneously for all $n \geq 1$, we have:*

$$\|\widehat{\mu}_n(0, \lambda, 0) - \mu\| \leq 2v\lambda^{p-1}(\mathfrak{C}_p(\mathbb{B}) + K_p 2^{p-1}) + \frac{\log(2/\delta)}{\lambda n}. \tag{3.A.1}$$

Moreover, if we want to optimize the bound at a particular sample size n^* and we set

$$\lambda = \left(\frac{\log(2/\delta)}{2n^* v(\mathfrak{C}_p(\mathbb{B}) + K_p 2^{p-1})} \right)^{1/p},$$

then with probability $1 - \delta$,

$$\|\widehat{\mu}_n(0, \lambda, 0) - \mu\| \leq (2v(\mathfrak{C}_p(B) + K_p 2^{p-1}))^{1/p} \left(\frac{\log(1/\delta)}{n} \right)^{1-1/p}. \tag{3.A.2}$$

Proof. Apply Theorem 3.2.1 with $k = 0$ and $\widehat{Z}_k = 0$. Then note that we can take $r(\delta, 0) = v^{1/p}$ for all δ since $\|\mu\| \leq (\mathbb{E}\|X\|^p)^{1/p} \leq v^{1/p}$ by Jensen's inequality. ■

Part II

Applications of Martingale Concentration

Chapter 4

Fully Adaptive Composition in Differential Privacy

Composition is a key feature of differential privacy. Well-known advanced composition theorems allow one to query a private database quadratically more times than basic privacy composition would permit. However, these results require that the privacy parameters of all algorithms be fixed before interacting with the data. To address this, Rogers et al. [133] introduced fully adaptive composition, wherein both algorithms and their privacy parameters can be selected adaptively. They defined two probabilistic objects to measure privacy in adaptive composition: privacy filters, which provide differential privacy guarantees for composed interactions, and privacy odometers, time-uniform bounds on privacy loss. There are substantial gaps between advanced composition and existing filters and odometers. First, existing filters place stronger assumptions on the algorithms being composed. Second, these odometers and filters suffer from large constants, making them impractical. We construct filters that match the rates of advanced composition, including constants, despite allowing for adaptively chosen privacy parameters. En route we also derive a privacy filter for approximate zCDP. We also construct several general families of odometers. These odometers match the tightness of advanced composition at an arbitrary, preselected point in time, or at all points in time simultaneously, up to a doubly-logarithmic factor. We obtain our results by leveraging advances in martingale concentration. In sum, we show that fully adaptive privacy is obtainable at almost no loss.

4.1 Introduction

Differential privacy [55] is an algorithmic criterion that provides meaningful guarantees of individual privacy for analyzing sensitive data. Intuitively, an algorithm is differentially private if similar inputs induce similar distributions on outputs. More formally, an algorithm $A : \mathcal{X} \to \mathcal{Y}$ is differentially private if, for any set of outcomes $G \subset \mathcal{Y}$ and any *neighboring* inputs $x, x' \in \mathcal{X}$,

$$\mathbb{P}(A(x) \in G) \leq e^{\epsilon}\mathbb{P}(A(x') \in G) + \delta, \tag{4.1.1}$$

where ϵ and δ are the privacy parameters of the algorithm.

A key property of differential privacy is graceful composition. Suppose $A_1, \ldots, A_n$ are algorithms such that each A_m is (ϵ_m, δ_m)-differentially private. Advanced composition [57, 84] states that, for any $\delta' > 0$, the *composed* sequence of algorithms is (ϵ, δ)-differentially private, where $\delta = \delta' + \sum_{m \leq n} \delta_m$, and

$$\epsilon = \sqrt{2 \log\left(\frac{1}{\delta'}\right) \sum_{m \leq n} \epsilon_m^2} + \sum_{m \leq n} \epsilon_m \left(\frac{e^{\epsilon_m} - 1}{e^{\epsilon_m} + 1}\right). \tag{4.1.2}$$

When all privacy parameters are the same and small, we roughly have $\epsilon = O(\sqrt{n}\epsilon_m)$. Hence, analysts can make use of sensitive datasets with a slow degradation of privacy.

However, there is a major disconnect between most existing results on privacy composition and modern data analysis. As analysts view the outputs of algorithms, the future manner in which they interact with the data changes. Advanced composition allows analysts to adaptively select algorithms, but not privacy parameters. In many cases, analysts may wish to choose the subsequent privacy parameters based on the outcomes of the previous private algorithms. For example, if an analyst learns, from past computations, that they only need to run one more computation, they should be able to use the remainder of their privacy budget in the final round. Likewise, if an analyst is having a hard time deriving conclusions, they should be allowed to adjust privacy parameters to extend the allowable number of computations.

This desideratum has motivated the study of *fully adaptive* composition, wherein one is allowed to adaptively select the privacy parameters of the algorithms. Rogers et al. [133] define two probabilistic objects which can be used to ensure privacy guarantees in fully adaptive composition. The first, called a *privacy filter*, is an adaptive stopping condition that ensures an entire interaction between an analyst and a dataset retains a pre-specified target privacy level, even when the privacy parameters are chosen adaptively. The second, called a *privacy odometer*, provides a sequence of high-probability upper bounds on how much privacy has been lost up to any point in time. While this work took the first steps towards fully adaptive composition, their filters and odometers suffered from large constants and the latter suffered from sub-optimal asymptotic rates.

We show that, as long as a target privacy level is pre-specified, one can obtain the same rate as advanced composition, including constants. We also construct families of privacy odometers that are not only tighter than the originals, but can be optimized for various target levels of privacy. Overall, we show that full adaptivity is not a cost—but rather a feature—of differential privacy.

4.1.1 Related Work

Privacy Composition: There is a long line of work on privacy composition. The "basic composition" theorem states that, when composing private algorithms, the privacy parameters (both ϵ and δ) add up linearly [55, 54, 51]. The "advanced composition" theorem allows the total ϵ to grow sublinearly with a small degradation on δ [57]. Later work [84, 123] studies "optimal" composition, a computationally intractable formula that tightly characterizes the overall privacy of composed mechanisms.

More recently, several variants of privacy have been studied including (zero)-concentrated differential privacy (zCDP) [20, 53], Renyi differential privacy (RDP) [122], and f-differential

privacy (f-DP) [46]. These all exhibit tighter composition results than differential privacy, but for restricted classes of mechanisms. These results do not allow adaptive choices of privacy parameters.

Privacy Filters and Odometers: Rogers et al. [133] originally introduced privacy filters and odometers, which allow privacy composition with adaptively selected privacy parameters. While their contributions provide a decent approximation of advanced composition, their bounds suffer from large constants, which prevents practical usage. Our work directly improves over these initial results. First, we construct privacy filters essentially matching advanced composition. We also provide flexible families of privacy odometers that outperform those of Rogers et al. [133].

Feldman and Zrnic [61] leverage RDP to construct Rényi filters, where they require individual mechanisms to satisfy RDP. Since our proof establishes a new privacy filter for approximate zCDP [20], our results also extend to approximate RDP [127], which directly generalizes their Rényi filter. Even though it is also possible to obtain a privacy filter for (ϵ, δ)-DP through Rényi filters [61], this result requires a stronger assumption that algorithms being composed satisfy *probabilistic* (i.e. point-wise) differential privacy [85]. Since converting from differential privacy to probabilistic differential privacy can be costly (see Lemma 4.4.2), our filters demonstrate an improvement by avoiding the conversion cost.

More recently, Koskela et al. [93] and Smith and Thakurta [146] provide privacy filters for Gaussian DP (GDP) [46]. However, their results do not hold for more general mechanisms under f-DP and therefore cannot handle algorithms with rare "catastrophic" privacy failure events, in which the privacy loss goes to infinity. Both of our (ϵ, δ)-filter and approximate zCDP filters can handle such events.

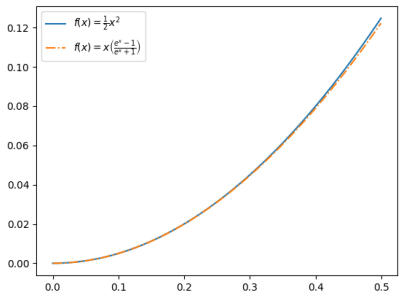
Feldman and Zrnic [61] and Lécuyer [105] construct RDP odometers. The former work sequentially composes Rényi filters and the latter work simultaneously runs multiple Rényi filters and takes a union bound. Neither odometer provides high probability, time-uniform bounds on privacy loss, making these results incomparable to our own. We believe our notion of odometers, which aligns with that of Rogers et al. [133], is more natural.

To prove our results, we leverage time-uniform concentration results for martingales [73, 76]. The bounds in these papers directly improve over related self-normalized concentration results [40, 27]. These latter bounds were leveraged in Rogers et al. [133] to construct filters and odometers.

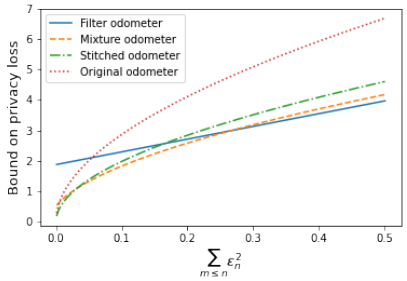
4.1.2 Summary of Contributions

In this work, we provide two primary contributions. We present these results in full rigor following a brief discussion of privacy basics and martingale theory in Section 4.2.

In Theorem 4.3.3 of Section 4.3, we construct *privacy filters* that match the rate of advanced composition, significantly improving over results of Rogers et al. [133]. Our filter follows from a more general approximate zCDP/RDP filter [20, 127] presented in Theorem 4.3.1. In particular, this approximate zCDP/RDP filter greatly generalizes existing filters from the pure RDP setting [61]. This extension allows us to capture a broader class of algorithms and avoids the conversion



(a) Comparing lower order terms



(b) Comparing privacy odometers

Figure 4.1: Figure 4.1a compares the lower order terms of advanced composition and our privacy filter. Figure 4.1b compares the original odometer of Rogers et al. [133] with our odometers (filter, mixture, and stitched).

loss when translating bounds between pure RDP and (ϵ, δ)-differential privacy. We state an informal version of filter in the case of approximate differential privacy below[1].

Informal 4.1.1 (Improved Privacy Filter)**.** *Fix target privacy parameters $\epsilon > 0$ and $\delta > 0$, and suppose $(A_n)_{n \ge 1}$ is an adaptively selected sequence of algorithms. Assume that A_n is (ϵ_n, δ_n)-DP conditioned on the outputs of the first $n-1$ algorithms, where ϵ_n and δ_n may depend on outputs of $A_1, \ldots, A_{n-1}$. If a data analyst stops interacting with the data before $\sqrt{2 \log\left(\frac{1}{\delta}\right) \sum_{m \le n+1} \epsilon_m^2} + \frac{1}{2} \sum_{m \le n+1} \epsilon_m^2 > \epsilon$, then the entire interaction is (ϵ, δ)-DP.*

In Theorem 4.4.5 of Section 4.4, we construct improved *privacy odometers* — that is, sequences of upper bounds on privacy loss which are all simultaneously valid with high probability. Our three families of odometers theoretically and empirically outperform those of Rogers et al. [133]. See Figure 4.1b for a comparison.

For both results, our key insight is to view adaptive privacy composition as depending not on the number of algorithms being composed, but rather on the sums of squares of privacy parameters, $\sum_{m \le n} \epsilon_m^2$. This shift to looking at "intrinsic time" allows us to apply recent advances in time-uniform concentration [73, 76] to privacy loss martingales. Overall, our results show that there is essentially no cost for fully adaptive private data analysis.

[1] In Appendix 4.D, we provide an alternative proof for our privacy filter result through reductions to generalized randomized response. While it gives the exact same rates, we believe it could be of independent interest. For example, it may be useful for obtaining filters with rates like the optimal composition [123, 84], which used a similar reduction to randomized response in their analysis.

4.2 Background on Differential Privacy

Throughout, we assume all algorithms map from a space of datasets $\mathcal{X}$ to outputs in a measurable space, typically either denoted $(\mathcal{Y}, \mathcal{G})$ or $(\mathcal{Z}, \mathcal{H})$. For a sequence of algorithms $(A_n)_{n\geq 1}$, we often consider the composed algorithm $A_{1:n} := (A_1, \ldots, A_n)$. For more background on measure-theoretic matters, as well as on the notion of neighboring datasets, see Appendix 4.A.

We start by formalizing a generalization of differential privacy in which the privacy parameters of an algorithm A_n can be functions of the outputs of $A_1, \ldots, A_{n-1}$. In particular, we replace the probabilities in Equation (4.1.1) with conditional probabilities given relevant random variables.

Definition 4.2.1 (Conditional Differential Privacy)**.** Suppose A and B are algorithms mapping from a space $\mathcal{X}$ to measurable spaces $(\mathcal{Y}, \mathcal{G})$ and $(\mathcal{Z}, \mathcal{H})$ respectively. Suppose $\epsilon, \delta : \mathcal{Z} \to \mathbb{R}_{\geq 0}$ are measurable functions. We say the algorithm A is (ϵ, δ)-differentially private conditioned on B if, for any neighbors $x, x' \in \mathcal{X}$ and for all measurable sets $G \in \mathcal{G}$, we have

$$\begin{aligned}&\mathbb{P}\left(A(x) \in G \mid B(x)\right)\\ &\quad\leq e^{\epsilon(B(x))}\mathbb{P}\left(A(x') \in G \mid B(x)\right) + \delta(B(x)).\end{aligned}$$

For conciseness, we will write either ϵ or $\epsilon(x)$ for $\epsilon(B(x))$ and likewise δ or $\delta(x)$ for $\delta(B(x))$.

In the nth round of adaptive composition, we will set $A := A_n$ and $B := A_{1:n-1}$. In this setting, the analyst has functions $\epsilon_n, \delta_n : \mathcal{Y}^{n-1} \to \mathbb{R}_{\geq 0}$ and takes the nth round privacy parameters to be $\epsilon_n(A_{1:n-1}(x))$ and $\delta_n(A_{1:n-1}(x))$. In other words, the analyst uses the outcome of the first $n-1$ algorithms to decide the level of privacy for the nth algorithm, ensuring that A_n is (ϵ_n, δ_n)-differentially private conditioned on $A_{1:n-1}$.

We will also leverage the notion of *zero-concentrated differential privacy (zCDP)* [20], which often provides a cleaner analysis for privacy composition. First, we will recall the definition of Rényi divergence.

Definition 4.2.2. The Rényi divergence from P to Q of order $\lambda \geq 1$ is defined as

$$D_\lambda(P\|Q) := \frac{1}{\lambda - 1}\log\left(\mathbb{E}_{Y\sim P}\left[\left(\frac{P(Y)}{Q(Y)}\right)^{\lambda-1}\right]\right).$$

The notion of zCDP bounds the Rényi divergence from $A(x)$ to $A(x')$ for any neighbors x and x'. We will focus on a conditional version of a more general definition called approximate zCDP [20, 127] that permits a small probability of unbounded Rényi divergence. The conditional approximate zCDP definition we provides uses the convex mixture formulation adapted from Papernot and Steinke [127], since it is more convenient for our proof. In Appendix 4.C.1, we will show that in the case δ and ρ are constant, this definition is equivalent to the original definition in Bun and Steinke [20].

Definition 4.2.3 (Conditional Approximate zCDP)**.** Suppose $A : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ with outputs in a measurable space $(\mathcal{Y}, \mathcal{G})$. Suppose $\delta, \rho : \mathcal{Y} \to \mathbb{R}_{\geq 0}$. We say the algorithm A satisfies conditional $\delta(z)$-approximate $\rho(z)$-zCDP if, for all $z \in \mathcal{Z}$ and any neighboring datasets x, x', there exist probability transition kernels[2] $P', P'', Q', Q'' : \mathcal{Z} \times \mathcal{G} \to [0, 1]$ such that the conditional outputs are distributed according to the following mixture distributions:

$$A(x; z) \sim (1 - \delta(z))P'(\cdot \mid z) + \delta(z)P''(\cdot \mid z)$$
$$A(x'; z) \sim (1 - \delta(z))Q'(\cdot \mid z) + \delta(z)Q''(\cdot \mid z),$$

where for all $\lambda \geq 1$, $D_\lambda(P'(\cdot \mid z) \| Q'(\cdot \mid z)) \leq \rho(z)\lambda$ and $D_\lambda(Q'(\cdot \mid z) \| P'(\cdot \mid z)) \leq \rho(z)\lambda$ for all $z \in \mathcal{Z}$.

We will also use the notions of filtration and martingales.

Filtration and Martingales: A process $(X_n)_{n\in\mathbb{N}}$ is said to be a martingale with respect to a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ if, for all $n \in \mathbb{N}$, (a) X_n is $\mathcal{F}_n$-measurable, (b) $\mathbb{E}|X_n| < \infty$, and (c) $\mathbb{E}(X_n \mid \mathcal{F}_{n-1}) = X_{n-1}$. Correspondingly, $(X_n)_{n\in\mathbb{N}}$ is a supermartingale if $\mathbb{E}(X_n \mid \mathcal{F}_{n-1}) \leq X_{n-1}$. In our context, we will consider the natural filtration $(\mathcal{F}_n(x))_{n\in\mathbb{N}}$ generated by $(A_n(x))_{n\geq 1}$. In our proofs, we construct the appropriate (super)martingales so that we can leverage the optional stopping theorem and time-uniform concentration to obtain privacy filters and odometers [157, 73, 76]. We present a full exposition of the mathematical tools in Appendix 4.A and 4.B.

4.3 Privacy Filters

We now provide our main results on privacy filters. In general, a privacy filter is a function N that takes the privacy parameters of a sequence of private algorithms as input and decides to stop at some point so that the composition of these algorithms satisfies a pre-specified level of privacy. We will first present a privacy filter for approximate zCDP (Theorem 4.3.1), which will immediately imply the privacy filter result for (ϵ, δ)-DP (Theorem 4.3.3). Since approximate zCDP bounds Rényi divergence of all orders λ, our proof for Theorem 4.3.1 also directly implies a privacy fiter for approximate RDP [127], which generalizes the RDP filter by Feldman and Zrnic [61].

Our (ϵ, δ)-DP filter improves on the rate of the original filter presented in Rogers et al. [133] and matches the rate of advanced composition that requires pre-fixed choices of privacy parameters. Even though it is also possible to obtain an (ϵ, δ)-DP filter through the result of Feldman and Zrnic [61], our privacy filters avoid their conversion costs and provide a tighter bound.[3]

We can now state our general privacy filter in terms of approximate zCDP.

Theorem 4.3.1 (Approximate zCDP filter)**.** *Let $(A_n)_{n\geq 1}$ be an adaptive sequence of algorithms, where $A_n : \mathcal{X} \times \mathcal{Y}^{n-1} \to \mathcal{Y}$. Assume that $\delta_n, \rho_n : \mathcal{Y}^{n-1} \to \mathbb{R}_{\geq 0}$. For any $n \geq 1$, assume that*

[2] A *probability transition kernel* $P' : \mathcal{Z} \times \mathcal{G} \to [0, 1]$ is a mapping such that $P(\cdot \mid z) : \mathcal{G} \to [0, 1]$ is a probability measure for each $z \in \mathcal{Z}$.

[3] Feldman and Zrnic [61, Section 4.3] apply Rényi filters to algorithms which satisfy (conditional) probabilistic differential privacy (pDP). In general, a lossy conversion from (ϵ, δ)-DP to (ϵ, δ)-pDP is required to apply their filter.

$A_n(\cdot; y_{1:n-1})$ is conditionally $\delta_n(y_{1:n-1})$-approximate $\rho_n(y_{1:n-1})$-zCDP for any prior outcomes $y_{1:n-1}$. We define the function $N : \mathcal{Y}^\infty \to \mathbb{N}$ where

$$N(y_1, y_2, \cdots) = \inf\left\{n : \sum_{m=1}^{n+1} \rho_m(y_{1:m-1}) > \rho\right\} \wedge \inf\left\{n : \sum_{m=1}^{n+1} \delta_m(y_{1:m-1}) > \delta\right\}.$$

Then $A_{1:N(\cdot)}(\cdot)$ is δ-approximate ρ-zCDP, where $N(x) = N((A_n(x))_{n\geq 1})$.

We note that the argument used to prove the above theorem immediately implies a privacy filter for approximate RDP, and thus Theorem 4.3.1 can be viewed as a strict generalization of the work of Feldman and Zrnic [61]. Further, Theorem 4.3.1 implies a privacy filter under (ϵ, δ)-differential privacy. To show this implication, we will use the following conversion results.

Lemma 4.3.2 ([20])**.** *If A satisfies (ϵ, δ)-DP, then A satisfies δ-approximate $\frac{1}{2}\epsilon^2$-zCDP. If A satisfies δ-approximate ρ-zCDP, then A satisfies $(\rho + 2\sqrt{\rho \ln(1/\delta')}, \delta + (1-\delta)\delta')$-DP.*

We can now obtain our (ϵ, δ)-privacy filter by a conversion of individual approximate differential privacy parameters to approximate zCDP ones, application of the approximate zCDP filter, and the conversion of approximate zCDP back to approximate differential privacy.

Theorem 4.3.3 ((ϵ, δ)-DP filter)**.** *Suppose $(A_n)_{n\geq 1}$ is a sequence of algorithms such that, for any $n \geq 1$, A_n is (ϵ_n, δ_n)-differentially private conditioned on $A_{1:n-1}$. Let $\epsilon > 0$ and $\delta = \delta' + \delta''$ be target privacy parameters such that $\delta' > 0$ and $\delta'' \geq 0$. We define the function $N : \mathcal{Y}^\infty \to \mathbb{N}$ where*

$$N(y_1, y_2, \cdots) = \inf\left\{n : \sum_{m=1}^{n+1} \epsilon_\ell^2(y_{1:m-1})/2 > \rho\right\} \wedge \inf\left\{n : \sum_{m=1}^{n+1} \delta_m(y_{1:m-1}) > \delta\right\}.$$

Then, the algorithm $A_{1:N(\cdot)}(\cdot)$ is $(\rho + 2\sqrt{\rho \log(1/\delta)}, \delta)$-DP, where $N(x) := N((A_n(x))_{n\geq 1})$.

Proof of Theorem 4.3.1. In our proof, we assume that $\sum_{n=1}^{\infty} \delta_n(y_{1:n-1}) \leq \delta$ for all sequences $(y_n)_{n\geq 1}$ without loss of generality. Let $P_{1:n}$ and $Q_{1:n}$ denote the joint distributions of $(A_1, \ldots, A_n)$ with inputs x and x', respectively. We overload notation and write $P_{1:n}(y_1, \ldots, y_n)$ and $Q_{1:n}(y_1, \ldots, y_n)$ for the likelihood of $y_1, \ldots, y_n$ under input x and x' respectively. We similarly write $P_n(y_n \mid y_{1:n-1})$ and $Q_n(y_n \mid y_{1:n-1})$ for the corresponding conditional densities.

By Bayes rule, for any $n \in \mathbb{N}$, we have

$$P_{1:n}(y_1, \cdots, y_n) = \prod_{m=1}^{n} P_m(y_m \mid y_{1:m-1}),$$
$$Q_{1:n}(y_1, \cdots, y_n) = \prod_{m=1}^{n} Q_m(y_m \mid y_{1:m-1}).$$

By our assumption of approximate zCDP at each step n, we can write the conditional likelihoods of P_n and Q_n as the following convex combinations:

$$P_n(y_n \mid y_{1:n-1}) = (1 - \delta_n(y_{1:n-1}))P'_n(y_n \mid y_{1:n-1}) + \delta_n(y_{1:n-1})P''_n(y_n \mid y_{1:n-1}),$$

$$Q_n(y_n \mid y_{1:n-1}) = (1 - \delta_n(y_{1:n-1}))Q'_n(y_n \mid y_{1:n-1}) + \delta_n(y_{1:n-1})Q''_n(y_n \mid y_{1:n-1}),$$

such that for all $\lambda \geq 1$ and all prior outcomes $y_{1:n-1}$, we have both

$$D_\lambda\left(P'_n(\cdot \mid y_{1:n-1}) \,\|\, Q'_n(\cdot \mid y_{1:n-1})\right) \leq \rho_n(y_{1:n-1})\lambda, \tag{4.3.1}$$
$$D_\lambda\left(Q'_n(\cdot \mid y_{1:n-1}) \,\|\, P'_n(\cdot \mid y_{1:n-1})\right) \leq \rho_n(y_{1:n-1})\lambda. \tag{4.3.2}$$

Now, from Lemma 4.C.5, we can then write these distributions as a convex combination of "good" distributions for which Rényi divergence is small, and "bad" distributions for which the divergence may be unbounded. In more detail, using the assumption that $\sum_{n=1}^{\infty} \delta_n(y_{1:n-1}) \leq \delta$ for all seqeunces $(y_n)_{n\geq 1}$, we have, for all $n \geq 1$,

$$P_{1:n}(y_1, \cdots, y_n) = (1-\delta) \underbrace{\prod_{m=1}^{n} P'_m(y_m|y_{1:m-1})}_{P'_{1:n}(y_1,\cdots,y_n)} + \delta P''_{1:n}(y_1, \cdots, y_n) \tag{4.3.3}$$

$$Q_{1:n}(y_1, \cdots, y_n) = (1-\delta) \underbrace{\prod_{m=1}^{n} Q'_m(y_m|y_{1:m-1})}_{Q'_{1:n}(y_1,\cdots,y_n)} + \delta Q''_{1:n}(y_1, \cdots, y_n). \tag{4.3.4}$$

From the above, if $N : \mathcal{Y}^\infty \to \mathbb{N}$ is the time outlined in the theorem statement, it follows that the joint densities[4] $P_{1:N}$ of $A_1(x), \cdots A_{N(x)}(x)$ and $Q_{1:N}$ of $A_1(x'), \cdots A_{N(x')}(x')$, and both can be written as a convex combination of distributions $(P'_{1:N}, P''_{1:N})$ and $(Q'_{1:N}, Q''_{1:N})$:

$$P_{1:N}(y_1, y_2, \cdots, y_N) = (1-\delta) \underbrace{\prod_{n=1}^{N} P'_n(y_n|y_{1:n-1})}_{P'_{1:N}(y_1,y_2,\cdots,y_N)} + \delta P''_{1:N}(y_1, y_2, \cdots, y_N)$$

$$Q_{1:N}(y_1, y_2, \cdots, y_N) = (1-\delta) \underbrace{\prod_{n=1}^{N} Q'_n(y_n|y_{1:n-1})}_{Q'_{1:N}(y_1,y_2,\cdots,y_N)} + \delta Q''_{1:N}(y_1, y_2, \cdots, y_N)$$

In the above, we notate quantities in terms of "N" instead of "$N(x)$" or "$N(x')$" since N only depends on the underlying dataset x or x' *through* the observed sequence of iterates $(y_n)_{n\geq 1}$.

What remains now is to bound the Rényi divergence between P'_N and Q'_N. We do this using an optional stopping argument for non-negative supermartingales (Lemma 4.B.1). Suppose $(Y'_n)_{n\geq 1}$ is a process whose nth finite-dimensional distribution is given by P'_n. For any fixed $\lambda \geq 1$, define the process $(M_n^{(\lambda)})_{n\geq 0}$ by:

$$M_n^{(\lambda)} := \exp\left\{(\lambda - 1) \sum_{m \leq n} \left[\log\left(\frac{P'_m(Y'_m \mid Y'_{1:m-1})}{Q'_m(Y'_m \mid Y'_{1:m-1})}\right) - \lambda\rho_m(Y'_{1:m-1})\right]\right\}. \tag{4.3.5}$$

[4]We ignore measure-theoretic concerns about specifying which dominating measures these densities are defined with respect to.

It is clear that $M_n^{(\lambda)}$ is a non-negative supermartingale with respect to natural filtration $(\mathcal{F}'_n)_{n\geq 1}$ given by $\mathcal{F}'_n := \sigma(Y'_m : m \leq n)$, a fact that we confirm in Lemma 4.C.4. We emphasize that $(\mathcal{F}'_n)_{n\geq 1}$ is not in fact the data generating filtration, but rather a tool used for theoretical analysis. In more detail, we consider this filtration because, heuristically, approximate zCDP aims at bounding the moment generating function of a "good" portion of the joint distribution — the true joint distribution may allow some probability of catastrophic failure (i.e. unbounded privacy loss). We adopt the same convention that $N := N(y_1, y_2, \dots)$ with the explicit values of $(y_n)_{n\geq 1}$ clear from context. Observe that $N((Y'_n)_{n\geq 1})$ is a stopping time with respect to $(\mathcal{F}'_n)_{n\geq 0}$. We now invoke optional stopping (Lemma 4.B.1), which yields

$$\begin{aligned}
\mathbb{E}[M^{(\lambda)}_{N(Y'_1,Y'_2,\dots)}] \leq 1 &\implies \mathbb{E}\left[\exp\left((\lambda-1)\sum_{n\leq N(Y'_1,Y'_2,\cdots)}\left\{\log\left(\frac{P'_n(Y'_n \mid Y'_{1:n-1})}{Q'_n(Y'_n \mid Y'_{1:n-1})}\right) - \lambda\rho_n(Y'_{1:n-1})\right\}\right)\right] \leq 1 \\
&\implies \mathbb{E}\left[\exp\left((\lambda-1)\sum_{n\leq N(Y'_1,Y'_2,\cdots)}\log\left(\frac{P'_n(Y'_n \mid Y'_{1:n-1})}{Q'_n(Y'_n \mid Y'_{1:n-1})}\right)\right)\right] \leq e^{\lambda(\lambda-1)\rho} \\
&\implies \mathbb{E}\left[\exp\left((\lambda-1)\log\left(\frac{P'_{1:N}(Y'_{1:N})}{Q'_{1:N}(Y'_{1:N})}\right)\right)\right] \leq e^{\lambda(\lambda-1)\rho}.
\end{aligned}$$

What we have just showed is precisely that

$$D_\lambda\left(P'_{1:N} \mid Q'_{1:N}\right) \leq \rho\lambda,$$

which is precisely the desired result. A symmetric argument yields an identical bound on $D_\lambda(Q'_{1:N} \mid P'_{1:N})$. Thus, we have showed the desired result. ■

4.4 Privacy Odometers

Previously, we constructed privacy filters that matched the rate of advanced composition while allowing *both* algorithms and privacy parameters to be chosen adaptively. While privacy filters require the total level of privacy to be fixed in advance, it is desirable to track the privacy loss at all steps without a pre-fixed budget [109]. We now study privacy odometers which provide sequences of upper bounds on accumulated privacy loss that are valid at all points in time simultaneously with high probability.

4.4.1 Background on Privacy Loss and Odometers

To formally introduce privacy odometers, we will first revisit the notion of *privacy loss*, which measures how much information is revealed about the underlying input dataset. For neighbors $x, x' \in \mathcal{X}$, let p^x and $p^{x'}$ be the densities of $A(x)$ and $A(x')$ respectively. The privacy loss between $A(x)$ and $A(x')$ is defined as

$$\mathcal{L}(x, x') := \log\left(\frac{p^x(A(x))}{p^{x'}(A(x))}\right). \tag{4.4.1}$$

By Equation (4.4.1), a negative privacy loss suggests that the input is more likely to be x', and likewise a positive privacy loss suggests that the input is more likely to be x. We now generalize privacy loss to its conditional counterpart.

Definition 4.4.1 (Conditional Privacy Loss)**.** Suppose A and B are as in Definition 4.2.1. Suppose $x, x' \in \mathcal{X}$ are neighbors. Let $p^x(\cdot|\cdot), p^{x'}(\cdot|\cdot) : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ be conditional densities for $A(x)$ and $A(x')$ respectively given $B(x)$.[5] The privacy loss between $A(x)$ and $A(x')$ conditioned on B is given by

$$\mathcal{L}_B(x, x') := \log\left(\frac{p^x(A(x)|B(x))}{p^{x'}(A(x)|B(x))}\right).$$

Suppose A_n is the nth algorithm being run and we have already observed $A_{1:n-1}(x)$ for some unknown input $x \in \mathcal{X}$. If we are trying to guess whether x or a neighbor x' produced the data, we would consider the privacy loss between $A_n(x)$ and $A_n(x')$ conditioned on $A_{1:n-1}(x)$. It is straightforward to characterize the privacy loss of a composed algorithm $A_{1:n}$ in terms of the privacy loss of each constituent algorithm $A_1, \cdots, A_n$. Namely, from Bayes rule,

$$\mathcal{L}_{1:n}(x, x') = \sum_{m \leq n} \mathcal{L}_m(x, x'), \tag{4.4.2}$$

where $\mathcal{L}_m(x, x')$ is shorthand for the conditional privacy loss between $A_m(x)$ and $A_m(x')$ given $A_{1:m-1}(x)$, per Definition 4.4.1. Equation (4.4.2) also holds at arbitrary random times $N(x)$ that only depend on the dataset $x \in \mathcal{X}$ through observed algorithm outputs.

The simple decomposition of privacy loss noted above motivates the study of an "alternative", probabilistic definition of differential privacy. Intuitively, an algorithm should be differentially private if, with high probability, the privacy loss is small. More formally, an algorithm $A : \mathcal{X} \to \mathcal{Y}$ is said to be (ϵ, δ)-*probabilistically differentially private*, or (ϵ, δ)-pDP for short, if, for all neighboring inputs $x, x' \in \mathcal{X}$, we have $\mathbb{P}(|\mathcal{L}(x, x')| > \epsilon) \leq \delta$. In the previous line (as well as in the remainder of the section), the randomness in $\mathcal{L}(x, x')$ comes from the randomized algorithm A.

Unfortunately, as noted by Kasiviswanathan and Smith [85] (in which pDP is called *pointwise indistinguishability*), pDP is a strictly stronger notion than DP. In particular, if an algorithm is (ϵ, δ)-pDP, it is also (ϵ, δ)-DP. The converse in general requires a costly conversion.

Lemma 4.4.2 (Conversions between DP and pDP [85])**.** *If A is (ϵ, δ)-pDP, then A is also (ϵ, δ)-DP. Conversely, if A is (ϵ, δ)-DP, then A is $(2\epsilon, \frac{2\delta}{\epsilon e^\epsilon})$-pDP.*

We note that that Guingona et al. [66] have recently shown that other possible conversion rates from probabilistic differential privacy to approximate differential privacy are possible. However, we note that these conversions require trading off tightness in the approximation parameter ϵ and the approximation parameter δ. In particular, a fully tight conversion from probabilistic differenial privacy to approximate differential privacy is not possible. We will work with the conditional counterpart of probabilistic differential privacy (pDP).

[5] To ensure the existence of conditional densities, it suffices to assume that $\mathcal{Y}$ and $\mathcal{Z}$ are *Polish spaces* under some metrics $d_{\mathcal{Y}}$ and $d_{\mathcal{Z}}$, and that $\mathcal{G}$ and $\mathcal{H}$ are the corresponding Borel σ-algebras associated with $d_{\mathcal{Y}}$ and $d_{\mathcal{Z}}$ [50]. These measurability assumptions are not restrictive, as Euclidean spaces, countable spaces, and Cartesian products of the two satisfy these assumption.

Definition 4.4.3 (Conditional Probabilistic Differential Privacy)**.** Suppose $A : \mathcal{X} \to \mathcal{Y}$ and $B : \mathcal{X} \to \mathcal{Z}$ are algorithms, and $\epsilon, \delta : \mathcal{Z} \to \mathbb{R}_{\geq 0}$ are measurable. Then, A is said to be (ϵ, δ)-*probabilistically differentially private conditioned on* B if, for any neighbors $x, x' \in \mathcal{X}$, we have

$$\mathbb{P}\left(|\mathcal{L}_B(x, x')| > \epsilon(B(x))|B(x)\right) \leq \delta(B(x)).$$

While in Theorem 4.3.3 we assumed that the algorithms being composed were *conditionally differentially private*, here, we need to assume *conditional probabilistic privacy*. This is because our goal is not differential privacy, but rather tight control over privacy loss. We conjecture that a version of our privacy odometer (in Theorem 4.4.5) that replaces pDP by DP and leaves all else identical does not hold. Our intuition for this conjecture is that there exist simple examples of algorithms satisfying (ϵ, δ)-DP that don't satisfy (ϵ, δ)-pDP (see Appendix 4.F, for instance). We believe that, by sequentially composing such algorithms and using anti-concentration results, one can show that some odometers fail to be valid. We leave this as potential future work. In sequential composition, we would assume the nth algorithm A_n is (ϵ_n, δ_n)-pDP conditioned on $A_{1:n-1}$. The privacy parameters would be given as functions of $A_{1:n-1}(x)$. Now we state the definition of privacy odometer, which provides bounds on privacy loss under arbitrary stopping conditions (e.g. conditions based on model accuracy).

Definition 4.4.4 (Privacy Odometer [133])**.** Let $(A_n)_{n\geq 1}$ be an adaptive sequence of algorithms such that, for all $n \geq 1$, A_n is (ϵ_n, δ_n)-pDP conditioned on $A_{1:n-1}$. Let $(u_n)_{n\geq 1}$ be a sequence of functions where $u_n : \mathbb{R}_{\geq 0}^{n-1} \times \mathbb{R}_{\geq 0}^{n-1} \to \mathbb{R}_{\geq 0}$. Let $\delta \in (0, 1)$ be a target confidence parameter. For $x \in \mathcal{X}, n \geq 1$, define $U_n(x) := u_n(\epsilon_{1:n-1}(x), \delta_{1:n-1}(x))$. Then, $(u_n)_{n\geq 1}$ is called a δ-privacy odometer if, for all $x, x' \in \mathcal{X}$ neighbors, we have

$$\mathbb{P}\left(\exists n \geq 1 : \mathcal{L}_{1:n}(x, x') > U_n(x)\right) \leq \delta.$$

4.4.2 Improved Privacy Odometers

We construct our privacy odometers in Theorem 4.4.5. Our technical centerpiece is time-uniform concentration inequalities for martingales [157, 73, 76]. For a martingale $(M_n)_{n\in\mathbb{N}}$ and confidence level $\delta > 0$, time-uniform concentration inequalities provides bounds $(U_n)_{n\in\mathbb{N}}$ satisfying $\mathbb{P}(\exists n \in \mathbb{N} : M_n > U_n) \leq \delta$. Thus, if we can create a martingale from privacy loss, we can use time-uniform concentration to construct odometers. Our proof first considers the case where each A_n is $(\epsilon_n, 0)$-pDP and the *privacy loss martingale* $(M_n)_{n\in\mathbb{N}}$ [57] is given by $M_0 = 0$ and:

$$M_n := M_n(x, x') := \mathcal{L}_{1:n}(x, x') - \sum_{m\leq n} \mathbb{E}\left(\mathcal{L}_m(x, x')|\mathcal{F}_{n-1}(x)\right) \tag{4.4.3}$$

We then extend to the case of $\delta_n \geq 0$ via conditioning.

To construct their filters and odometers, Rogers et al. [133] use self-normalized concentration inequalities [40, 27]. We instead use advances in time-uniform martingale concentration [73, 76], which yields tighter results.

Theorem 4.4.5. *Suppose $(A_n)_{n\geq 1}$ is a sequence of algorithms such that, for any $n \geq 1$, A_n is (ϵ_n, δ_n)-pDP conditioned on $A_{1:n-1}$. Let $\delta = \delta' + \delta''$ be a target approximation parameter such that $\delta' > 0, \delta'' \geq 0$. Define $N := N((\delta_n)_{n\geq 1}) := \inf\left\{n \in \mathbb{N} : \delta'' < \sum_{m\leq n+1} \delta_m\right\}$ and $V_n := \sum_{m\leq n} \epsilon_m^2$. Define the following:*

1. ***Filter odometer.*** *For any $\epsilon > 0$, let $y^* := \left(-\sqrt{2\log\left(\frac{1}{\delta'}\right)} + \sqrt{2\log\left(\frac{1}{\delta'}\right) + \epsilon}\right)^2$. Define functions $(u_n^F)_{n\geq 1}$ by*

$$u_n^F(\epsilon_{1:n}, \delta_{1:n}) := \begin{cases} \infty & n > N \\ \frac{\sqrt{2y^*\log\left(\frac{1}{\delta'}\right)}}{2} + \frac{\sqrt{2\log\left(\frac{1}{\delta'}\right)}}{2\sqrt{y^*}} V_n + \frac{1}{2}V_n & \textit{otherwise}. \end{cases}$$

2. ***Mixture odometer.*** *For any $\gamma > 0$, define the sequence of functions $(u_n^M)_{n\geq 1}$ by*

$$u_n^M(\epsilon_{1:n}, \delta_{1:n}) := \begin{cases} \infty & n > N \\ \sqrt{2\log\left(\frac{1}{\delta'}\sqrt{\frac{V_n+\gamma}{\gamma}}\right)(\gamma + V_n)} + \frac{1}{2}V_n & \textit{otherwise}. \end{cases}$$

3. ***Stitched odometer.*** *For any $v_0 > 0$, define the sequence of functions $(u_n^S)_{n\geq 1}$ by*

$$u_n^S(\epsilon_{1:n}, \delta_{1:n}) := \begin{cases} \infty & n > N \textit{ or } V_n < v_0 \\ 1.7\sqrt{V_n\left(\log\log\left(\frac{2V_n}{v_0}\right) + 0.72\log\left(\frac{5.2}{\delta'}\right)\right)} + \frac{1}{2}V_n & \textit{else}. \end{cases}$$

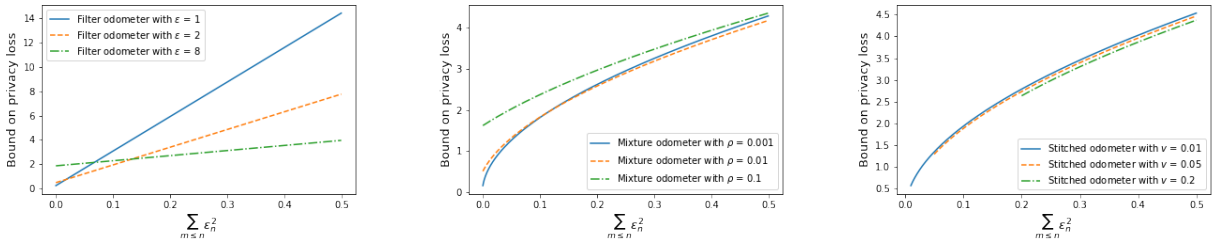
Then, any of the sequences $(u_n^F)_{n\geq 1}$, $(u_n^M)_{n\geq 1}$, or $(u_n^S)_{n\geq 1}$ is a δ-privacy odometer.

The proof of Theorem 4.4.5 can be found in Appendix 4.E. We now provide intuition for our odometers, which are plotted in Figure 4.3. Our insight is to view odometers not as functions of the number of algorithms being composed, but rather as functions of the intrinsic time $\sum_{m\leq n}\epsilon_m^2$. This reframing allows us to leverage the various time-uniform concentration inequalities discussed in Appendix 4.B. The filter odometer is the tightest odometer when the value $\sum_{m\leq n}\epsilon_m^2$ is close to a fixed accumulated variance y^*, but the tightness drops off precipitously when $\sum_{m\leq n}\epsilon_m^2$ is far from y^*. The mixture odometer, which is named after the *the method of mixtures* [132, 45, 76], sacrifices tightness at any fixed point in time to obtain overall tighter bounds on privacy loss. This odometer can be numerically optimized, in terms of ρ, for tightness at a predetermined value $\sum_{m\leq n}\epsilon_m^2$. The stitched odometer, whose name derives from Theorem 4.B.4, is similarly tight across time. This odometer requires that $\sum_{m\leq n}\epsilon_m^2$ exceed some pre-selected "variance" v_0 before becoming nontrivial (i.e. finite). Larger values of v_0 will yield tighter odometers, albeit at the cost of losing bound validity when accumulated variance is small. With this intuition, we can compare our odometers to the original presented in Rogers et al. [133].

Lemma 4.4.6 (Theorem 6.5 in Rogers et al. [133])**.** *Assume the same setup as Theorem 4.4.5, and fix $\delta = \delta' + \delta''$, where $\frac{1}{e} \geq \delta' > 0$ and $\delta'' \geq 0$. Define the sequence of functions $(u_n^R)_{n\geq 1}$ by*

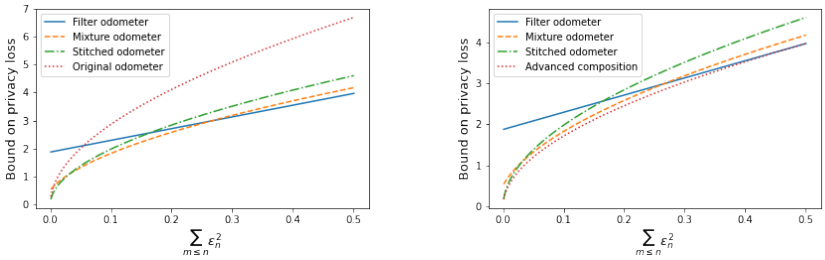
$$u_n^R(\epsilon_{1:n}, \delta_{1:n}) := \begin{cases} \infty, & n > N \\ \sqrt{2V_n\left(\log(110e) + 2\log\left(\frac{\log(|x|)}{\delta'}\right)\right)} + \frac{1}{2}V_n & n \leq N, V_n \in \left[\frac{1}{|x|^2}, 1\right] \\ \sqrt{2\left(\frac{1}{|x|^2} + V_n\right)\left(1 + \frac{1}{2}\log\left(1 + |x|^2 V_n\right)\right)\log\log\left(\frac{4}{\delta'}\log_2(|x|)\right)} & +\frac{1}{2}V_n, \\ & \textit{otherwise} \end{cases},$$

where $|x|$ denotes the number of elements in dataset x. Then, $(u_n^R)_{n\geq 1}$ is a δ-privacy odometer.



(a) Comparing filter odometers (b) Comparing mixture odometers (c) Comparing stitched odometers

Figure 4.2: Comparison of filter, mixture, and stitched odometers plotted as functions of $\sum_{m \leq n} \epsilon_m^2$. We set $\delta' = 10^{-6}$ and assume all algorithms being composed are purely differentially private for simplicity.



(a) New odometers vs. original (b) New odometers vs. pointwise advanced composition

Figure 4.3: Figure 4.3a compares our odometers to the original. Figure 4.3b compares them with advanced composition optimized point-wise. The curve plotted for advanced composition is valid at any fixed time, but not uniformly over time. Our odometers nevertheless provide a close approximation.

Our new odometers improve over the one presented in Lemma 4.4.6. First, the above odometer has an explicit dependence on dataset size. In learning settings, datasets are large, degrading the quality of the odometer. Secondly, the tightness of the odometer drops off outside of the interval $\left[\frac{1}{|x|^2}, 1\right]$. If *any* privacy parameter of an algorithm being composed exceeds 1, the bound becomes significantly looser. Lastly, and perhaps most simply, the form of the odometer is complicated. Our odometers all have relatively straightforward dependence on the intrinsic time $\sum_{m \leq n} \epsilon_m^2$.

We now examine the rates of all odometers. For simplicity, let $v := \sum_{m \leq n} \epsilon_m^2$. The stitched odometer has a rate of $O(\sqrt{v \log \log(v)})$ in its leading term, asymptotically matching the law of the iterated logarithm [132] up to constants. Both the original privacy odometer and the mixture odometer have a rate of $O(\sqrt{v \log(v)})$, demonstrating worse asymptotic performance. The filter odometer has the worst asymptotic performance, growing linearly as $O(v)$. This does not mean the stitched odometer is the best odometer, since target levels of privacy are often kept small.

To empirically compare odometers, it suffices to consider the setting of *pure* differential pri-

vacy, as the odometers identically depend on $(\delta_n)_{n\geq 1}$. Each presented odometer can be viewed as a function of v, allowing us to compare odometers by plotting their values for a continuum of v. Figure 4.3a shows that there is no clearly tightest odometer. All odometers, barring the original, dominate for some window of values of v. While the stitched odometer is asymptotically best, the mixture odometer is tighter for small values of v. Likewise, if one knows an approximate target privacy level, the filter odometer is tightest. This behavior is expected from our understanding of martingale concentration [73, 76]: there is no uniformly tightest boundary containing (with probability $1-\delta$) the entire path of a martingale; boundaries that are tight early must be looser later, and vice versa. In fact, we conjecture that our bounds are essentially unimprovable in general — this conjecture stems from the fact that the time-uniform martingale boundaries employed have error probability *essentially* equal to δ, which in turn stems from the deep fact that for continuous-path (and thus continuous-time) martingales, Ville's inequality (Theorem 1.0.2)—that underlies the derivation of these boundaries—holds with exact equality. Since we operate in discrete-time, the only looseness in Ville's inequality stems from lower-order terms that reflect the possibility that at the stopping time, the value of the stopped martingale may not be *exactly* the value at the boundary.

In Figure 4.3b, we compare our odometers with advanced composition optimized in a pointwise sense for all values of v simultaneously. This boundary *is not a valid odometer*, as advanced composition only holds at a prespecified point in intrinsic time v. Our odometers are almost tight with advanced composition for the values of v plotted. Our filter odometer lies tangent to the advanced composition curve, as expected from Section 5.2 of Howard et al. [73].

4.5 Future Directions

There are many open problems related to fully adaptive composition. For example, even though privacy filters have been studied under the notion of Gaussian DP [146, 93], privacy filters and odometers have not been studied for general f-DP [46]. It also has not been investigated whether adaptivity in privacy parameter selection improves the performance of iterative algorithms such as private SGD. Intuitively, it should be beneficial to let the iterates of an algorithm guide future choices of privacy parameters. Optimal composition results [84, 123, 170] have yet to be considered in a setting where privacy parameters are adaptively selected. In Appendix 4.D, we provide another proof of Theorem 4.3.3, which leverages a reduction of private algorithms to generalized randomized response. Since such a reduction was used in the proofs of Kairouz et al. [84] and Murtagh and Vadhan [123], we believe this proof can be useful for optimal composition with adaptively chosen privacy parameters.

4.A Measure-Theoretic Formalism

Below, we provide some measure-theoretic formalisms and details regarding datasets and neighboring relations.

Neighboring Datasets: Roughly speaking, an algorithm is differentially private if it difficult to distinguish between output distributions when the algorithm is run on similar inputs. In general, this notion of similarity amongst inputs is defined as a *neighboring relation* $\sim$ between elements on the input space $\mathcal{X}$. In particular, if two inputs (also referred to as datasets or databases) $x, x' \in \mathcal{X}$ satisfy the neighboring relation $x \sim x'$, the we say x and x' are *neighbors*.

There are several canonical examples of neighboring relations on the space of inputs $\mathcal{X}$. One example is where $\mathcal{X} = \mathbb{X}^n$ for some data domain $\mathbb{X}$. The data domain can be viewed as the set of all possible individual entries for a dataset, and the space $\mathbb{X}^n$ correspondingly contains all possible n element datasets. In this setting, databases $x, x' \in \mathcal{X}$ may be considered neighbors if x and x' differ in exactly one entry. Another slightly more general setting is when $\mathcal{X} = \mathbb{X}^*$, i.e., all possible datasets of finite size. In this situation, the earlier notion of neighboring still makes sense. However, in addition, we may say input datasets x and x' are neighbors if x can be obtained from x' by either adding or deleting an element. This is a very natural notion of neighboring, as under such a relation an algorithm would be differentially private if it were difficult to determine the presence or absence of an individual. Our work is agnostic to the precise choice of neighboring relation. As such, we choose to leave the notion as general as possible.

Algorithms and Random Variables: We will consider algorithms as randomized mappings $A : \mathcal{X} \to \mathcal{Y}$ taking inputs from $\mathcal{X}$ to some output space $\mathcal{Y}$. To be fully formal, we consider the output space $\mathcal{Y}$ as a *measurable space* $(\mathcal{Y}, \mathcal{G})$, where $\mathcal{G}$ is some σ-algebra denoting possible events. Recall that a σ-algebra $\mathcal{S}$ for a set S is simply a subset of 2^S containing S and $\emptyset$ that is closed under countable union, intersection, and complements. When we say A is an algorithm having inputs in some space $\mathcal{X}$, we really mean $A(x)$ is a $\mathcal{Y}$-valued random variable for any $x \in \mathcal{X}$. The space $\mathcal{X}$ need not have an associated σ-algebra, as algorithm inputs are essentially just indexing devices. Given a sequence of algorithms $(A_n)_{n\geq 1}$, $(A_n(x))_{n\geq 1}$ is a sequence of $\mathcal{Y}$-valued random variables, for any $x \in \mathcal{X}$.[6]

Since we are dealing with the composition of algorithms, we write $A_{1:n}(x)$ as shorthand for the random vector of the first n algorithm outputs, i.e. $A_{1:n}(x) = (A_1(x), \ldots, A_n(x))$. Formally, the random vector $A_{1:n}(x)$ takes output values in the product measurable space $(\mathcal{Y}^n, \mathcal{G}^{\otimes n})$ where $\mathcal{G}^{\otimes n}$ denotes the n-fold product σ-algebra of $\mathcal{G}$ with itself. Likewise, since the number of algorithm outputs one views in fully-adaptive composition may be random, if N is a random time (i.e. a $\mathbb{N}$-valued random variable), we will often consider the random vector $A_{1:N}(x) = (A_1(x), \ldots, A_N(x))$.

Filtrations and Stopping Times: Since privacy composition involves sequences of random outputs, we will use the measure-theoretic notion of a *filtration*. If we have fixed an input $x \in \mathcal{X}$, we can assume the random sequence $(A_n(x))_{n\geq 1}$ is defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Given such a probability space, a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ of $\mathcal{F}$ is a sequence of σ-algebras satisfying: (i) $\mathcal{F}_n \subset \mathcal{F}_{n+1}$ for all $n \in \mathbb{N}$, and (ii) $\mathcal{F}_n \subset \mathcal{F}$ for all $n \in \mathbb{N}$. Given an arbitrary $\mathcal{Y}$-valued discrete-time stochastic process $(X_n)_{n\geq 1}$, it is often useful to consider the *natural filtration* $(\mathcal{F}_n)_{n\in\mathbb{N}}$ given by $\mathcal{F}_n := \sigma(X_m : m \leq n)$ and $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Intuitively, a filtration formalizes the notion

[6]Even if algorithms have different types of outputs (maybe some algorithms have categorical outputs while others output real-valued vectors), $\mathcal{Y}$ can still be made appropriately large to contain all possible outcomes.

of accumulating information over time. In particular, in the context of the natural filtration generated by a stochastic process, the nth σ-algebra in the filtration $\mathcal{F}_n$ essentially represents the entirety of information contained in the first n random variables. In other words, if one is given $\mathcal{F}_n$, they would know all possible events/outcomes that could have occurred up to and including timestep n.

Lastly, we briefly mention the notion of a *stopping time*, as this measure-theoretic object is necessary to define privacy filters. Given a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$, a random time N is said to be a stopping time with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}}$ if, for any n, the event $\{N \leq n\} \in \mathcal{F}_n$. In words, a random time N is a stopping time if given the information in $\mathcal{F}_n$ we can determine whether or not we should have stopped by time n. Stopping times are essential to the study of fully-adaptive composition, as a practitioner of privacy will need to use the adaptively selected privacy parameters to determine whether or not to stop interacting with the underlying sensitive database.

4.B Martingale Inequalities

In this appendix, we provide a thorough exposition into the concentration inequalities leveraged in this paper. First, at the heart of supermartingale concentration is Ville's inequality [157], which was stated in Theorem 1.0.2.

We do not directly leverage Ville's inequality in this work, but all inequalities we use can be directly proven from Theorem 1.0.2 [73, 76]. In short, each inequality in this supplement is proved by carefully massaging a martingale of interest into a non-negative supermartingale.

Another useful tool we will leverage is Doob's optional stopping theorem.

Lemma 4.B.1 (Optional stopping theorem [49])**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a nonnegative supermartingale with respect to some filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$. Then $\mathbb{E}[X_\tau] \leq \mathbb{E}[X_0]$ for all stopping times τ that are potentially infinite.*

For our alternative proof of the privacy filter (in Section 4.D), we leverage the following special case of a recent advance in time-uniform martingale concentration [73]. The following Theorem 4.B.2 is just a special case of the main result in Howard et al. [73], and we include the proof for completeness. When we say a random variable X is σ^2-subGaussian conditioned on some sigma-algebra $\mathcal{G}$, we mean that, for all $\lambda \geq 0$,

$$\mathbb{E}\left(e^{\lambda X} \mid \mathcal{G}\right) \leq e^{\lambda^2\sigma^2/2}.$$

In particular, if X is σ^2-subGaussian as above, this does not imply that $-X$ is σ-subGaussian (because the condition is only assumed for $\lambda \geq 0$). In general, X can have different behaviors in its left and right tail, see for example the discussion of the differing tails of the empirical variance of Gaussians in Howard et al. [76].

Theorem 4.B.2. *Let $(M_n)_{n\in\mathbb{N}}$ be a martingale with respect to some filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ such that $M_0 = 0$ almost surely. Moreover, let $(\sigma_n)_{n\geq 1}$ be a $(\mathcal{F}_n)_{n\in\mathbb{N}}$-predictable sequence of random variables such that, conditioned on $\mathcal{F}_{n-1}$, $\Delta M_n := M_n - M_{n-1}$ is σ_n^2-subGaussian. Define $V_n := \sum_{m\leq n}\sigma_m^2$. Then, we have, for all $a, b > 0$,*

$$\mathbb{P}\left(\exists n \in \mathbb{N} : M_n \geq \frac{b}{2} + \frac{b}{2a}V_n\right) \leq \exp\left(\frac{-b^2}{2a}\right).$$

Proof of Theorem 4.B.2. Let $(M_n)_{n\in\mathbb{N}}$ be the martingale listed in the theorem statement. Observe that, for any $a, b > 0$, the process $(X_n)_{n\in\mathbb{N}}$ given by

$$X_n := \exp\left(\frac{b}{a}M_n - \frac{b^2}{2a^2}\sum_{m\leq n}\sigma_m^2\right)$$

is a non-negative supermartingale. As such, applying Ville's inequality (Theorem 1.0.2) yields

$$\mathbb{P}\left(\exists n \in \mathbb{N} : X_n > \exp\left(\frac{b^2}{2a}\right)\right) \leq \exp\left(-\frac{b^2}{2a}\right).$$

Now, on such event, taking logs and rearranging yields

$$\frac{b}{a}M_n \leq \frac{b^2}{2a} + \frac{b^2}{2a^2}\sum_{m\leq n}\sigma_m^2.$$

Multiplying both sides by $\frac{a}{b}$ finishes the proof. ■

The predictable process $(V_n)_{n\in\mathbb{N}}$ is a proxy for the accumulated variance of $(M_n)_{n\in\mathbb{N}}$ up to any fixed point in time. In particular, the process $(V_n)_{n\in\mathbb{N}}$ can be thought of as yielding the "intrinsic time" of the process. The free parameters a and b thus allow us to optimize the tightness of the boundary for some intrinsic moment in time. This is ideal for us, as, for the sake of composition, the target privacy parameter ϵ can guide us in finding a point in intrinsic time (that is, in terms of the process $(V_n)_{n\in\mathbb{N}}$) to optimize for. We discuss how to apply this inequality to prove privacy composition results both in this supplement and in Section 4.3.

We also leverage the following martingale inequalities from Howard et al. [76] in Section 4.4, where we construct various families of time-uniform bounds on privacy loss in fully-adaptive composition. These inequalities take on a more complicated form than Theorem 4.B.2, but we explain the intuition behind them in the sequel. The first bound we present relies on the method of mixtures for martingale concentration, which stems back to Robbins' work in the 1970s [132]. There are many good resources providing an introduction to the method of mixtures [45, 87, 76].

Theorem 4.B.3. *Let $(M_n)_{n\in\mathbb{N}}$ be a martingale with respect to some filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ such that $M_0 = 0$ almost surely. Moreover, let $(\sigma_n)_{n\geq 1}$ be a $(\mathcal{F}_n)_{n\in\mathbb{N}}$-predictable sequence of random variables such that, conditioned on $\mathcal{F}_{n-1}$, $\Delta M_n := M_n - M_{n-1}$ is σ_n^2-subGaussian. Define $V_n := \sum_{m\leq n}\sigma_m^2$ and choose a tuning parameter $\gamma > 0$. Then, for any $\delta > 0$, we have*

$$\mathbb{P}\left(\exists n \in \mathbb{N} : M_n \geq \sqrt{2(V_n+\gamma)\log\left(\frac{1}{\delta}\sqrt{\frac{V_n+\gamma}{\gamma}}\right)}\right) \leq \delta.$$

The next inequality relies on the recent technique of boundary stitching, first presented in Howard et al. [76]. Intuitively, the technique works by breaking intrinsic time — that is, time according to the accumulated variance process $(V_n)_{n\in\mathbb{N}}$ — into roughly geometrically spaced pieces. Then, one optimizes a tight-boundary in each region and takes a union bound. The actual details are more technical, but are not needed in this work.

Theorem 4.B.4. *Let $(M_n)_{n\in\mathbb{N}}$ be a martingale with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}}$ such that $M_0 = 0$ almost surely. Moreover, let $(\sigma_n)_{n\geq 1}$ be a $(\mathcal{F}_n)_{n\in\mathbb{N}}$-predictable sequence of random variables such that, conditioned on $\mathcal{F}_{n-1}$, both $\Delta M_n := M_n - M_{n-1}$ and $-\Delta M_n$ are σ_n^2-subGaussian. Define $V_n := \sum_{m\leq n} \sigma_m^2$ and choose a starting intrinsic time $v_0 > 0$. Then, for any $\delta \in (0,1)$, we have*

$$\mathbb{P}\left(\exists n \in \mathbb{N} : M_n \geq 1.7\sqrt{V_n\left(\log\log\left(\frac{2V_n}{v_0}\right) + .72\log\left(\frac{5.2}{\delta}\right)\right)} \quad \textit{and} \quad V_n \geq v_0\right) \leq \delta.$$

Note that the original version of Theorem 4.B.4 as found in Howard et al. [76] has more free parameters to optimize over, but we have already simplified the expression to make the result more readable. The free parameter $v_0 > 0$ in the above boundary gives the intrinsic time at which the boundary becomes non-trivial (i.e., the tightest available upper bound before $V_n \geq v_0$ is ∞).

We qualitatively compare these bounds in Section 4.4, wherein we construct various time-uniform bounds on privacy loss processes. For now, Theorem 4.B.2 can be thought of as providing a tight upper bound on a martingale at a single point in intrinsic time, providing loose guarantees elsewhere. On the other hand, Theorems 4.B.3 and 4.B.4 provide decently tight control over a martingale at all points in intrinsic time simultaneously, although at the cost of sacrificing tightness at any given fixed point.

4.C Details in Proof of Approx-zCDP Filter

4.C.1 Equivalence of Approximate zCDP Definitions

We will show that our definition of approximate zCDP is equivalent to the original definition of approximate zCDP due to Bun and Steinke [20]. Let us first restate their definition as a condition on a private algorithm A.

Condition 4.C.1 (Original definition of Bun and Steinke [20])**.** For any neighboring datasets x, x', there exist events E and E' such that for all $\lambda \geq 1$,

$$\begin{aligned}
&D_\lambda(A(x) \mid E \| A(x') \mid E') \leq \rho\lambda, \\
&D_\lambda(A(x') \mid E' \| A(x) \mid E) \leq \rho\lambda, \\
&\mathbb{P}(A(x) \in E) \geq 1-\delta, \text{ and} \\
&\mathbb{P}(A(x') \in E') \geq 1-\delta.
\end{aligned}$$

Our definition is adapted from the approximate Rényi differential privacy definition due to Papernot and Steinke [127]. We restate the (unconditional) definition below.

Condition 4.C.2 (Adapted from Papernot and Steinke [127])**.** For any neighboring datasets x, x', there exist distributions P', P'', Q', Q'' such that the outputs are distributed according to the following mixture distributions:

$$A(x) \sim (1-\delta)P' + \delta P'', \qquad A(x') \sim (1-\delta)Q' + \delta Q''$$

with for all $\lambda \geq 1,\ D_\lambda(P' \| Q') \leq \rho\lambda$ and $D_\lambda(P' \| Q') \leq \rho\lambda$.

Theorem 4.C.3. *Conditions 4.C.1 and 4.C.2 are equivalent.*

Proof of Theorem 4.C.3. Fix any neighbors x, x'. Suppose an algorithm A satisfies Condition 4.C.1 for some events E, E'. Then we could let P' and Q' be the conditional distributions $\mathbb{P}(A(x) \in \cdot \mid A(x) \in E)$ and $\mathbb{P}(A(x') \in \cdot \mid A(x') \in E')$ respectively. Then let

$$\begin{aligned} P''(\cdot) &= \frac{1}{\delta}\Big(\mathbb{P}(A(x) \in \cdot \mid A(x) \in E^c)\mathbb{P}(A(x) \in E^c) \\ &\quad + P'(\cdot)\left(\mathbb{P}(A(x) \in E) - (1-\delta)\right)\Big), \\ Q''(\cdot) &= \frac{1}{\delta}\Big(\mathbb{P}(A(x') \in \cdot \mid A(x') \in E'^c)\mathbb{P}(A(x') \in E'^c) \\ &\quad + Q'(\cdot)\left(\mathbb{P}(A(x') \in E') - (1-\delta)\right)\Big). \end{aligned}$$

Then $A(x)$ is distributed according to the mixture $(1-\delta)P' + \delta P''$, and $A(x')$ is distributed according to the mixture $(1-\delta)Q' + \delta Q''$. Thus, A also satisfies condition 4.C.2 given that $D_\lambda(P' \| Q') \le \lambda\rho$ and $D_\lambda(Q' \| P') \le \lambda\rho$ by our assumption of Condition 4.C.1.

Now suppose A satisfies Condition 4.C.2 for some pairs of distributions (P', P'') and (Q', Q''). Then we can view the output distribution of $A(x)$ as generating a Bernoulli random variable C such that with probability $(1-\delta)$, $C = 1$ and $A(x)$ draws an outcome from P' and with probability $C = 0$ and $A(x)$ draws an outcome from P''. Similarly, we can view $A(x')$ as flipping a coin C' such that $A(x')$ draws an outcome from Q' when $C' = 1$. Then letting the events E be all the randomness of $A(x)$ such that $C = 1$ and E' be all the randomness of $A(x')$ such that $C' = 1$ satisfies condition 4.C.1. ■

4.C.2 Missing Proofs

The following proof technique was used in prior works, including [24, 61]

Lemma 4.C.4. *Let $(M_n^{(\lambda)})_{n\ge1}$ be as defied in Equation* (4.3.5)*. Then, $(M_n^{(\lambda)})_{n\ge1}$ is a non-negative supermartingale with respect to its natural filtration $(\mathcal{F}'_n)_{n\ge1}$ given by $\mathcal{F}'_n := \sigma(Y'_m : m \le n)$.*

Proof. For any $k \ge 1$,

$$\begin{aligned} \mathbb{E}[M_n^{(\lambda)} \mid \mathcal{F}'_{n-1}] &= \mathbb{E}\left[M_{n-1}^{(\lambda)} \exp\left((\lambda-1)\log\left(\frac{P'_n(Y'_n \mid Y'_{1:n-1})}{Q'_n(Y'_n \mid Y'_{1:n-1})}\right) - \lambda(\lambda-1)\rho_n(Y'_{1:n-1})\right) \mid \mathcal{F}'_{n-1}\right] \\ &= M_{n-1}^{(\lambda)}\,\mathbb{E}\left[\left(\frac{P'_n(Y'_n \mid Y'_{1:n-1})}{Q'_n(Y'_n \mid Y'_{1:n-1})}\right)^{(\lambda-1)} \mid \mathcal{F}'_{n-1}\right] \cdot \exp\big(-\lambda(\lambda-1)\rho_n(Y'_{1:n-1})\big) \\ &\le M_{n-1}^{(\lambda)} \exp\big(\lambda(\lambda-1)\rho_n(Y'_{1:n-1})\big) \exp\big(-\lambda(\lambda-1)\rho_n(Y'_{1:n-1})\big) \\ &= M_{n-1}^{(\lambda)}, \end{aligned}$$

where the last inequality follows from the R[́enyi divergence bound due to approximate zCDP. ■

Lemma 4.C.5. *Let the distributions $P_{1:n}, Q_{1:n}, P'_{1:n}, Q'_{1:n}$ be defined in* (4.3.3)*,* (4.3.4) *for any $n \geq 1$. Then there exists distributions $P''_{1:n}$ and $Q''_{1:n}$ such that*

$$P_{1:n} = (1-\delta)P'_{1:n} + \delta P''_{1:n},$$
$$Q_{1:n} = (1-\delta)Q'_{1:n} + \delta Q''_{1:n}.$$

Proof. We will show the decomposition for $P_{1:n}$, and the proof follows identically for the decomposition of $Q_{1:n}$. First, we can express $P_{1:n}(y_1, \cdots, y_n)$ for any $y_1, \cdots y_n$ as follows:

$$\begin{aligned}
&P_{1:n}(y_1, \cdots, y_n) = \prod_{m=1}^{n} P_m(y_m \mid y_{1:m-1}) \\
&= \prod_{m=1}^{n} \left[(1-\delta_m(y_{1:m-1}))P'_m(y_m \mid y_{1:m-1}) + \delta_m(y_{1:m-1})P''_m(y_m \mid y_{1:m-1})\right] \\
&= \sum_{S \subseteq [n]} \underbrace{\left(\prod_{m \in S} \delta_m(y_{1:m-1}) \prod_{m \in S^c} (1-\delta_m(y_{1:m-1}))\right)}_{w_S(y_{1:m})} \cdot \underbrace{\prod_{m \in S} P''_m(y_m \mid y_{1:m-1}) \prod_{m \leq n, m \notin S} P'_m(y_m \mid y_{m-1})}_{f_S(y_{1:m})}
\end{aligned}$$

It suffices to show that $w_\emptyset(y_{1:m}) \geq 1-\delta$ for all $y_{1:m}$. To see this, we have the following by assumption

$$w_\emptyset = \prod_{m \leq n} (1-\delta_m(y_{1:m-1})) \geq 1 - \sum_{m \leq n} \delta_m(y_{1:m-1}) \geq 1-\delta.$$

■

4.D An Alternative Proof for Theorem 4.3.3

We begin by providing an alternative statement to Theorem 4.3.3, which is fully stated in terms of ϵ's and δ's. Straightforward calculations can confirm the equivalence of the two statements.

Theorem 4.D.1. *Suppose $(A_n)_{n \geq 1}$ is a sequence of algorithms such that, for any $n \geq 1$, A_n is (ϵ_n, δ_n)-differentially private conditioned on $A_{1:n-1}$. Let $\epsilon > 0$ and $\delta = \delta' + \delta''$ be target privacy parameters such that $\delta' > 0, \delta'' \geq 0$. Consider the function $N : \mathbb{R}_{\geq 0}^{\infty} \times \mathbb{R}_{\geq 0}^{\infty} \to \mathbb{N}$ given by*

$$N((\epsilon_n)_{n\geq 1}, (\delta_n)_{n \geq 1}) := \inf \left\{ n : \epsilon < \sqrt{2 \log\left(\frac{1}{\delta'}\right) \sum_{m \leq n+1} \epsilon_m^2} + \frac{1}{2} \sum_{m \leq n+1} \epsilon_m^2 \quad \textit{or} \quad \delta'' < \sum_{m \leq n+1} \delta_m \right\}.$$

Then, the algorithm $A_{1:N(\cdot)}(\cdot) : \mathcal{X} \to \mathcal{Y}^{\infty}$ is (ϵ, δ)-DP, where $N(x) := N((\epsilon_n(x))_{n \geq 1}, (\delta_n(x))_{n \geq 1})$. In other words, N is an (ϵ, δ)-privacy filter.

We first prove Theorem 4.D.1 under a stronger assumption on the algorithms being composed.

Lemma 4.D.2. *Theorem 4.D.1 holds under the stronger assumption that, for any $n \geq 1$, A_n is (ϵ_n, δ_n)-pDP conditioned on $A_{1:n-1}$.*

To prove Lemma 4.D.2, we need to following bound on the conditional expectation of privacy loss, which can be immediately obtained from the bound on expected privacy loss presented in Bun and Steinke [20].

Lemma 4.D.3 (Proposition 3.3 in Bun and Steinke [20])**.** *Suppose A and B are algorithms such that A is ϵ-differentially private conditioned on B. Then, for any input dataset $x \in \mathcal{X}$ and neighboring dataset $x' \sim x$, we have that*

$$\mathbb{E}\left(\mathcal{L}(x, x')|B(x)\right) \leq \frac{1}{2}\left(\epsilon(B(x))\right)^2.$$

Now, we prove Lemma 4.D.2.

Proof of Lemma 4.D.2. To begin, we assume that the algorithms $(A_n)_{n\geq 1}$ satisfy $(\epsilon_n, 0)$-pDP conditioned on $A_{1:n-1}$. We will show how to alleviate this assumption on the approximation parameter in the second half of the proof. Fix an input database $x \in \mathcal{X}$. For convenience, we denote by $(\mathcal{F}_n(x))_{n\in\mathbb{N}}$ the natural filtration generated by $(A_n(x))_{n\geq 1}$. Since we have fixed $x \in \mathcal{X}$, for notational simplicity, we write ϵ_n for the random variable $\epsilon_n(A_{1:n-1}(x))$ and define δ_n similarly. Additionally, by N we mean the stopping time $N((\epsilon_n)_{n\in\mathbb{N}}, (\delta_n)_{n\in\mathbb{N}})$. Recall that we have already argued that, for any neighboring dataset $x' \sim x$, the process

$$M_n := M_n(x, x') = \mathcal{L}_{1:n}(x, x') - \sum_{m\leq n} \mathbb{E}\left(\mathcal{L}_m(x, x')|\mathcal{F}_{m-1}(x)\right)$$

is a martingale with respect to $(\mathcal{F}_n(x))_{n\in\mathbb{N}}$. Further observe that its increments $\Delta M_n := \mathcal{L}_n(x, x') - \mathbb{E}\left(\mathcal{L}_n(x, x')|\mathcal{F}_{n-1}(x)\right)$ are ϵ_n^2-subGaussian conditioned on $\mathcal{F}_{n-1}(x)$.

Thus, by Theorem 4.B.2, we know that, for any $b, a > 0$, we have

$$\mathbb{P}\left(\exists n \in \mathbb{N} : M_n \geq \frac{b}{2} + \frac{b}{2a}V_n\right) \leq \exp\left(\frac{-b^2}{2a}\right),$$

where the process $(V_n)_{n\in\mathbb{N}}$ given by $V_n := \sum_{m\leq n} \epsilon_m^2$ is the accumulated variance up to and including time n. Thus, it suffices to optimize the free parameters a and b to prove the result.

To do this, consider the following function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ given by

$$f(y) = \sqrt{2\log\left(\frac{1}{\delta'}\right)y} + \frac{1}{2}y.$$

Clearly, f is a quadratic polynomial in $\sqrt{y}$ which is strictly increasing. In particular, one can readily check that

$$y^* := \left(-\sqrt{2\log\left(\frac{1}{\delta'}\right)} + \sqrt{2\log\left(\frac{1}{\delta'}\right) + \epsilon}\right)^2 \tag{4.D.1}$$

solves the equation $f(y) = \epsilon$, where $\epsilon > 0$ is the target privacy parameter.

As such, setting $a := y^*$ and $b := \sqrt{2\log\left(\frac{1}{\delta'}\right)y^*}$ yields

$$\exp\left(\frac{-b^2}{a}\right) = \exp\left(\frac{-2y^*\log\left(\frac{1}{\delta'}\right)}{y^*}\right) = \delta'.$$

Furthermore, expanding the definition of $(M_n)_{n\in\mathbb{N}}$, we see that for the selected parameters the parameters yield, with probability at least $1-\delta'$, for all $n \leq N$ we have:

$$\begin{aligned}
\mathcal{L}_{1:n}(x,x') &\leq \frac{b}{2} + \frac{b}{2a}V_n + \sum_{m\leq n}\mathbb{E}\left(\mathcal{L}_m(x,x') \mid \mathcal{F}_{m-1}\right) \\
&\leq \frac{b}{2} + \frac{b}{2a}\sum_{m\leq n}\epsilon_m^2 + \frac{1}{2}\sum_{m\leq n}\epsilon_m^2 \\
&= \frac{1}{2}\sqrt{2\log\left(\frac{1}{\delta'}\right)y^*} + \frac{1}{2}\frac{\sqrt{2\log\left(\frac{1}{\delta'}\right)y^*}}{y^*}\sum_{m\leq n}\epsilon_m^2 + \frac{1}{2}\sum_{m\leq n}\epsilon_m^2 \\
&\leq \frac{1}{2}\sqrt{2\log\left(\frac{1}{\delta'}\right)y^*} + \frac{1}{2}\sqrt{2\log\left(\frac{1}{\delta'}\right)y^*} + \frac{1}{2}\sum_{m\leq n}\epsilon_m^2 \\
&= \sqrt{2\log\left(\frac{1}{\delta'}\right)y^*} + \frac{1}{2}\sum_{m\leq n}\epsilon_m^2 \leq \sqrt{2\log\left(\frac{1}{\delta'}\right)y^*} + \frac{1}{2}y^* = \epsilon.
\end{aligned}$$

Thus, we have proven the desired result in the case where all algorithms have $\delta_n = 0$.

Now, we show how to generalize our result to the case where the approximation parameters δ_n are not identically zero. Define the events

$$\begin{aligned}
A &:= \left\{\exists n \leq N : \mathcal{L}_{1:n}(x,x') > \epsilon\right\}, \text{ and} \\
B &:= \left\{\exists n \leq N : \mathcal{L}_n(x,x') > \epsilon_n\right\}.
\end{aligned}$$

Our goal is to show that, with N defined as in the statement of Theorem 4.3.3, that $\mathbb{P}(A) \leq \delta$. Simply using Bayes rule, we have that

$$\mathbb{P}(A) = \mathbb{P}(A \cap B^c) + \mathbb{P}(A \cap B) \leq \mathbb{P}(A|B^c) + \mathbb{P}(B) \leq \delta' + \mathbb{P}(B),$$

where the second inequality follows from our already-completed analysis in the case that $\delta_n = 0$. Now, we show that $\mathbb{P}(B) \leq \delta''$, which suffices to prove the result as we have, by assumption, $\delta = \delta' + \delta''$.

Define the modified privacy loss random variables $(\widetilde{\mathcal{L}}_n(x,x'))_{n\in\mathbb{N}}$ by

$$\widetilde{\mathcal{L}}_n(x,x') := \begin{cases} \mathcal{L}_n(x,x') & n \leq N \\ 0 & \text{otherwise} \end{cases}.$$

Likewise, define the modified privacy parameter random variables $\widetilde{\epsilon}_n$ and $\widetilde{\delta}_n$ in an identical manner. Then, we can bound $\mathbb{P}(B)$ in the following manner:

$$\mathbb{P}(\exists n \leq N : \mathcal{L}_n(x,x') > \epsilon_n) = \mathbb{P}\left(\exists n \in \mathbb{N} : \widetilde{\mathcal{L}}_n(x,x') > \widetilde{\epsilon}_n\right)$$

$$\leq \sum_{n=1}^{\infty} \mathbb{P}\left(\widetilde{\mathcal{L}}_n(x, x') > \widetilde{\epsilon}_n\right) = \sum_{n=1}^{\infty} \mathbb{E}\mathbb{P}\left(\widetilde{\mathcal{L}}_n(x, x') > \widetilde{\epsilon}_n | \mathcal{F}_{n-1}\right)$$
$$\leq \sum_{n=1}^{\infty} \mathbb{E}\widetilde{\delta}_n = \mathbb{E}\left[\sum_{n=1}^{\infty} \widetilde{\delta}_n\right] = \mathbb{E}\left[\sum_{n \leq N} \delta_n\right] \leq \delta''.$$

Thus, we have have proven the desired result in the general case. ■

Our key insight above is to view filters as functions of the "intrinsic time" determined by privacy parameters, $\sum_{m \leq n} \epsilon_m^2$. Lemma 4.D.2 can also be obtained leveraging the analysis for Rényi filters [61]. However, our approach to proving Lemma 4.D.2 has the advantage that it does not require reductions between different modes of privacy. While Lemma 4.D.3, which bounds expected privacy loss, does require some complicated analysis, we only ever need to apply Lemma 4.D.2 to instances of randomized response, in which case computing the privacy loss bound is trivial.

We now use Lemma 4.D.2 to prove Theorem 4.D.1. Recall that Lemma 4.4.2 shows that algorithms that satisfy pDP also satisfy DP, but the converse is not true and may require a conversion cost. To avoid this cost, we define following generalization of randomized response.

Definition 4.D.4 (Conditional Randomized Response)**.** Let $\mathcal{R} := \{0, 1, \top, \bot\}$ and $2^{\mathcal{R}}$ be the corresponding power set of $\mathcal{R}$. Then, R taking inputs in $\{0, 1\}$ to outputs in the measurable space $(\mathcal{R}, 2^{\mathcal{R}})$ is an instance of (ϵ, δ)-randomized response if, for $b \in \{0, 1\}$, $R(b)$ outputs the following:

$$R(b) = \begin{cases} b & \text{with probability } (1-\delta)\frac{e^\epsilon}{1+e^\epsilon} \\ 1-b & \text{with probability } (1-\delta)\frac{1}{1+e^\epsilon} \\ \top & \text{with probability } \delta \text{ if } b = 1 \\ \bot & \text{with probability } \delta \text{ if } b = 0. \end{cases}$$

More generally, suppose $B : \{0, 1\} \to \mathcal{Z}$ is a randomized algorithm. For functions $\epsilon, \delta : \mathcal{Z} \to \mathbb{R}_{\geq 0}$, we say R is an instance of (ϵ, δ)-randomized response conditioned on B if, for any true input $b' \in \{0, 1\}$ and hypothesized alternative $b \in \{0, 1\}$, the conditional probability $\mathbb{P}(R(b) \in \cdot | B(b') = z)$ is the same as the law of $(\epsilon(z), \delta(z))$-randomized response with input bit b.

Conditional (ϵ, δ)-randomized response satisfies both conditional (ϵ, δ)-DP and conditional (ϵ, δ)-pDP. We will leverage the fact that it satisfies both privacy definitions with the same parameters. A surprising result in the nonadaptive setting is that *any* (ϵ, δ)-DP algorithm can be viewed as a randomized post-processing of (ϵ, δ)-randomized response [84]. We generalize this result to the adaptive conditional setting below. In the language of Blackwell's comparison of experiments [17], instances of randomized response are "sufficient" for instances of arbitrary DP algorithms, and we prove that the same is true for conditional randomized response and conditionally DP algorithms. In what follows, by a transition kernel ν, we mean that for any $b \in \mathcal{Z}$ and $r \in \mathcal{R}$, $\nu(\cdot, r \mid b)$ is a probability measure on $(\mathcal{Y}, \mathcal{G})$.

Lemma 4.D.5 (Reduction to Conditional Randomized Response)**.** *Let A and B map from $\mathcal{X}$ to measurable spaces $(\mathcal{Y}, \mathcal{G})$ and $(\mathcal{Z}, \mathcal{H})$, respectively. Suppose A is (ϵ, δ)-differentially private*

conditioned on B. Fix neighbors $x_0, x_1 \in \mathcal{X}$, and let R be an instance of (ϵ, δ)-randomized response conditioned on B', where $B' : \{0,1\} \to \mathcal{Z}$ is the restricted algorithm satisfying $B'(b) = B(x_b)$. Then, there is a transition kernel $\nu : \mathcal{G} \times \mathcal{R} \times \mathcal{Z} \to [0,1]$ such that, for all $b, b' \in \{0,1\}$, $\mathbb{P}\left(A(x_b) \in \cdot \mid B'(b')\right) = \nu_{b,b'}$, where $\nu_{b,b'} = \mathbb{E}\left(\nu(\cdot, R(b) \mid B'(b')) \mid B'(b')\right)$.[7]

Lemma 4.D.5 tells us that the conditional distribution obtained by averaging the kernel $\nu(\cdot, R(b) \mid B'(b'))$ over the randomness in $R(b)$ matches the conditional distribution of $A(x_b)$. To prove Lemma 4.D.5, first recall the important fact that *any* differentially private algorithm can be viewed as a post-processing of randomized response [84], as stated in Lemma 4.D.6 below.

Lemma 4.D.6 (Reduction to Randomized Response [84])**.** *Let algorithm $A : \mathcal{X} \to \mathcal{Y}$ be (ϵ, δ)-DP. Let R be an instance of (ϵ, δ)-randomized response. Then, for any neighbors $x_0, x_1 \in \mathcal{X}$, there is a transition kernel $\nu : \mathcal{G} \times \mathcal{R} \to [0,1]$ such that for $b \in \{0,1\}$, we have $\mathbb{P}(A(x_b) \in \cdot) = \nu_b$, where*[8] *$\nu_b = \mathbb{E}\nu(\cdot, R(b))$.*

In Lemma 4.D.5 of Section 4.3, we generalized Lemma 4.D.6 to the case of conditional differential privacy. To do this, we introduced *conditional randomized response* in Definition 4.D.4. In conditional randomized response, on the event $\{B = z\}$, the conditional laws of $R(0)$ and $R(1)$ just become that of regular randomized response with some known privacy parameters $\epsilon(z)$ and $\delta(z)$. We now prove Lemma 4.D.5.

Proof of Lemma 4.D.5. Let $b, b' \in \{0,1\}$ be arbitrary. For any outcome $\{B'(b') = z\}$, let $\mathbb{P}_z(A(x_b) \in \cdot)$ be the probability measure $\mathbb{P}(A(x_b) \in \cdot | B'(b') = z)$. In particular, this measure does not depend on the input bit b'. By the assumptions of conditional differential privacy (Definition 4.2.1), it follows that under the probability measure $\mathbb{P}_z$, $A(x_b)$ is $(\epsilon(z), \delta(z))$-differentially private. Moreover, it also follows that R is an instance of $(\epsilon(z), \delta(z))$-randomized response under $\mathbb{P}_z$. Consequently, Lemma 4.D.6 yields the existence of a kernel ν_z such that $\mathbb{P}_z(A(x_b) \in \cdot) = \mathbb{E}_z \nu_z(\cdot, R(b))$, where the averaged measure is as defined in Footnote 8. Setting $\nu(\cdot, R(b)|z) := \nu_z(\cdot, R(b))$, we see that

$$\mathbb{P}(A(x_b) \in \cdot \mid B'(b') = z) = \mathbb{E}\left(\nu(\cdot, R(b) \mid z) \mid B'(b') = z\right),$$

[7] By $\nu_{b,b'}(\cdot) := \mathbb{E}\left(\nu(\cdot, R(b) \mid B'(b')) \mid B'(b')\right)$, we mean that $\nu_{b,b'}$ is the (random) averaged probability measure:

$$\begin{aligned}\nu_{b,b'}(\cdot) = {} & \mathbb{P}(R(b) = 1 \mid B'(b'))\nu(\cdot, 1 \mid B'(b')) \\ & + \mathbb{P}(R(b) = 0 \mid B'(b'))\nu(\cdot, 0 \mid B'(b')) \\ & + \mathbb{P}(R(b) = \perp \mid B'(b'))\nu(\cdot, \perp \mid B'(b')) \\ & + \mathbb{P}(R(b) = \top \mid B'(b'))\nu(\cdot, \top \mid B'(b')).\end{aligned}$$

[8] By $\nu_b(\cdot) := \mathbb{E}\nu(\cdot, R(b))$, we mean ν_b is the averaged probability measure given by

$$\begin{aligned}\nu_b(\cdot) = {} & \mathbb{P}(R(b) = 1)\nu(\cdot, 1) + \mathbb{P}(R(b) = 0)\nu(\cdot, 0) \\ & + \mathbb{P}(R(b) = \perp)\nu(\cdot, \perp) + \mathbb{P}(R(b) = \top)\nu(\cdot, \top).\end{aligned}$$

which thus yields

$$\mathbb{P}(A(x_b) \in \cdot \mid B'(b')) = \mathbb{E}\left(\nu(\cdot, R(b) \mid B'(b')) \mid B'(b')\right),$$

where the conditionally averaged measure is as described in Footnote 7 in the main body of the paper. This proves the desired result. ■

Lastly, before proving Theorem 4.D.1, we need the following lemma. This lemma essentially tells us that if A is (ϵ, δ)-pDP conditioned on B, and A' is a randomized post-processing algorithm, then releasing the vector (A, A') is also (ϵ, δ)-pDP conditioned on B. Note that this is *not* in contradiction with the converse direction of Lemma 4.4.2, as releasing the output of A' alone may not satisfy conditional (ϵ, δ)-pDP. But once we observe A, since A' is a post-processing, we can gleam no more information about the true underlying dataset.

Lemma 4.D.7. *Suppose A, B are algorithms with inputs in $\mathcal{X}$ and outputs in measurable spaces $(\mathcal{Y}, \mathcal{G})$ and $(\mathcal{Z}, \mathcal{H})$ respectively. Assume A is (ϵ, δ)-pDP conditioned on B. Let $(S, \mathcal{S})$ be a measurable space and suppose $\mu : \mathcal{S} \times \mathcal{Y} \times \mathcal{Z} \to [0, 1]$ is a conditional transition kernel. Suppose $A' : \mathcal{X} \to S$ is an algorithm satisfying*

$$\mathbb{P}\left(A'(x) \in \ \cdot \ \mid A(x') = y, B(x') = z\right) = \mu(\cdot, y \mid z), \tag{4.D.2}$$

for all $y \in \mathcal{Y}, z \in \mathcal{Z}$, and $x, x' \in \mathcal{X}$. Then, the joint algorithm $(A, A') : \mathcal{X} \to \mathcal{Y} \times S$ is also (ϵ, δ)-pDP conditioned on B.

Proof of Lemma 4.D.7. Let $x, x' \in \mathcal{X}$ be arbitrary neighboring datasets. Let $q_B^x, q_B^{x'}$ be the corresponding conditional joint densities of $(A(x), A'(x))$ and $(A(x'), A'(x'))$ given $B(x)$ respectively. Likewise, let $p_B^x, p_B^{x'}$ be the corresponding conditional densities of $A(x)$ and $A(x')$ respectively conditioned on $B(x)$, and $q_{B,A}^x, q_{B,A}^{x'}$ the conditional densities of $A'(x)$ and $A'(x')$ given $A(x)$ and $B(x)$. Let $\mathcal{L}_B^{(A,A')}(x, x')$ denote the joint privacy loss between $(A(x), A'(x))$ and $(A(x'), A'(x'))$ given $B(x)$, while $\mathcal{L}_B^A(x, x')$ denotes the privacy loss between $A(x)$ and $A(x')$ given $B(x)$. We have, using Bayes rule,

$$\begin{aligned}
\mathcal{L}_B^{(A,A')}(x, x') &= \log\left(\frac{q_B^x(A(x), A'(x) \mid B(x))}{q_B^{x'}(A(x), A'(x) \mid B(x))}\right) \\
&= \log\left(\frac{p_B^x(A(x) \mid B(x))}{p_B^{x'}(A(x) \mid B(x))} \cdot \frac{q_{B,A}^x(A'(x) \mid B(x), A(x))}{q_{B,A}^{x'}(A'(x) \mid B(x), A(x))}\right) \\
&= \log\left(\frac{p_B^x(A(x) \mid B(x))}{p_B^{x'}(A(x) \mid B(x))}\right) = \mathcal{L}_B^{(A)}(x, x'),
\end{aligned}$$

The first equality on the second line follows from the assumption outlined in Equation (4.D.2). More specifically, since we have

$$\begin{aligned}
&\mathbb{P}\left(A'(x) \in \cdot | A(x), B(x)\right) = \mu(\cdot, A(x) \mid B(x)) = \\
&\mathbb{P}\left(A'(x') \in \cdot | A(x), B(x)\right),
\end{aligned}$$

it follows that the conditional densities $q_{B,A}^x$ and $q_{B,A}^{x'}$ are equal almost surely. Since A is (ϵ, δ)-pDP conditioned on B, the result now follows. ■

We now can prove Theorem 4.D.1 using these tools.

Proof of Theorem 4.D.1. Fix arbitrary neighbors $x_0, x_1 \in \mathcal{X}$. Let $(R_n)_{n\geq 1}$ be a sequence of algorithms such that R_n is an instance of (ϵ_n, δ_n)-randomized response conditioned on $A'_{1:n-1} : \{0,1\} \to \mathcal{Y}^{n-1}$, where $A'_m : \{0,1\} \to \mathcal{Y}$ is the restricted algorithm given by $A'_m(b) := A_m(x_b)$, for all $m \geq 1$. Lemma 4.D.5 guarantees the existence of a sequence of transition kernels $(\nu_n)_{n\geq 1}$, $\nu_n : \mathcal{G} \times \mathcal{R} \times \mathcal{Y}^{n-1} \to [0,1]$ such that, for all $n \geq 1$ and $b, b' \in \{0,1\}$, we have $\mathbb{P}(A'_n(b) \in \cdot \mid A'_{1:n-1}(b')) = \nu^{(n)}_{b,b'}$ almost surely. Here, $\nu^{(n)}_{b,b'}$ is the averaged conditional probability, as defined in terms of ν_n in Lemma 4.D.5 and Footnote 7. This equality means we can find an underlying probability space (i.e. a coupling) such that the random post-processing draws from the kernel $\nu_n(\cdot, R_n(b) \mid A'_{1:n-1}(b'))$ equal $A'_n(b)$ almost surely, for all $n \geq 1$.

Now, for any $n \geq 1$, since R_n is an instance of (ϵ_n, δ_n)-randomized response conditioned on $A'_{1:n-1}$, it follows that R_n is in fact (ϵ_n, δ_n)-pDP conditioned on $A'_{1:n-1}$. Moreover, this also implies that R_n is (ϵ_n, δ_n)-pDP conditioned on $(A'_{1:n-1}, R_{1:n-1})$, since, by definition, ϵ_n and δ_n only depend on the realizations of $R_{1:n-1}$ through the outputs of $A'_{1:n-1}$. By Lemma 4.D.7, it follows that for all $n \geq 1$, the algorithm (R_n, A'_n) is (ϵ_n, δ_n)-pDP conditioned on $(R_{1:n-1}, A'_{1:n-1})$. Thus, by Lemma 4.D.2, it follows that the composed algorithm $(R_{1:N'(\cdot)}(\cdot), A'_{1:N'(\cdot)}(\cdot))$ is (ϵ, δ)-DP, where $N'(b) := N(x_b)$ and ϵ, δ and N, are as outlined in the statement of Theorem 4.3.3.

Lastly, since differential privacy is closed under arbitrary post-processing [52], it follows that $A'_{1:N'(\cdot)}(\cdot)$ is (ϵ, δ)-differentially private. Since x_0 and x_1 were arbitrary neighboring inputs, the result follows, i.e. $A_{1:N(\cdot)}(\cdot) : \mathcal{X} \to \mathcal{Y}^\infty$ is (ϵ, δ)-differentially private. ■

4.E Proof for Privacy Odometers in Theorem 4.4.5

We now show the formal proof for our privacy odometers presented in Theorem 4.4.5 in Section 4.4.

Theorem 4.4.5. As in the proof of Theorem 4.D.2, we first consider the case where $\delta_n = 0$ for all $n \geq 1$. In this case, fix an input dataset $x \in \mathcal{X}$ and a neighboring dataset $x' \in \mathcal{X}$. Let $(M_n)_{n\in\mathbb{N}}$ be the corresponding privacy loss martingale as outlined in Equation (4.4.3), where we implicitly hide the dependence on x, x', which are fixed. Let $(u_n)_{n\geq 1}$ be one of the sequences outlined in the theorem statement, and define $U_n := u_n(\epsilon_{1:n}, \delta_{1:n})$ for all $n \geq 1$, where once again we write ϵ_n and δ_n for $\epsilon_n(A_{1:n-1}(x))$ and $\delta_n(A_{1:n-1}(x))$ respectively. It follows from Theorems 4.B.2, 4.B.3, and 4.B.4 that

$$\mathbb{P}\left(\exists n \in \mathbb{N} : M_n > B_n\right) \leq \delta,$$

for $B_n = U_n - \frac{1}{2}\sum_{m\leq n} \epsilon_m^2$. Recalling that $M_n = \sum_{m\leq n}\{\mathcal{L}_m(x, x') - \mathbb{E}(\mathcal{L}_m(x, x')|\mathcal{F}_{n-1}(x))\}$ and that $\mathbb{E}(\mathcal{L}_n(x, x')|\mathcal{F}_{n-1}(x)) \leq \frac{1}{2}\epsilon_n^2$ for all $n \in \mathbb{N}$, it thus follows that

$$\mathbb{P}\left(\exists n \in \mathbb{N} : \mathcal{L}_{1:n}(x, x') > U_n\right) \leq \delta,$$

where $(\mathcal{F}_n(x))_{n\geq 1}$ is again the natural filtration generated by $(A_n(x))_{n\geq 1}$. Thus, since $x \sim x'$ were arbitrary, we have shown that $(u_n)_{n\geq 1}$ is a δ-privacy odometer in the case $\delta_n = 0$ for all $n \geq 1$.

To generalize to the case where δ_n may be nonzero, we can apply precisely the same argument used in the second part of the proof of Lemma 4.D.2, thus proving the general result. ■

4.F An Algorithm Satisfying (ϵ, δ)-DP but not (ϵ, δ)-pDP

In this appendix, we construct a simple algorithm taking binary inputs that satisfies (ϵ, δ)-DP but not (ϵ, δ)-pDP. In particular, this provides intuition as to why we conjecture our odometers constructed in Section 4.4 would not hold under the assumption that the algorithms being composed satisfy (ϵ, δ)-DP in general.

To this end, fix a privacy parameter $\epsilon > 0$ and an approximation parameter $\delta \in (0, 1)$. Let $A : \{0, 1\} \to \{0, 1, \top, \bot\}$ be an instance of (ϵ, δ)-randomized response, and let $B : \{0, 1\} \to \{0, 1\}$ be defined by

$$B(b) := \begin{cases} 1 & \text{if } A(b) \in \{1, \top\}, \\ 0 & \text{otherwise.} \end{cases}$$

Since differential privacy is closed under arbitrary post-processing, it follows that the constructed algorithm B is (ϵ, δ)-differentially private. On the other hand, setting $x = 1$, $x' = 0$, we note that on the event $\{B(1) = 1\}$,

$$\begin{aligned} \mathcal{L}_B(1, 0) &= \log\left(\frac{\mathbb{P}(B(1) = 1)}{\mathbb{P}(B(0) = 1)}\right) = \log\left(\frac{\mathbb{P}(A(1) = 1) + \mathbb{P}(A(1) = \top)}{\mathbb{P}(A(0) = 1) + \mathbb{P}(A(0) = \top)}\right) \\ &= \log\left(\frac{\delta + (1 - \delta)\frac{e^\epsilon}{1+e^\epsilon}}{(1 - \delta)\frac{1}{1+e^\epsilon}}\right) \\ &= \log\left(\frac{\delta + e^\epsilon}{1 - \delta}\right) > \epsilon. \end{aligned}$$

Since straightforward calculation yields

$$\mathbb{P}(B(1) = 1) = (1 - \delta)\frac{e^\epsilon}{1 + e^\epsilon} + \delta > \delta,$$

we see that B does not satisfy (ϵ, δ)-pDP.

Chapter 5

Brownian Noise Reduction: Maximizing Privacy Subject to Accuracy Constraints

There is a disconnect between how researchers and practitioners handle privacy-utility tradeoffs. Researchers primarily operate from a privacy first perspective, setting strict privacy requirements and minimizing risk subject to these constraints. Practitioners often desire an accuracy first perspective, possibly satisfied with the greatest privacy they can get subject to obtaining sufficiently small error. Ligett et al. [109] have introduced a "noise reduction" algorithm to address the latter perspective. The authors show that by adding correlated Laplace noise and progressively reducing it on demand, it is possible to produce a sequence of increasingly accurate estimates of a private parameter while only paying a privacy cost for the least noisy iterate released. In this work, we generalize noise reduction to the setting of Gaussian noise, introducing the Brownian mechanism. The Brownian mechanism works by first adding Gaussian noise of high variance corresponding to the final point of a simulated Brownian motion. Then, at the practitioner's discretion, noise is gradually decreased by tracing back along the Brownian path to an earlier time. Our mechanism is more naturally applicable to the common setting of bounded ℓ_2-sensitivity, empirically outperforms existing work on common statistical tasks, and provides customizable control of privacy loss over the entire interaction with the practitioner. We complement our Brownian mechanism with $\mathrm{ReducedAboveThreshold}$*, a generalization of the classical* $\mathrm{AboveThreshold}$ *algorithm that provides adaptive privacy guarantees. Overall, our results demonstrate that one can meet utility constraints while still maintaining strong levels of privacy.*

5.1 Introduction

Over the past decade, differential privacy has seen industry-wide adoption as a means of protecting sensitive information [58, 65]. By injecting appropriate amounts of noise, differentially private algorithms allow the computation of population-level quantities of interest while guaranteeing individual-level privacy. Of the private mechanisms used in industry, those relating to private empirical risk minimization (ERM) are perhaps the most impactful, in part due to

their application in machine learning tasks [1, 148]. Researchers have developed many private ERM mechanisms, ranging from least squares minimzation [138, 25] to subsampled gradient descent [1, 11, 160]. Despite this vast literature, most existing results take the same broad approach: they aim to minimize error (statistical risk) subject to strict privacy guarantees. While this strict adherence to privacy constraints may be necessary in some applications, it often provides weak utility guarantees [63] and can make some learning tasks impossible [56]. Industry applications of differential privacy may desire an *accuracy first* perspective, setting desired risk requirements for models used in production. Privacy may still be a desirable aspect of computation, but it is by no means the only goal; minimizing risk may take center stage.

The main existing approach to this accuracy-oriented perspective on privacy was given by Ligett et al. [109]. These authors introduce a *noise reduction mechanism* for gradually releasing a private, high-dimensional parameter. By leveraging a Laplace-based Markov process [94], they construct a mechanism for which the privacy loss of releasing arbitrarily many estimates of a parameter only depends on the privacy loss of the least noisy parameter viewed. This is in contrast to results about the composition of private algorithms, in which privacy degrades according to the total number of parameters witnessed [57, 84, 123]. The authors also demonstrate how to privately query the utility of observed parameters on private data by coupling their Laplace-based mechanism with $\mathrm{AboveThreshold}$, a classical differentially private algorithm [52, 114].

While the above mechanism provides significant privacy loss savings over a baseline method that doubles the privacy loss each round, Laplace noise is unfit for many settings in which ℓ_2-sensitivity is used for calibrating noise. Since converting from ℓ_2-sensitivity to ℓ_1-sensitivity[1] incurs a dimension-dependent cost, it is important to develop a noise reduction technique with Gaussian noise.

Contributions and paper outline. We introduce the *Brownian mechanism*, a novel approach for privately releasing a parameter vector subject to accuracy constraints. The Brownian mechanism adds correlated Gaussian noise to a risk-minimizing parameter through a Brownian motion. Noise is then iteratively stripped by moving adaptively backwards along the random walk until a suitable stopping condition is met, such as meeting a target accuracy on a public dataset. In Section 5.3, we define the Brownian mechanism and characterize its privacy loss. Using machinery from martingale theory, we construct *privacy boundaries* for the Brownian mechanism — upper bounds on privacy loss that hold simultaneously with high probability. In particular, the failure probability of these bounds does not depend on the number of outcomes observed, overcoming a seeming need for a union bound faced by Ligett et al. [109]. These privacy boundaries yield provable, high-probability bounds on privacy loss under data-dependent stopping conditions.

If private data is used to evaluate risk, then the data-dependent stopping conditions can themselves leak information. To counter this, we introduce $\mathrm{ReducedAboveThreshold}$ in Section 5.5, a generalization of the classical $\mathrm{AboveThreshold}$ algorithm for privately querying accuracy on sensitive data. We show how to couple $\mathrm{ReducedAboveThreshold}$ and the Brownian mechanism so that a data analyst only ever incurs *twice* the privacy loss they would incur if they had queried accuracy on a public dataset. This is in contrast to the results in Ligett et al. [109], which note that the privacy loss of $\mathrm{AboveThreshold}$ often dominates the privacy loss incurred from using noise reduction.

[1]The ℓ_p sensitivity of f is defined as $\sup_{x \sim x'} ||f(x) - f(x')||_p$ for $p \geq 1$.

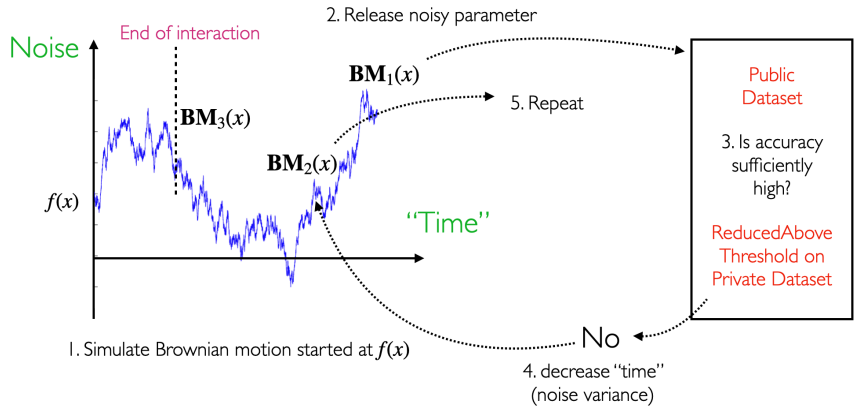


Figure 5.1: An example of running the Brownian mechanism to gradually release a statistic $f(x)$. First, a very noisy version of the hidden parameter $\mathrm{BM}_1(x)$ is viewed. Then, loss is measured, either on a public dataset, or on a private dataset using a method such as $\mathrm{ReducedAboveThreshold}$. If a target loss is met, the process stops. Otherwise, noise is removed and the process repeats.

We empirically evaluate the Brownian mechanism and $\mathrm{ReducedAboveThreshold}$ in Section 5.6, finding that the Brownian mechanism can offer privacy loss savings over the Laplace noise reduction method introduced by Ligett et al. [109]. In our view, these results demonstrate that the Brownian mechanism is a practical, intuitive mechanism for meeting accuracy requirements in private ERM.

Lastly, we derive other new mechanisms for noise reduction, of independent interest. We generalize the Laplace process of Koufogiannis et al. [94] to continuous time in Section 5.4, thus making the Laplace noise reduction mechanism of Ligett et al. [109] more flexible and adaptive to data-dependent privacy levels. We also briefly mention a noise reduction mechanism for Skellam noise in Section 5.4, a discrete distribution used in count queries [6].

5.2 Preliminaries

Differential privacy, privacy loss, and ex-post privacy. An algorithm $A : \mathcal{X} \to \mathcal{Y}$ is (ϵ, δ)-differentially private if, for any measurable set $E \subset \mathcal{Y}$ and any neighboring inputs $x \sim x'$,

$$\mathbb{P}(A(x) \in E) \leq e^{\epsilon}\mathbb{P}(A(x') \in E) + \delta. \tag{5.2.1}$$

In the above [55], $\sim$ denotes some arbitrary neighboring relation. Typically $x \sim x'$ indicates x and x' differ in one entry, but any other relation suffices. While differential privacy has proven itself a mainstay of private computation, condition (5.2.1) is too rigid to allow data analysts to achieve a minimum desired accuracy. In other words, it embraces a *privacy first* perspective, fixing a strict condition in terms of parameters ϵ and δ that must be met. We are interested in the *accuracy first* perspective, setting a target accuracy and correspondingly optimizing privacy parameters.

The above definition of differential privacy is qualitatively focused on bounding the information-theoretic quantity of *privacy loss* [55, 57, 52].

Definition 5.2.1 (Privacy Loss). Let $A : \mathcal{X} \to \mathcal{Y}$ be an algorithm, and fix neighbors $x \sim x'$ in $\mathcal{X}$. Let p^x and $p^{x'}$ be the respective densities of $A(x)$ and $A(x')$ on the space $\mathcal{Y}$ with respect to some reference measure[2]. Then, the privacy loss between $A(x)$ and $A(x')$ is the random variable

$$\mathcal{L}(x, x') := \log\left(\frac{p^x(A(x))}{p^{x'}(A(x))}\right).$$

We think of $A(x)$ as the true outcome, and $\mathcal{L}(x, x')$ measures how much more likely this outcome is under the true input x versus an alternative x'. Privacy loss provides a *probabilistic* definition of privacy. Namely, A is (ϵ, δ)-*probabilistically differentially private* if, for all neighbors $x \sim x'$,

$$\mathbb{P}\left(\mathcal{L}(x, x') > \epsilon\right) \leq \delta. \tag{5.2.2}$$

While probabilistic differential privacy is not equivalent to differential privacy [85], (ϵ, δ)-probabilistically differential privacy implies (ϵ, δ)-differential privacy. Probabilistic differential privacy emerged as a means for studying privacy composition, and has been leveraged in proving many results [84, 123, 133, 163]. A natural extension of privacy to the accuracy-oriented regime is *ex-post* privacy, which allows the bound in condition (5.2.2) to depend the observed algorithm output.

Definition 5.2.2 (Ligett et al. [109])**.** Let $A : \mathcal{X} \to \mathcal{Y}$ be an algorithm and $\mathcal{E} : \mathcal{Y} \to \mathbb{R}_{\geq 0}$ a function. We say A is $(\mathcal{E}, \delta)$-ex-post private if, for any neighboring inputs $x \sim x'$, we have

$$\mathbb{P}\left(\mathcal{L}(x, x') > \mathcal{E}(A(x))\right) \leq \delta.$$

While any algorithm is trivially ex-post private with $\mathcal{E}(A(x)) := \infty$, the goal is to make $\mathcal{E}(A(x))$ as small as possible. We describe theoretical tools for obtaining ex-post privacy guarantees in Section 5.3, and empirically compute the ex-post privacy distributions of various mechanisms in Section 5.6.

Background on Noise Reduction. Heuristically, a noise reduction mechanism allows a data analyst to view multiple, increasingly accurate estimates of a risk minimizing parameter while only paying an ex-post privacy cost for the *least* noisy iterate observed. Pinning down a general definition of a noise reduction mechanism is difficult, as any definition would need to depend on how the released parameter estimates were produced. In this paper, we consider the relevant case of additive noise mechanisms. Below, we provide an explicit definition of noise reduction mechanisms for this setting.

In the following definition, we let $(A_t)_{t\geq 0}$ be some collection of potentially correlated noise variables. In particular, A_t should be thought of as marginally having either a multivariate normal distribution $\mathcal{N}(0, tI_d)$ or multivariate Laplace distribution $\mathrm{Lap}(t)$. The index t can be viewed as either "time" or "variance", with larger values of t indicating greater variance of noise added. Further, when we refer to a sequence of *time functions* $(T_n)_{n\geq 1}$, we mean a sequence of functions $T_n : (\mathbb{R}^d)^{n-1} \to \mathbb{R}_{>0}$ such that, for all $n \geq 1$ and $\beta_{1:n} \in (\mathbb{R}^d)^n$,

$$T_{n+1}(\beta_{1:n}) \leq T_n(\beta_{1:n-1}). \tag{5.2.3}$$

[2]For instance, if μ_x and $\mu_{x'}$ are the laws of $A(x)$ and $A(x')$ respectively, the reference measure can be taken to be $\mu_x + \mu_{x'}$.

Intuitively, the nth time function gives the adaptively chosen variance of noise that will be added to the nth parameter based on the first $n-1$ observed parameters.

Let $M : \mathcal{X} \to \mathcal{Y}^\infty$ be an algorithm mapping databases for sequences of outputs. Let $M_n : \mathcal{X} \to \mathcal{Y}$ give the nth element of the sequence and $M_{1:n} : \mathcal{X} \to \mathcal{Y}^n$ the first n elements. We assume $M_n(x) := f(x) + A_{T_n(x)}$, where $f : \mathcal{X} \to \mathcal{Y}$ is some function that should be thought of as producing a true, risk-minimizing parameter, $(T_n)_{n\geq 1}$ is a sequence of time functions, and $T_n(x) := T_n(M_{1:n-1}(x))$.

Definition 5.2.3 (Noise Reduction Mechanism). Let $(A_t)_{t\geq 0}$ and $M : \mathcal{X} \to \mathcal{Y}^\infty$ be as above, $a \in \mathcal{Y}$ any constant, and suppose $A_t + a$ has marginal density p_t^a. We say M is a *noise reduction* mechanism if, for any $n \geq 1$ and any neighboring datasets $x \sim x'$, we have

$$\mathcal{L}_{1:n}(x, x') = \frac{p_{T_n(x)}^{f(x)}(M_n(x))}{p_{T_n(x)}^{f(x')}(M_n(x'))},$$

where $\mathcal{L}_{1:n}(x, x')$ denotes the privacy loss between $M_{1:n}(x)$ and $M_{1:n}(x')$.

The only noise reduction mechanism in the literature uses a Markov process with Laplace marginals [94] to gradually release a sensitive parameter [109]. As originally presented, this *Laplace Noise Reduction* mechanism is nonadaptive, requiring a data analyst to fix a finite sequence of privacy parameters $(\epsilon_n)_{n\in[K]}$ in advance. Instead of presenting this method as background, we describe it in Section 5.4, in which we construct an adaptive generalization of this mechanism. We then leverage this generalization as a subroutine in $\mathrm{ReducedAboveThreshold}$, a generalization of $\mathrm{AboveThreshold}$ with adaptive privacy guarantees.

Background on Brownian Motion. We now provide a brief background on Brownian motion, perhaps the best-known example of a continuous time stochastic process [104].

Definition 5.2.4. A continuous time real-valued process $(B_t)_{t\geq 0}$ is called a standard Brownian motion if (1) $B_0 = 0$, (2) $(B_t)_{t\geq 0}$ has continuous sample paths, (3) (B_t) has independent increments, i.e. $B_{t+s} - B_s$ is independent of B_s for all $s, t \geq 0$, and (4) $B_t \sim \mathcal{N}(0, t)$ for all $t \geq 0$.

We say a process $(B_t)_{t\geq 0}$ is a d-dimensional standard Brownian motion if each coordinate process is an independent standard Brownian motion.

We use many properties of Brownian motion to construct the Brownian mechanism and analyze its privacy loss in Section 5.3. One important property of Brownian motion is that it is a continuous time martingale. This property allow us to use time-uniform supermartingale concentration to characterize and bound the privacy loss of the Brownian mechanism at data-dependent stopping times [73, 76]. We do not go into detail about martingale concentration in this background section, but rather defer it to Appendix 5.A. Additionally, $(B_t)_{t\geq 0}$ is a Markov process. This tells us that if we inspect the Brownian motion at times $0 \leq t_1 < t_2 < \cdots < t_n$, then $B_{t_2}, \ldots, B_{t_n}$ can be viewed as a randomized post-processing of B_{t_1} that *does not* depend on B_s for any $s < t_1$. This property allows us to show that the privacy loss of the Brownian mechanism — which adds noise to a parameter via a Brownian motion — only depends on the least noisy parameter observed.

5.3 The Brownian Mechanism: a Gaussian Noise Reduction Mechanism

The Brownian mechanism works by simulating a Brownian motion starting at some multivariate parameter; this parameter should be thought of as the risk-minimizing output if there were no privacy constraints. The data analyst first observes the random walk at some large time. Then, if so desired, the analyst "rewinds" time to an earlier point on the Brownian path, reducing noise to obtain a more accurate estimate. Due to the Markovian nature of Brownian motion, the analyst will only pay a privacy cost proportional to variance of the random walk at the earliest inspected time.

Definition 5.3.1. Let $f : \mathcal{X} \to \mathbb{R}^d$ be a function and $(T_n)_{n\geq 1}$ a sequence of time functions. Let $(B_t)_{t\geq 0}$ be a standard d-dimensional Brownian motion. The Brownian mechanism associated with f and $(T_n)_{n\geq 1}$ is the algorithm $\mathrm{BM} : \mathcal{X} \to (\mathbb{R}^d)^\infty$ given by

$$\mathrm{BM}(x) := \left(f(x) + B_{T_n(x)}\right)_{n\geq 1},$$

where we set $T_n(x) := T_n\left(f(x) + B_{T_1(x)}, \ldots, f(x) + B_{T_{n-1}(x)}\right)$ with $T_1(x)$ being constant.

We have chosen $T_n(x)$ as indexing notation to denote dependence on x, even if this is only through observed parameters. In the context of ERM, one can think of f as computing a risk minimizing parameter associated with a private dataset $x \in \mathcal{X}$. The data analyst uses T_n along with the previous iterate to determine how far to rewind time to obtain the nth iterate.

The Brownian mechanism, as defined above, produces an infinite sequence of parameters. In practice, a data analyst will only view finitely many iterates, stopping when some utility condition has been met or a minimum privacy level is reached. We introduce *stopping functions* to model how a data analyst adaptively interacts with noise reduction mechanisms.

Definition 5.3.2 (Stopping Function). Let $M : \mathcal{X} \to \mathcal{Y}^\infty$ be a an algorithm. For $x \in \mathcal{X}$, let $(\mathcal{F}_n(x))_{n\in\mathbb{N}}$ be the filtration given by $\mathcal{F}_n(x) := \sigma(M_i(x) : i \leq n)$.[3] A function $N : \mathcal{Y}^\infty \to \mathbb{N}$ is called a stopping function if for any $x \in \mathcal{X}$, $N(x) := N(M(x))$ is a stopping time with respect to $(\mathcal{F}_n(x))_{n\geq 1}$.

A stopping function N is a rule used to decide when to stop viewing parameters that *only* depends on the observed iterates of the noise reduction mechanism. N could heuristically be "stop at the first time a parameter achieves an accuracy of 95% on a held-out dataset." Recall from Figure 5.1 and equation (5.2.3) that the later iterations of BM correspond to smaller noise variances, meaning that T_n is a decreasing sequence in the number of iterations n. Further, the filtration $\mathcal{F}$ defined above is quite different from the usual filtrations considered for Brownian motions. In some cases, an analyst may want the stopping function to depend on the underlying private dataset through more than just the released parameters, e.g. they may want their rule to be "stop at the first time a parameter achieves an accuracy of 95% on the private dataset." In this

[3]The notation $\sigma(X)$ denotes the σ-algebra generated by X. N is said to be a stopping time with respect to (X_n) if $\{N \leq n\} \in \sigma(X_m : m \leq n)$ for all $n \in \mathbb{N}$. This definition can be extended to allow for N to depend on independent, external randomization, but we omit this for simplicity.

case, additional privacy may be lost due to observing $N(x)$. We detail how to handle this more subtle case in Section 5.5.

Due to the Markovian nature of Brownian motion, we get the following lemma. We include a proof in Appendix 5.B for completeness.

Lemma 5.3.3. *Let $x \sim x'$ be neighbors and $(T_n)_{n\geq 1}$ a sequence of time functions. Then, for any $n \geq 1$, letting $\mathcal{L}^{\mathrm{BM}}_{1:n}(x, x')$ denote the privacy loss between $\mathrm{BM}_{1:n}(x)$ and $\mathrm{BM}_{1:n}(x')$, we have*

$$\mathcal{L}^{\mathrm{BM}}_{1:n}(x, x') = \log\left(\frac{p^{f(x)}_{T_n(x)}(\mathrm{BM}_n(x))}{p^{f(x')}_{T_n(x)}(\mathrm{BM}_n(x))}\right),$$

where p^{μ}_t is the density of a $\mathcal{N}(\mu, tI_d)$ random variable. Furthermore, the above equality holds if n is replaced by an almost surely bounded stopping function $N(x)$.

Lemma 5.3.3 just tells us that the Brownian mechanism is a noise reduction mechanism, i.e. that the privacy lost by viewing the first n iterates is exactly the privacy lost by viewing the nth iterate in isolation.

The following theorem characterizes the privacy loss of the Brownian mechanism.

Theorem 5.3.4. *Let BM be the Brownian mechanism associated with $(T_n)_{n\geq 1}$, a function $f : \mathcal{X} \to \mathbb{R}^d$, and stopping function N. For neighbors $x \sim x'$, the privacy loss between $\mathrm{BM}_{1:N(x)}(x)$ and $\mathrm{BM}_{1:N(x')}(x')$ is given by*

$$\mathcal{L}^{\mathrm{BM}}_{1:N(x)}(x, x') = \frac{||f(x) - f(x')||_2^2}{2T_{N(x)}(x)} + \frac{||f(x) - f(x')||_2}{T_{N(x)}(x)} W_{T_{N(x)}(x)},$$

where $(W_t)_{t\geq 0}$ is a standard, univariate Brownian motion. Suppose f has ℓ_2-sensitivity at most Δ_2. Then, letting $a^+ := \max(0, a)$, we have

$$\mathcal{L}^{\mathrm{BM}}_{1:N(x)}(x, x') \leq \frac{\Delta_2^2}{2T_{N(x)}(x)} + \frac{\Delta_2}{T_{N(x)}(x)} W^+_{T_{N(x)}(x)}.$$

The above theorem can be viewed as a process-level equivalent of the well-known fact that the privacy loss of the Gaussian mechanism has an uncentered Gaussian distribution [11]. We prove the Theorem 5.3.4 in Appendix 5.B. Given the clean characterization of privacy loss above, we now show how to construct high-probability, time-uniform privacy loss bounds. We define *privacy boundaries*, which map the variance of BM to high-probability bounds on privacy loss.

Definition 5.3.5. A function $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a *δ-privacy boundary* for the Brownian mechanism associated with time functions $(T_n)_{n\geq 1}$ if for any neighboring datasets $x \sim x'$, we have

$$\mathbb{P}\left(\exists n \geq 1 : \mathcal{L}^{\mathrm{BM}}_{1:n}(x, x') \geq \psi(T_n(x))\right) \leq \delta$$

Since the privacy loss of BM is a deterministic function of a Brownian motion, we can apply results from martingale theory to construct general families of privacy boundaries.

Theorem 5.3.6. *Assume the same setup as in Theorem 5.3.4. Let $\delta > 0$ and f be a function with ℓ_2-sensitivity Δ_2. The following classes of functions form δ-privacy boundaries.*

1. ***(Mixture boundary)*** *For any $\rho > 0$, ψ_ρ^M given by*

$$\psi_\rho^M(t) := \frac{\Delta_2^2}{2t} + \frac{\Delta_2}{t}\sqrt{2(t+\rho)\log\left(\frac{1}{\delta}\sqrt{\frac{t+\rho}{\rho}}\right)}.$$

2. ***(Linear boundary)*** *For any $a, b > 0$ such that $2ab = \log(1/\delta)$, $\psi_{a,b}^L$ given by*

$$\psi_{a,b}^L(t) := \frac{\Delta_2}{t}\left(\frac{\Delta_2}{2} + b\right) + \Delta_2 a.$$

We prove Theorem 5.3.6 in Appendix 5.B. In the same appendix, we plot the boundaries in Figure 5.4.

Privacy boundaries serve a dual purpose for the Brownian mechanism. First, since time-uniform concentration bounds are valid at arbitrary data-dependent times, that need not be stopping times with respect to the standard forward Brownian Motion filtration [76], privacy boundaries provide ex-post privacy guarantees. Second, in many settings, it may be more natural for a data analyst to adaptively specify target privacy levels instead of noise levels. This is, for instance, the case in our experiments in Section 5.6. By inverting privacy boundaries, data analysts can compute the proper amount of noise to remove at each step to meet target privacy levels.

We make the above precise in Corollary 5.3.7. In what follows, when we refer to a sequence $(\mathcal{E}_n)_{n\geq 1}$ of *privacy functions*, we mean a sequence of functions $\mathcal{E}_n : (\mathbb{R}^d)^{n-1} \to \mathbb{R}_{\geq 0}$ such that, for all n and $\beta_{1:n} \in (\mathbb{R}^d)^n$, $\mathcal{E}_{n+1}(\beta_{1:n}) \geq \mathcal{E}_n(\beta_{1:n-1})$.

Corollary 5.3.7. *Let N be a stopping function, as in Definition 5.3.2. If ψ is a δ-privacy boundary for* BM*, we have*

$$\sup_{x\sim x'} \mathbb{P}\left(\mathcal{L}_{N(x)}^{\mathrm{BM}}(x, x') \geq \psi\left(T_{N(x)}(x)\right)\right) \leq \delta,$$

i.e. the algorithm $\mathrm{BM}_{1:N(\cdot)}(\cdot)$ is $\big(\psi(T_{N(\cdot)}(\cdot)), \delta\big)$-ex post private, where $(\cdot)$ denotes a positional argument for an input $x \in \mathcal{X}$. Further, let $(\mathcal{E}_n)_{n\geq 1}$ be a sequence of privacy functions, and define

$$T_n(\beta_{1:n-1}) := \inf\left\{t \geq 0 : \psi(t) \geq \mathcal{E}_n(\beta_{1:n-1})\right\}.$$

Then $\mathrm{BM}_{1:N(\cdot)}(\cdot)$ is $(\mathcal{E}_{N(\cdot)}(\cdot), \delta)$-ex post private, where $\mathcal{E}_n(x)$ is defined analogously to $T_n(x)$.

Again, N should be thought of as a stopping rule based on parameter accuracy. $\mathcal{E}_n$ should be thought of as a rule for choosing the nth privacy parameter given $\mathrm{BM}_{1:n-1}(x)$.

5.4 An Adaptive, Continuous-Time Extension of Laplace Noise Reduction

Here, we generalize the original noise reduction mechanism of Ligett et al. [109], which will be used as a subroutine in Algorithm 1 in the following section. We first describe the original Laplace-based Markov process of Koufogiannis et al. [94]. Fix any positive integer K and any

finite, increasing sequence of times $(t_n)_{n\in[K]}$. Let $(\zeta_n)_{n=0}^K$ be the d-dimensional process given by $\zeta_0 = 0$ and

$$\zeta_n = \begin{cases} \zeta_{n-1} & \text{with probability } \left(\frac{t_{n-1}}{t_n}\right)^2 \\ \zeta_{n-1} + \mathrm{Lap}(t_n) & \text{otherwise.} \end{cases} \tag{5.4.1}$$

Koufogiannis et al. [94] show that $\zeta_n \sim \mathrm{Lap}(t_n)$ and that $(\zeta_n)_{n=0}^K$ is Markovian. Ligett et al. [109] use the above process to construct a noise reduction mechanism. Namely, they define the the *Laplace Noise Reduction* mechanism associated with $f : \mathcal{X} \to \mathbb{R}^d$ and $(t_n)_{n\in[K]}$ to be the algorithm $\mathrm{LNR} : \mathcal{X} \to (\mathbb{R}^d)^K$ given by $\mathrm{LNR}(x) := (f(x) + \zeta_K, \ldots, f(x) + \zeta_1)$. If $t_n := \Delta_1/\epsilon_n$, then releasing nth component $\mathrm{LNR}_n(x)$ in isolation is equivalent to running the classical Laplace mechanism with privacy level ϵ_n.

We now extend the process $(\zeta_n)_{n\in[K]}$ to a continuous time process with the same finite-dimensional distributions. Let $\eta > 0$ be arbitrary, and let $(P_t)_{t\geq\eta}$ be an inhomogeneous Poisson process with intensity function $\lambda(t) := \frac{2}{t}$. For $n \geq 1$, let $\mathcal{T}_n := \inf\{t \geq \eta : P_t \geq n\}$ be the nth jump of $(P_t)_{t\geq\eta}$ and set $\mathcal{T}_0 := \eta$. Noting that P_t must be a nonnegative integer, define the process $(Z_t)_{t\geq\eta}$ by

$$Z_t := \sum_{n=0}^{P_t} \mathrm{Lap}(\mathcal{T}_n). \tag{5.4.2}$$

It is immediate that $(Z_t)_{t\geq\eta}$ is Markovian. We show in Appendix 5.D that $Z_t \sim \mathrm{Lap}(t)$. With $(Z_t)_{t\geq\eta}$, one can make LNR fully adaptive, meaning that the times $(t_n)_{n\in[K]}$ at which it is invoked need not be prespecified, and can depend on the underlying input database x by using time functions.

Definition 5.4.1. Let $f : \mathcal{X} \to \mathbb{R}^d$ be a function and $(T_n)_{n\geq1}$ a sequence of time functions. Let $(Z_t)_{t\geq\eta}$ be the process defined in Equation (5.4.2). The *Laplace noise reduction* mechanism associated with f and $(T_n)_{n\geq1}$ is the algorithm $\mathrm{LNR} : \mathcal{X} \to (\mathbb{R}^d)^\infty$ given by

$$\mathrm{LNR}(x) := (f(x) + Z_{T_n(x)})_{n\geq1},$$

where again $T_n(x) := T_n(f(x) + Z_{T_1(x)}, \ldots, f(x) + Z_{T_{n-1}(x)})$ and $T_1(x)$ is constant.

If the analyst would prefer instead to specify privacy functions $(\mathcal{E}_n)_{n\geq1}$, they can do so by leveraging the corresponding time functions $T_n(x) := \Delta_1/\mathcal{E}_n(x)$, where $\mathcal{E}_n(x)$ is defined analogously to $T_n(x)$. We leverage LNR in our experiments in Section 5.6 and the process $(Z_t)_{t\geq0}$ as a subroutine in constructing $\mathrm{ReducedAboveThreshold}$. An analogous argument to the one used in proving Lemma 5.3.3 can be used to show LNR enjoys the following ex-post privacy guarantee.

Proposition 5.4.2. *Let* LNR *be associated with* $(T_n)_{n\geq1}$ *and a function* f *with* ℓ_1*-sensitivity* Δ_1*. If* N *is stopping function, the algorithm* $\mathrm{LNR}_{1:N(\cdot)}(\cdot)$ *is* $(\Delta_1/T_{N(\cdot)}(\cdot), 0)$*-ex post private.*

Skellam Noise Reduction. Last, we briefly discuss how to generate a noise reduction mechanism for Skellam noise [6]. Recall that a random variable X has a Skellam distribution with parameters λ_1 and λ_2 if $X =_d Y_1 - Y_2$, where $Y_1 \sim \mathrm{Poisson}(\lambda_1)$ and $Y_2 \sim \mathrm{Poisson}(\lambda_2)$ are independent Laplace random variables. For succinctness, we write $X \sim \mathrm{Skell}(\lambda_1, \lambda_2)$.

Let $(P_1(t))_{t\geq 0}$ and $(P_2(t))_{t\geq 0}$ be two independent, homogeneous Poisson process with rates λ_1 and λ_2 respectively. Observe that the continuous time process $(X_t)_{t\geq 0}$ given by $X_t := P_1(t) - P_2(t)$ is clearly Markovian, has independent increments, and has $X_t \sim \text{Skell}(t\lambda_1, t\lambda_2)$. Thus, $(X_t)_{t\geq 0}$ can be used to define a Skellam noise reduction mechanism by releasing $(f(x) + X_{T_n(x)})_{n\geq 1}$ for some sequence of time functions $(T_n)_{n\geq 1}$.

5.5 Privately Checking if Accuracy is Above a Threshold

In Section 5.3 we presented the Brownian mechanism, characterized its privacy loss, and showed how to obtain ex-post privacy guarantees for arbitrary stopping functions. In particular, these stopping functions could be based on the accuracy of the observed iterates on public held-out data.

However, one may desire to privately check the accuracy of observed iterates on the dataset $x \in \mathcal{X}$. [109] were able to accomplish this goal by coupling LNR with AboveThreshold, a classical algorithm for privately answering threshold queries [52]. In the context of ERM, AboveThreshold iteratively checks if the empirical risk of each parameter is below a target threshold, stopping at the first such occurrence. The downside to AboveThreshold is that it requires a prefixed privacy level. In empirical studies, Ligett et al. [109] found this fixed privacy cost dominated the ex-post privacy guarantees, showing little benefit to using noise reduction.

Below, we construct ReducedAboveThreshold, a generalization of AboveThreshold which provides ex-post privacy guarantees. We show how to couple BM with ReducedAboveThreshold to obtain tighter ex-post privacy guarantees than coupling with AboveThreshold would permit. In particular, if BM is run using parameters $(\epsilon_n)_{n\geq 1}$ and ReducedAboveThreshold indicates the Nth parameter obtains sufficiently high accuracy, the privacy loss of the net procedure will be at most $2\epsilon_N$ — only twice the privacy loss that would be accrued by testing on public data.

Algorithm 1 ReducedAboveThreshold (via Laplace Noise Reduction)

Input: Algorithm $\text{Alg} : \mathcal{X} \to \mathcal{Y}^\infty$, parameter $\epsilon_{\max} > 0$, threshold τ, database $x \in \mathcal{X}$, utility $u : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$ where $u(\beta, \cdot)$ is Δ-sensitive $\forall \beta$, privacy functions $(\mathcal{E}_n)_{n\geq 1}$ with $\mathcal{E}_n \leq \epsilon_{\max}$ $\forall n$.

for $n \geq 1$ **do**

$\epsilon_n := \mathcal{E}_n(\text{Alg}_{1:n-1}(x))$, $T_n := 2\Delta/\epsilon_n$

$\zeta_n := Z_{T_n}$, where $(Z_t)_{t\geq \eta}$ in Eq. (5.4.2) defines the LNR mechanism with $\eta := 2\Delta/\epsilon_{\max}$.

$\xi_n \sim \text{Lap}\left(\frac{4\Delta}{\epsilon_n}\right)$

if $u(\text{Alg}_n(x), x) + \xi_n \geq \tau + \zeta_n$ **then**

Print 1 and HALT

else

Print 0

τ should be seen as a target accuracy, Alg as a mechanism for releasing a parameter (e.g. BM, LNR), and u as evaluating the accuracy of $\text{Alg}_n(x)$ on x. $\epsilon_{\max}$ is an arbitrarily large constant, representing the minimum level of privacy required, used to prevent the user from examining

(Z_t) at arbitrarily small times. The above generalizes to sequences of thresholds $(\tau_n)_{n\geq 1}$ and sequences $(u_n)_{n\geq 1}$ of functions $u_n : \mathcal{Y}^n \times \mathcal{X} \to \mathbb{R}$ that are Δ-sensitive in their second argument, but the added generality yields only marginal benefits. When $\mathcal{E}_n = \epsilon$ for all n, Algorithm 1 recovers AboveThreshold as a special case. The intuition behind ReducedAboveThreshold is that by gradually removing Laplace noise from the threshold, a data analyst can ensure that privacy of the whole procedure only depends on the magnitude of Laplace noise added when the algorithm halts. The following characterizes the privacy loss of Algorithm 1.

Theorem 5.5.1. *For any $n \geq 1$ and neighboring datasets $x \sim x'$, let $\mathcal{L}^{\mathrm{Alg}}_{1:n}(x, x')$ denote the privacy between $\mathrm{Alg}_{1:n}(x)$ and $\mathrm{Alg}_{1:n}(x')$. For any $x \in \mathcal{X}$, define $N(x)$ to be the first round where* ReducedAboveThreshold *run on input $x \in \mathcal{X}$ outputs 1, that is*

$$N(x) := \inf\{n \geq 1 : \mathrm{ReducedAboveThreshold}_n(x) = 1\}.$$

Then, the privacy loss between $\mathrm{ReducedAboveThreshold}(x)$ *and* $\mathrm{ReducedAboveThreshold}(x')$*, denoted $\mathcal{L}^{\mathrm{RAT}}(x, x')$, is bounded by*

$$\mathcal{L}^{\mathrm{RAT}}(x, x') \leq \mathcal{L}^{\mathrm{Alg}}_{1:N(x)}(x, x') + \mathcal{E}_{N(x)}(\mathrm{Alg}_{1:N(x)-1}(x)).$$

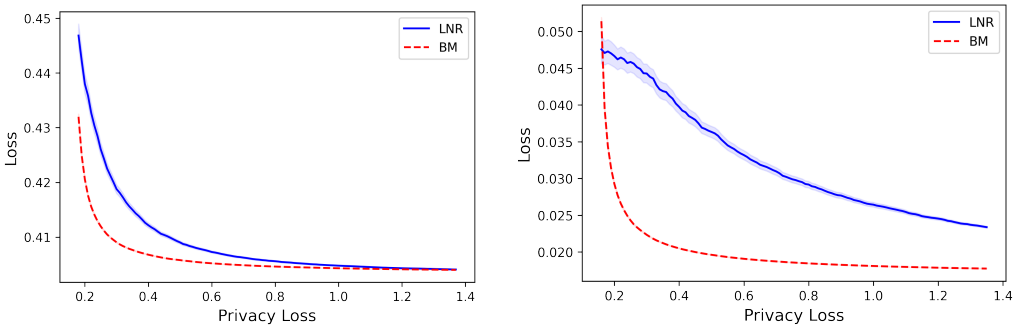
We prove Theorem 5.5.1 in Appendix 5.C, where we also provide a utility guarantee for ReducedAboveThreshold. This utility guarantee, much like the utility guarantee for AboveThreshold, is in practice weak as it derives from a union bound. Using Theorem 5.5.1, we can simply choose $\mathrm{Alg} = \mathrm{BM}$ as a means of adaptively generating parameters. The following corollary, which follows immediately from the above theorem, provides the ex-post privacy guarantees of combining ReducedAboveThreshold and BM.

Corollary 5.5.2. *Let* BM *be the Brownian mechanism associated with a function f, decreasing time functions $(T_n)_{n\geq 1}$, and a a δ-privacy boundary ψ. Let* ReducedAboveThreshold *be run with privacy functions $(\psi(T_n))_{n\geq 1}$, threshold τ, and algorithm* BM*. Then,* ReducedAboveThreshold *is $\big(2\psi(T_{N(\cdot)}(\cdot)), \delta\big)$-ex post private.*

5.6 Experiments

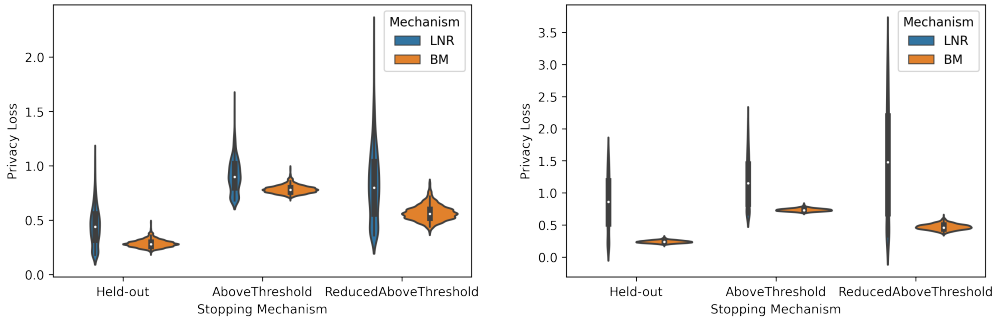
Choice of tasks: We compare the performance of BM and LNR on the tasks of regularized logistic regression via output perturbation [25] and ridge regression via covariance perturbation [145].[4] For logistic regression, we leveraged the KDD-99 dataset [90] with $d = 38$ features, predicting whether network events can be classified as "normal" or "malicious". For ridge regression, we used the Twitter dataset [89] with $d = 77$ features to predict log-popularity of posts. In each case, we ran our experiments on $n = 10,000$ randomly sub-sampled data points. In order to guarantee bounded sensitivity, we normalized each data point to have unit ℓ_2 norm. We note that this aspect differs from the experimentation conducted by Ligett et al. [109], who normalized by the *maximum* ℓ_2 norm, a non-private operation.

[4]The two tasks use the logistic loss $\ell(y, z) := \log(1 + \exp(-yz))$ and the squared loss $\ell(y, z) := \frac{1}{2}(z - y)^2$. The regularized loss on a dataset $\mathcal{D} := \{(x_i, y_i)\}_{i\in[n]}$ is $L(\beta, \mathcal{D}) := \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \beta^T x_i) + \frac{\lambda||\beta||_2^2}{2}$.



(a) Regularized Logistic Regression (b) Ridge Regression

Figure 5.2: Privacy loss plotted against loss (respectively regularized logistic and ridge loss) for the statistical tasks of regularized logistic regression and ridge regression.



(a) Regularized Logistic Regression (b) Ridge Regression

Figure 5.3: Empirical privacy loss distributions for logistic regression and ridge regression with loss assessed either (left) on the training data treated as a public, held-out dataset, (middle) via AboveThreshold, or (right) via ReducedAboveThreshold.

Experiments: For each task, we conducted two experiments. We discuss the specific parameter settings for these experiments in Appendix 5.E. In the first experiment, we plotted guaranteed (in the case of LNR) or high-probability (in the case of BM) privacy loss on the x-axis against average loss (either logistic or ridge) on the y-axis. We conduct such a comparison as probability 1 privacy loss bounds cannot be provided for the Gaussian mechanism. Likewise, adding a probability δ of minimally improves privacy loss for the Laplace mechanism. We computed the average loss curve for each mechanism over 1,000 trials, and have included point-wise valid 95% confidence intervals.

In the second experiment, we plotted the empirical privacy loss distributions for BM and LNR under the stopping conditions of loss being at most 0.41 for logistic regression and 0.025 for ridge regression. For each mechanism, we evaluated this empirical distribution using three approaches for testing empirical loss: treating the training data as a held-out dataset, using AboveThreshold, and using our mechanism, ReducedAboveThreshold. In AboveThreshold, we set the privacy parameter to be fixed at $\epsilon = 0.5$. In ReducedAboveThreshold, we took the sequence of privacy parameters to be the same as the sequence of privacy parameters used by BM and LNR. We once

again computed these empirical distributions over 1,000 runs of each mechanism.

Findings: The findings of the two experiments are summarized in Figure 5.2 and Figure 5.3. For both tasks, BM obtains significant improvements in loss over LNR near the privacy loss level that was optimized for. For both tasks, the privacy loss distribution for BM has lower median privacy loss than that of LNR. In addition, the privacy loss distribution for BM is more tightly concentrated around the median, indicating more consistent performance. The privacy loss distribution for LNR has a heavy tail, demonstrating that many runs do not attain the target loss until high privacy loss costs are incurred. Comparing $\mathrm{ReducedAboveThreshold}$ and $\mathrm{AboveThreshold}$, we see that the privacy loss distribution for $\mathrm{ReducedAboveThreshold}$ has higher variance than that of $\mathrm{AboveThreshold}$. However, $\mathrm{ReducedAboveThreshold}$ attains a significantly lower median level of privacy loss when coupled with BM. This latter point reflects the observations of Ligett et al. [109], who note that when $\mathrm{AboveThreshold}$ is used to determine stopping conditions on private data, it contributes the bulk of the privacy loss to the empirical distributions. On the other hand, our figures demonstrate that $\mathrm{ReducedAboveThreshold}$ results in a more mild privacy loss at target stopping conditions.

5.7 Conclusion

In this paper, we constructed the Brownian mechanism (BM), a novel approach to noise reduction that adds noise to a hidden parameter via a Brownian motion. We not only precisely characterized the privacy loss of the Brownian mechanism, but also bounded it through applying machinery from continuous time martingale theory. We then demonstrated how the utility of the iterates produced by BM can be assessed on private data via $\mathrm{ReducedAboveThreshold}$, a generalization of the classical $\mathrm{AboveThreshold}$ algorithm. This was itself accomplished by a continuous-time generalization of the original Laplace noise reduction (LNR) mechanism. Last, we empirically demonstrated that BM outperforms LNR on common statistical tasks, such as regularized logistic and ridge regression.

We comment on several limitations and open problems related to our work. We considered noise reduction mechanisms in the setting of one-shot privacy, in which only a single mechanism is run on private data. Traditional composition results, such as those for fixed privacy parameters [57, 84, 123] or adaptively selected parameters [133, 62, 163] are not directly applicable to algorithms satisfying ex-post privacy; additional machinery needs to be developed to handle composition in this case. A naive approach to composition is possible, which involves summing the ex-post privacy guarantees of composed algorithms and summing the corresponding δ's, but we expect this approach to be loose. Finally, noise reduction is currently only applicable to output perturbation methods; it remains open to see how to combine noise reduction with other prominent methods for private computation, such as objective perturbation.

5.A Background on Martingale Concentration

In this section, we provide a background on the basics of martingale concentration needed throughout this paper. While standard Brownian motion $(B_t)_{t\geq 0}$ is not a nonnegative super-

martingale, geometric Brownian motion given by $Y_t^\lambda := \exp\left(\lambda B_t - \frac{\lambda^2}{2}t\right)$ is a nonnegative martingale for any $\lambda \in \mathbb{R}$, and hence Ville's inequality (Theorem 1.0.2) can be applied. In fact, the probability in the lemma above becomes exactly δ when it is applied to a nonnegative martingale with continuous paths like Y_t^λ. From Ville's inequality, the following *line-crossing inequality* for Brownian motion can be obtained.

Lemma 5.A.1 (Line-Crossing Inequality)**.** *For $\delta \in (0,1)$ and $a, b > 0$ satisfying $e^{-2ab} = \delta$, we have*

$$\mathbb{P}\left(\exists t \geq 0 : B_t \geq at + b\right) = \delta.$$

A proof of the above fact can be found in any standard book on continuous time martingale theory [104, 50]. The above also follows from a special case of the more general time-uniform Chernoff bound presented in Howard et al. [73], as discussed earlier in this document.

The above inequality can be seen as optimizing the tightness of the time-uniform boundary at one preselected point in time. However, due to the adaptive nature of the Brownian mechanism presented in Section 5.3, it is sometimes desirable to construct a time-uniform boundary which sacrifices tightness at a fixed point in time to obtain greater tightness over all of time.

The aforementioned *method of mixtures* provides one such approach for constructing tighter time-uniform boundaries [86, 76]. We discuss this concept briefly in the context of Brownian motion. Observe that, since $(Y_t^\lambda)_{t\geq 0}$ is a nonnegative martingale, for any probability measure π on $\mathbb{R}$, the process $(X_t^\pi)_{t\geq 0}$ given by

$$X_t^\pi := \int_{\mathbb{R}} Y_t^\lambda \pi(d\lambda)$$

is also nonnegative martingale. By appropriately choosing the probability measure π and applying Ville's inequality, one obtains the following concentration inequality [76].

Lemma 5.A.2 (Mixture Inequality)**.** *Let $\rho > 0$ and $\delta \in (0,1)$ be arbitrary. Then,*

$$\mathbb{P}\left(\exists t \geq 0 : B_t \geq \sqrt{2(t+\rho)\log\left(\frac{1}{\delta}\sqrt{\frac{t+\rho}{\rho}}\right)}\right) = \delta.$$

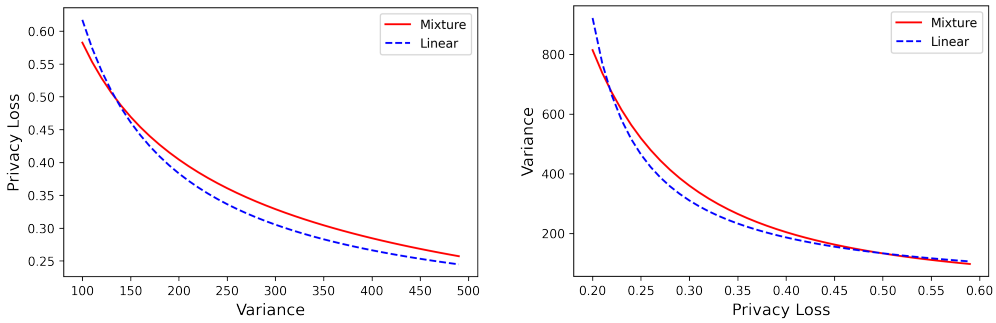
We leverage Lemmas 5.A.1 and 5.A.2 to construct the privacy boundaries in Theorem 5.3.6 in Appendix 5.B.

5.B Proofs From Section 5.3

Here, we prove the results from Section 5.3. We start by showing that BM is in fact a noise-reduction mechanism, which is claimed in Lemma 5.3.3. To prove the cited lemma, it suffices to show the following result.

Proposition 5.B.1. *For $\nu \in \mathbb{R}^d$, let $(B_t^\nu)_{t\geq 0}$ be a standard d-dimensional Brownian motion starting at ν. Let $(T_n)_{n\geq 1}$ be a sequence of decreasing time functions*[5] *$T_n : \mathbb{R}^{(n-1)d} \to \mathbb{R}$,*

[5] As before, T_1 is implicitly a constant, independent of $(B_t^\nu)_{t\geq 0}$



(a) Variance of Noise vs. Privacy Loss (b) Privacy Loss vs. Variance of Noise

Figure 5.4: A comparison of the linear and mixture boundaries, both optimized for tightness at $\epsilon = 0.3$ with $\delta = 10^{-6}$. The first plot directly plots the corresponding bounds as in Theorem 5.3.6. The second plot inverts the boundaries, showing the variance necessary to meet a target privacy level.

$N : \mathbb{R}^\infty \to \mathbb{N}$ a bounded stopping function, and define $T_n^\nu := T_n\left(B^\nu_{T_1^\nu}, \ldots, B^\nu_{T_{n-1}^\nu}\right)$ and $N^\nu := N\left((B^\nu_{T_n^\nu})_{n\geq 1}\right)$. Let $p^\nu_{1:N}$ denote the joint density of $\left(B^\nu_{T_1^\nu}, \ldots, B^\nu_{T_{N^\nu}^\nu}\right)$. Then, with probability 1, we have

$$\frac{p^\nu_{1:N}\left(B^\nu_{T_1^\nu}, \ldots, B^\nu_{T_{N^\nu}^\nu}\right)}{p^\mu_{1:N}\left(B^\nu_{T_1^\nu}, \ldots, B^\nu_{T_{N^\nu}^\nu}\right)} = \frac{\exp\left(-\frac{(B^\nu_{T_{N^\nu}^\nu}-\nu)^2}{2T_{N^\nu}}\right)}{\exp\left(-\frac{(B^\nu_{T_{N^\nu}^\nu}-\mu)^2}{2T_{N^\nu}}\right)},$$

which is just the ratio between the density of a $\mathcal{N}(\nu, T^\nu_{N^\nu})$ random variable and a $\mathcal{N}(\mu, T^\nu_{N^\nu})$ random variable evaluated at $B^\nu_{T_{N^\nu}^\nu}$.

A key part of proving the above proposition will be developing a strong Markov property for Brownian bridges. Recall that a *Brownian bridge* is, in essence, a Brownian motion that has been "pinned down" at some initial and terminating value. More rigorously, for a random variable $A \in \mathbb{R}^d$ and a constant $b \in \mathbb{R}^d$, a Brownian bridge $(X_t)_{0\leq t\leq T}$ with initial value $X_0 = A$ and terminating value $X_T = b$ is a process that can be written in the form $X_t = \frac{T-t}{T}A + B_t - \frac{t}{T}(B_T - b)$, where $(B_t)_{0\leq t\leq T}$ is a standard d-dimensional Brownian motion that is independent of A. The following properties of Brownian bridges follow from the definition.

Lemma 5.B.2 (Properties of Brownian Bridges)**.** *Let $(X_t)_{0\leq t\leq T}$ be a d-dimensional Brownian bridge with $X_0 = A$, for A being a random vector in $\mathbb{R}^d$, and $X_1 = b$, with $b \in \mathbb{R}^d$ fixed. Then, the following hold:*

1. *If $A' \in \mathbb{R}^d$ is independent of $(X_t)_{t\geq 0}$, A and $b' \in \mathbb{R}^d$ is constant, the process $(X'_t)_{0\leq t\leq T}$ given by*

$$X'_t := X_t + \frac{T-t}{T}A' + \frac{t}{T}b'$$

is a d-dimensional Brownian bridge on $[0, T]$ with initial value $X'_0 = A + A'$ and terminating value $X'_T = b + b'$.

2. $\mu(t) := \mathbb{E}X_t = \frac{T-t}{T}\mathbb{E}A + \frac{t}{T}b$ *for all* $0 \leq t \leq T$.
3. $k(s,t) := \mathrm{Cov}(X_s, X_t) = \frac{(T-t)(T-s)}{T^2}\mathrm{Cov}(A) + \left(s \wedge t - \frac{st}{T}\right) I_d$.
4. *For any* $C > 0$, *the process* $(X'_t)_{0 \leq t \leq CT}$ *given by* $X'_t := \sqrt{C}X_{t/C}$ *is a* d*-dimensional Brownian bridge with initial point* $\sqrt{C}A$ *and terminal point* $\sqrt{C}b$ *on* $[0, CT)$.
5. *If* $A \sim \mathcal{N}(\mu, \Sigma)$, *then* $(X_t)_{0 \leq t \leq T}$ *is a continuous Gaussian process on* $[0,T]$, *and hence it's law is uniquely determined by* μ *and* k.

If $(B_t)_{t \geq 0}$ is a d-dimensional Brownian motion and τ is a stopping time with respect to the natural filtration $(\mathcal{F}_t)_{t \geq 0}$, the strong Markov property for Brownian motion (see Theorem 2.20 of Le Gall [104]) tells us that the process $(B_{\tau+t} - B_\tau)_{t \geq 0}$ is also a d-dimensional Brownian motion that is independent of $\mathcal{F}_\tau$. While we need to be a little more careful with scaling in the setting of Brownian bridges, we can show a similar strong Markov property.

Lemma 5.B.3. *Let* $(X_t)_{0 \leq t \leq 1}$ *be a standard* d*-dimensional Brownian bridge with* $X_0 = A$ *and* $X_1 = b$, *and let* $(\mathcal{G}_t)_{0 \leq t \leq 1}$ *be the corresponding natural filtration. Let* τ *be a* $(\mathcal{G}_t)$ *stopping time. Let* $(X_t^{(\tau)})_{0 \leq t \leq 1-\tau}$ *be the process defined by* $X_t^{(\tau)} := X_{t+\tau} - \frac{1-\tau-t}{1-\tau}X_\tau - \frac{t}{1-\tau}b$, *and define the rescaled process* $(Y_t^{(\tau)})_{0 \leq t \leq 1}$ *by*

$$Y_t^{(\tau)} := \sqrt{1-\tau}X^{(\tau)}_{t/(1-\tau)}.$$

Then, $(Y_t)_{0 \leq t \leq 1}$ *is a standard Brownian bridge with* $Y_0 = Y_1 = 0$ *independent of* $\mathcal{G}_\tau$.

Proof. **Step 1: reduction to the case** $a = b = 0$**:** First, we note that it suffices to prove the result when $A = a$ is a constant. If we prove the result in this case, we note we have by the tower rule for conditional expectations that, for any event E,

$$\mathbb{P}(Y^{(\tau)} \in E) = \mathbb{E}\left[\mathbb{P}\left(Y^{(\tau)} \in E \mid A\right)\right] = \mathbb{E}\left[\mathbb{P}\left(Z \in E\right)\right] = \mathbb{P}\left(Z \in E\right),$$

where $(Z_t)_{0 \leq t \leq 1}$ is a Brownian bridge with $Z_0 = Z_1 = 0$. Next, note it suffices to prove the result in the case $a = b = 0$. Let $(X_t)_{0 \leq t \leq 1}$ be a Brownian bridge satisfying $X_0 = a$ and $X_1 = b$. Define another process $(X'_t)_{t \geq 0}$ on the same probability space by $X'_t := X_t - (1-t)a - tb$. By the first part of Lemma 5.B.2, $(X'_t)_{t \geq 0}$ is a Brownian bridge on $[0,1]$ with initial point $X'_0 = 0$ and $X'_1 = 0$. Clearly, the natural filtration $(\mathcal{G}_t)_{0 \leq t \leq 1}$ for $(X_t)_{0 \leq t \leq 1}$ is also the natural filtration for $(X'_t)_{0 \leq t \leq 1}$. Further, a simple calculation yields that for any fixed $0 \leq s \leq t \leq 1$, $X_t^{(s)'} = X_t^{(s)}$. Thus it also follows that $Y_t^{(\tau)'} = Y_t^{(\tau)}$ for all $(\mathcal{G}_t)$ stopping times τ and all $0 \leq t \leq 1$.

Step 2: considering when $\tau = T$ **is deterministic:** Thus, going forward we consider the case where $(X_t)_{0 \leq t \leq 1}$ is a Brownian bridge with $X_0 = X_1 = 0$. Clearly it suffices to consider $(X_t)_{0 \leq t \leq 1}$ to be one-dimensional in what follows, as in the multivariate case the coordinates of X are independent one-dimensional Brownian bridges. We first consider the case where $\tau = T$ is a constant time. In this case, the process $(Z_t)_{0 \leq t \leq 1}$ given by

$$Z_t := \begin{cases} X_t & \text{for } 0 \leq t < T, \\ X_{t-T}^{(T)} & \text{for } T \leq t \leq 1 \end{cases}$$

is clearly a Gaussian process on $[0,1]$ that is continuous on $[0,T)$ and $[T,1]$. To show the result, we must show (1) for any $s \in [0,T), t \in [T,1]$, $k(s,t) := \mathrm{Cov}(Z_s, Z_t) = 0$ (this implies $X^{(T)}$, and hence $Y^{(T)}$ is independent of $\mathcal{G}_T$), (2) $\mu(t) := \mathbb{E}Z_t = 0$ for all $t \in [T,1]$, and (3) $k(s,t) := \mathrm{Cov}(Z_s, Z_t) = (s-T) \wedge (t-T) - \frac{(s-T)(t-T)}{1-T}$ for all $s,t \in [T,1]$ (these final two points show the law of $X^{(T)}$ is that of a Brownian bridge since we already have sample path continuity).

We now check each of these properties. In what follows, recall that $X_t = B_t - tB_1$ for some (now one-dimensional) Brownian motion $(B_t)_{0\le t\le 1}$, and remember that $\mathrm{Cov}(B_s, B_t) = s \wedge t$.

1. For $s \in [0,T)$ and $t \in [0, 1-T]$, we have (assuming for now that $\mathbb{E}[X_t^{(T)}] = 0$, which we confirm in a later point)

$$\begin{aligned}\mathrm{Cov}(X_s, X_t^{(T)}) &= \mathbb{E}\left[X_s\left(X_{t+T} - \frac{1-T-t}{1-T}X_T\right)\right] \\ &= \mathbb{E}\left[X_s X_{t+T}\right] - \frac{1-T-t}{1-T}\mathbb{E}\left[X_s X_t\right] \quad = s(1-t-T) + \frac{1-T-t}{1-T}s(1-T) \\ &= 0,\end{aligned}$$

   which confirms the first point.

2. For any $t \in [0, 1-T]$, we have

$$\mathbb{E}\left[X_t^{(T)}\right] = \mathbb{E}\left[B_{t+T} - (t+T)B_1 - \frac{1-T-t}{1-T}B_T + \frac{1-T-t}{1-T}TB_1\right] = 0,$$

   proving the second point.

3. Lastly, using property 3 of Lemma 5.B.2, for $s,t \in [0,1-T]$ s.t. $s < t$, we have

$$\begin{aligned}\mathrm{Cov}\left(X_s^{(T)}, X_t^{(T)}\right) &= \mathbb{E}\left[\left(X_{s+T} - \frac{1-T-s}{1-T}X_T\right)\left(X_{t+T} - \frac{1-T-t}{1-T}X_T\right)\right] \\ &= \{(s+T) - (s+T)(t+T)\} - \frac{1-T-t}{1-T}\{T - (s+T)T\} \\ &\quad - \frac{1-T-s}{1-T}\{T - (t+T)T\} + \frac{(1-T-t)(1-T-s)}{(1-T)^2}\left\{T - T^2\right\} \\ &= \frac{1}{1-T}\Big[(s+T)(1-T-t)(1-T) - T(1-T-s)(1-T-t)\Big] \\ &= \frac{(1-T-t)s}{(1-T)} = s - \frac{st}{1-T}.\end{aligned}$$

Since we have shown that, for any $T \in [0,1]$, $Y^{(T)}$ is independent of $\mathcal{G}_T$, we have that, for any $E \in \mathcal{G}_t$ and any bounded, any fixed times $0 \le t_1 < t_2 < \cdots < t_p \le 1$, and continuous function $F : \mathbb{R}^{dp} \to \mathbb{R}_{\ge 0}$,

$$\mathbb{E}\mathbb{1}_E F(Y_{t_1}^{(T)}, \ldots, F_{t_p}^{(T)}) = \mathbb{P}(A)\mathbb{E}F(X_{t_1}, \ldots, X_{t_p}),$$

which is a fact we will use in the sequel.

Step 3: generalizing to general stopping times:

We now emulate a standard proof of the strong Markov property for Brownian motion to extend to the case where τ is a $(\mathcal{G}_t)_{0\le t\le 1}$ stopping time (in particular, the proof of Theorem 2.20 in Le Gall [104]).

It suffices to show that, for any $A \in \mathcal{G}_\tau$, $0 \le t_1 < t_2 < \cdots < t_p \le 1$, and $F : \mathbb{R}^{dp} \to \mathbb{R}_{\ge 0}$ continuous and bounded that

$$\mathbb{E}\mathbb{1}_A F(Y^{(\tau)}_{t_1}, \ldots, Y^{(\tau)}_{t_p}) = \mathbb{P}(A)\mathbb{E}F(X_{t_1}, \ldots, X_{t_p}).$$

As noted in Le Gall [104], this not only proves the independence of $(Y^{(\tau)}_t)$ and $\mathcal{G}_\tau$, but also demonstrates by taking $A = \Omega$ (where $(\Omega, \mathcal{F}, \mathbb{P})$ is the underlying probability space) that $(Y^{(\tau)}_t)$ and (X_t) have the same finite-dimensional distributions, and hence $(Y^{(\tau)}_t)$ is a standard d-dimensional Brownian bridge since sample paths are continuous.

For n a positive integer and $T \in \mathbb{R}$, define $T|_n := \min\{k2^{-n} : k \in \mathbb{Z},\ k2^{-n} \ge T\}$, i.e. $T|_n$ is the smallest real of the form $k2^{-n}$ that is greater than or equal to T. A straightforward expansion of $Y^{(\tau|_n)}_t$ yields that, for any $t \in [0,1]$, we have $Y^{(\tau|_n)}_t \xrightarrow[n\to\infty]{} Y^{(\tau)}_t$, and thus bounded convergence yields

$$\begin{aligned}
\mathbb{E}\mathbb{1}_A F(Y^{(\tau)}_{t_1}, \ldots, Y^{(\tau)}_{t_p}) &= \lim_{n\to\infty} \mathbb{E}\mathbb{1}_A F(Y^{(\tau|_n)}_{t_1}, \ldots, Y^{(\tau|_n)}_{t_p}) \\
&= \lim_{n\to\infty} \sum_{k=0}^{2^n} \mathbb{E}\mathbb{1}_A \mathbb{1}_{E^k_n} F(Y^{(\tau|_n)}_{t_1}, \ldots, Y^{(\tau|_n)}_{t_p}) \\
&= \lim_{n\to\infty} \sum_{k=1}^{2^n} \mathbb{E}\mathbb{1}_A \mathbb{1}_{E^k_n} F(Y^{(k2^{-n})}_{t_1}, \ldots, Y^{(k2^{-n})}_{t_p}) \\
&= \lim_{n\to\infty} \sum_{k=1}^{2^n} \mathbb{P}(A \cap E^k_n)\mathbb{E}F(Y^{(k2^{-n})}_{t_1}, \ldots, Y^{(k2^{-n})}_{t_p}) \\
&= \mathbb{P}(A)\mathbb{E}F(X_{t_1}, \ldots, X_{t_p}),
\end{aligned}$$

where $(X_t)_{0\le t\le 1}$ is a standard d-dimensional Brownian bridge, proving the desired result. In the above, $E^k_n := \{(k-1)2^{-n} < \tau \le k2^{-n}\}$, and we use the identity $\mathbb{1}_{E^k_n} F(Y^{(\tau|_n)}_{t_1}, \ldots, Y^{(\tau|_n)}_{t_p}) = \mathbb{1}_{E^k_n} F(Y^{(k2^{-n})}_{t_1}, \ldots, Y^{(k2^{-n})}_{t_p})$. The second to last inequality follows from applying the result where t is a deterministic time, noting that the event $A \cap E^k_n$ is $\mathcal{G}_{k2^{-n}}$-measurable.

Thus, we have shown the desired result. ■

Corollary 5.B.4. *Let $(X_t)_{0\le t\le 1}$ be a d-dimensional Brownian bridge with $X_0 = A$ and $X_1 = b$, where A is a random variable. Let $(\mathcal{G}_t)_{0\le t\le 1}$ be the corresponding natural filtration. Let τ be a $(\mathcal{G}_t)$ stopping time. Then, for any $G \in \mathcal{G}_\tau$, the conditional law of the process $(X_t)_{\tau\le t\le 1}$ given $\{\tau = T, X_\tau = x\} \cap G$ is that of a Brownian bridge on $[T,1]$ with initial value $X_T = x$ and terminal value $X_1 = b$, i.e.*

$$\mathbb{P}(X \in \cdot \mid \tau = T, X_\tau = x, G) = \mathbb{P}(S \in \ \cdot\ ),$$

where $(S_t)_{T\le t\le 1}$ is a Brownian bridge on $[T,1]$ with $S_T = x$ and $S_1 = b$.

Proof. Let $(Z_t)_{0\le t\le 1}$ be a Brownian bridge wth $Z_0 = Z_1 = 0$. Applying the tower rule for conditional expectation alongside Lemma 5.B.3 gives us, for all $E \in \mathcal{F}$,

$$\mathbb{P}\left(Y^{(\tau)} \in E \mid \tau, X_\tau, \mathbb{1}_G\right) = \mathbb{E}\left[\mathbb{P}\left(Y^{(\tau)} \in E \mid \mathcal{F}_\tau\right)\right] = \mathbb{P}(Z \in E).$$

Thus, with probability one over the joint distribution of $(X_\tau, \tau, \mathbb{1}_G)$, we have

$$\mathbb{P}\left(Y^{(\tau)} \in E \mid X_\tau = x, \tau = t, G\right) = \mathbb{P}(Z \in E).$$

With Lemma 5.B.2, we know that, since $Y^{(\tau)}$ is a Brownian bridge with $Y_0^{(\tau)} = Y_1^{(\tau)} = 0$ on this event, then, $\frac{1}{\sqrt{1-T}} Y^{(\tau)}_{t(1-T)} = X_{t+T} - \frac{1-T-t}{1-T}x + \frac{t}{1-T}b$ is a Brownian bridge with initial and terminal value 0 on $[0, 1-T]$. The remainder of the result follows by adding $\frac{1-T-t}{1-T}x - \frac{t}{1-T}b$, applying the first part of Lemma 5.B.2, and reindexing the process to be defined on $[T, 1]$. ■

Lemma 5.B.3 and Corollary 5.B.4 above show that the conditional distributions of Brownian bridges, even at stopping times, are very well-behaved — the conditional distributions are exactly that of another Brownian bridge. We aim to apply these results to our analysis of the privacy loss of the Brownian mechanism as follows. We will shortly that the distribution of the outputs of the Brownian mechanism, which can be viewed as a Brownian motion being run in reverse, can be equivalently viewed as a Brownian bridge with random (particularly, multivariate Gaussian) initial state and fixed terminating state. Coupling this with the above strong Markov property, we will show that even when an analyst picks arbitrarily complicated stopping functions, the privacy loss looks as if the inspection times were fixed in advance.

First, we show that, for a fixed number n of time functions, the privacy loss is exactly as outlined in the statement of Proposition 5.B.1.

Lemma 5.B.5. *Let $n \in \mathbb{N}$ be arbitrary, and let $T_1, \dots, T_n$ be decreasing (i.e. non-increasing) time functions. Let $p^\nu_{1:n}$ denote the joint density of $(B^\nu_{T_1}, \dots, B^\nu_{T_n})$, where $(B^\nu_t)_{t\ge 0}$ is a d-dimensional Brownian motion starting at $\nu \in \mathbb{R}^d$ and $T^\nu_m := T_m\left(B^\nu_{T_1}, \dots, B^\nu_{T_{m-1}}\right)$. Then, for any $y_1, \dots, y_n \in \mathbb{R}^d$, we have*[6]

$$p^\nu_{1:n}(y_1, \dots, y_n) \propto_\nu \exp\left(-\frac{\|y_n - \nu\|^2}{2T_n}\right) \prod_{m=2}^{n} \exp\left(\frac{-\|y_{m-1} - y_m\|^2}{2(T_{m-1} - T_m)}\right),$$

where $T_m = T_m(y_1, \dots, y_{m-1})$ for notational convenience and $\propto_\nu$ indicates that the constant of proportionality does not depend on ν.

Proof. We prove the result by induction on n, with the base case of $n = 1$ being trivial. Assume now the result holds for n. Recall that the first time function T_1 is simply a constant. Define the "backwards" process $(X^\nu_t)_{0\le t\le T_1}$ by $X^\nu_t := B^\nu_{T_1 - t}$, and let $(\mathcal{G}_t)_{0\le t\le T_1}$ be the corresponding natural filtration, i.e. $\mathcal{G}_t := \sigma(X^\nu_s : s \le t) = \sigma(B^\nu_{T_1 - s} : s \le t)$. Inspection yields that $(X^\nu_t)_{0\le t\le T_1}$ is a Brownian bridge with $X^\nu_0 \sim \mathcal{N}(\nu, T_1 I_d)$ and $X^\nu_{T_1} = \nu$.

[6]Since we may have $T_m = T_{m-1}$ for some m, we adopt the convention that when $y_m = y_{m-1}$, $\exp\left(\frac{-(y_m - y_{m-1})^2}{2(T_m - T_{m-1})}\right) = 1$. Likewise, when $y_m \ne y_{m-1}$ in this setting, we adopt $\exp\left(\frac{-(y_m - y_{m-1})^2}{2(T_m - T_{m-1})}\right) = 0$. After the proof of this lemma, only the former case will occur.

First, we note that the strong Markov property (in particular Corollary 5.B.4) yields that, for any $(\mathcal{G}_t)$ stopping times $\tau_1 \leq \cdots \leq \tau_n$, the law of of $(X_t^\nu)_{\tau_n \leq 1 \leq T_1}$ conditional on the event $\{X_{\tau_1}^\nu = y_1, \ldots, X_{\tau_n}^\nu = y_n\}$ is that of a Brownian bridge with initial point $X_{\tau_n}^\nu = y_n$ and terminal point $X_{T_1}^\nu = \nu$. Applying this in the case $\tau_m = T_1 - T_m$, this yields that the conditional law of $(X_t^\nu)_{T_1 - T_n \leq t \leq T_1}$ given $\{X_0^\nu = y_1, X_{T_1 - T_2}^\nu = y_2, \ldots, X_{T_1 - T_n}^\nu = y_n\}$ is a Brownian bridge with initial point $X_{T_1 - T_n}^\nu = y_n$ and terminal point $X_{T_1}^\nu = \nu$. But, this is equivalent to saying the conditional law of $(B_t^\nu)_{0 \leq t \leq T_n}$ given $\{B_{T_n}^\nu = y_n, \ldots, B_{T_1}^\nu = y_1\}$ is that of a Brownian bridge with initial value $B_0^\nu = \nu$ and terminal value $B_{T_n}^\nu = y_n$.

Next, note that, on the event $\{B_{T_n}^\nu = y_n, \ldots, B_{T_1}^\nu = y_1\}$, the time function $T_{n+1} = T_{n+1}(y_1, \ldots, y_n)$ is constant in value. Following the from the preceding paragraph, the conditional density $p_{1:n+1}^\nu(y_{n+1} \mid y_1, \ldots, y_n)$ is just that of a Brownian bridge with initial value $B_0^\nu = \nu$ and $B_{T_n}^\nu = y_n$ inspected at time T_{n+1}. That is, from using the covariance and mean expressions for a Brownian bridge along with the fact it is a Gaussian process, we have by Lemma 5.B.2

$$p_{1:n+1}^\nu(y_{n+1} \mid y_1, \ldots, y_n) \propto_\nu \exp\left(- \frac{\left\| y_{n+1} - \nu - \frac{T_{n+1}}{T_n}(y_n - \nu) \right\|^2}{2(T_n - T_{n+1})} \cdot \frac{T_{n+1}}{T_n} \right).$$

Thus, applying Bayes rule for densities alongside the inductive hypothesis, we have

$$\begin{aligned}
p_{1:n+1}^\nu(y_1, \ldots, y_{n+1}) &= p_{1:n}^\nu(y_1, \ldots, y_n) p_{1:n+1}^\nu(y_{n+1} \mid y_1, \ldots, y_n) \\
&\propto_\nu \exp\left(- \frac{\|y_n - \nu\|^2}{2T_n} \right) \cdot \left(\prod_{m=2}^{n} \exp\left(\frac{-\|y_{m-1} - y_m\|^2}{2(T_{m-1} - T_m)} \right) \right) \cdot \exp\left(- \frac{\left\| y_{n+1} - \nu - \frac{T_{n+1}}{T_n}(y_n - \nu) \right\|^2}{2(T_n - T_{n+1})} \cdot \frac{T_{n+1}}{T_n} \right) \\
&= \exp\left(- \frac{\|y_{n+1} - \nu\|^2}{2T_{n+1}} \right) \cdot \prod_{m=2}^{n+1} \exp\left(\frac{-\|y_{m-1} - y_m\|^2}{2(T_{m-1} - T_m)} \right),
\end{aligned}$$

which proves the desired claim. ■

With the above lemma, which shows that Proposition 5.B.1 holds when the number of time functions is constant, we can now prove that Proposition 5.B.1 holds in full generality. The idea behind the general proof is as follows. First, we consider the setting where an analyst has a sequence of time functions $T_1, T_2, \ldots$ and uses a stopping function N that satisfies $N((y_n)_{n \geq 1}) \leq n$ for all possible strings of inputs. We then construct a sequence of exactly n time functions $S_1, \ldots, S_n$ such that $p_{1:n}^\mu\left(B_{S_1^\nu}^\nu, \ldots, B_{S_n^\nu}^\nu\right) = p_{1:N}^\mu\left(B_{T_1^\nu}^\nu, \ldots, B_{T_{N^\nu}^\nu}^\nu\right)$. Then, in the general case where we only assume $N((y_n)_{n \geq 1}) < \infty$ for all sequences $(y_n)_{n \geq 1}$, for any $\delta > 0, \nu \in \mathbb{R}^d$, there is some n_δ^ν such that $\mathbb{P}\left(N\left(\left(B_{T_n^\nu}^\nu \right)_{n \geq 1} \right) \leq n_{\nu,\delta} \right) \geq 1 - \delta$, which will allow us to apply our argument from the setting where N is bounded alongside a limiting argument.

With the above brief description of our technique at hand, we now prove Proposition 5.B.1.

Proof of Proposition 5.B.1. By assumption, for all sequences $(y_m)_{m\geq 1}$ of elements of $\mathbb{R}^d$, we have $N\left((y_m)_{m\geq 1}\right) \leq n$ for some fixed natural number $n \in \mathbb{N}$. If $(T_m)_{m\geq 1}$ is the original sequence of stopping functions, define a new sequence by $S_m := T_{m\wedge N}$ for all $m \in [n]$.[7]

It is straightforward to see that, for any $\mu, \nu \in \mathbb{R}^d$,

$$p^{\mu}_{1:N}\left(B^{\nu}_{T^{\nu}_1}, \ldots, B^{\nu}_{T^{\nu}_{N^\nu}}\right) = p^{\mu}_{1:n}\left(B^{\nu}_{T_{1\wedge N^\nu}}, \ldots, B^{\nu}_{T^{\nu}_{n\wedge N^\nu}}\right) = p^{\mu}_{1:n}\left(B^{\nu}_{S^{\nu}_1}, \ldots, B^{\nu}_{S^{\nu}_n}\right).$$

Moreover, Lemma 5.B.5 yields that

$$\frac{p^{\nu}_{1:n}\left(B^{\nu}_{S^{\nu}_1}, \ldots, B^{\nu}_{S^{\nu}_n}\right)}{p^{\mu}_{1:n}\left(B^{\nu}_{S^{\nu}_1}, \ldots, B^{\nu}_{S^{\nu}_n}\right)} = \frac{\exp\left(-\frac{\left\|B^{\nu}_{S^{\nu}_n}-\nu\right\|^2}{2S^{\nu}_n}\right)}{\exp\left(-\frac{\left\|B^{\nu}_{S^{\nu}_n}-\mu\right\|^2}{2S^{\nu}_n}\right)} = \frac{\exp\left(-\frac{\left\|B^{\nu}_{T^{\nu}_{N^\nu}}-\nu\right\|^2}{2T^{\nu}_{N^\nu}}\right)}{\exp\left(-\frac{\left\|B^{\nu}_{T^{\nu}_{N^\nu}}-\mu\right\|^2}{2T^{\nu}_{N^\nu}}\right)},$$

which is just the ratio between the density of a $\mathcal{N}\left(\nu, T^{\nu}_{N^\nu}\right)$ random variable and a $\mathcal{N}\left(\mu, T^{\nu}_{N^\nu}\right)$ random variable evaluated at $B^{\nu}_{T^{\nu}_{N^\nu}}$, proving the desired result. ■

We now prove Theorem 5.3.4, which gives a closed form characterization of the Brownian mechanism. In what follows, we use the same notation for the density of Brownian motion as in the above proof.

Proof of Theorem 5.3.4. The second statement of the theorem is trivial and follows from our assumption of bounded ℓ_2 sensitivity. Hence, we only prove the first statement below.

From the results of Lemma 5.3.3, we have

$$\begin{aligned}\mathcal{L}^{\mathrm{BM}}_{1:N(x)}(x, x') &= \log\left(\frac{p^{f(x)}_{T_{N(x)}(x)}(\mathrm{BM}_n(x))}{p^{f(x')}_{T_{N(x)}(x)}(\mathrm{BM}_n(x))}\right) \\ &= -\frac{1}{2}\left[\frac{||B_{T_{N(x)}} - f(x)||_2^2}{T_{N(x)}(x)} - \frac{||B_{T_{N(x)}} - f(x')||_2^2}{T_{N(x)}(x)}\right]\end{aligned}$$

Without loss of generality, and for the sake of simplicity, $f(x) = 0$. The privacy loss can be written as

$$\begin{aligned}\mathcal{L}^{\mathrm{BM}}_{1:N(x)}(x, x') &= \frac{1}{2T_{N(x)}(x)}\left(-||B_{T_{N(x)}(x)}||_2^2 + ||B_{T_{N(x)}(x)} - f(x')||_2^2\right) \\ &= -\frac{1}{T_n(x)}\langle B_{T_{N(x)}(x)}, f(x')\rangle + \frac{1}{2T_{N(x)}(x)}||f(x')||_2^2 \\ &= -\frac{||f(x')||_2}{T_{N(x)}(x)}\left\langle B_{T_{N(x)}(x)}, \frac{f(x')}{||f(x')||_2}\right\rangle + \frac{1}{2T_{N(x)}(x)}||f(x')||_2^2\end{aligned}$$

[7]While N technically accepts an infinite sequence $(y_n)_{n\geq 1}$ of vectors as input, by definition, checking $N\left((y_n)\right) \leq m$ only requires examining the first m elements of the sequence $y_1, \ldots, y_m$.

$$= -\frac{||f(x')||_2}{T_{N(x)}(x)} \left\langle B_{T_{N(x)}(x)}, \frac{f(x')}{||f(x')||_2} \right\rangle + \frac{1}{2T_{N(x)}(x)} ||f(x')||_2^2$$
$$= -\frac{||f(x')||_2}{T_{N(x)}(x)} W_{T_n(x)} + \frac{1}{2T_{N(x)}(x)} ||f(x')||_2^2.$$

Note that the last inequality follows from the fact that if $(B_t)_{t\geq 0}$ is a d-dimensional Brownian motion and $z \in \mathbb{R}^d$ is a unit vector under the ℓ_2 norm, then the process $W_t := \langle z, B_t \rangle$ is a standard Brownian motion. Noting that $(-W_t)_{t\geq 0}$ is also a Brownian motion furnishes the result. ■

We now use the characterization of privacy loss in Theorem 5.3.4 alongside the time-uniform concentration results for continuous time martingales found in Appendix 5.A to construct two general families of privacy boundaries. We now prove Theorem 5.3.6.

Proof of Theorem 5.3.6. Recall from Theorem 5.3.4 that we have the following bound

$$\mathcal{L}^{\mathrm{BM}}_{1:N(x)}(x, x') \leq \frac{\Delta^2}{2T_{N(x)}(x)} + \frac{\Delta}{T_{N(x)}(x)} W^+_{T_{N(x)}(x)},$$

where $A^+ := \max(A, 0)$. First, by leveraging Lemma 5.A.2, we see that, with probability at least $1 - \delta$, we have

$$\mathcal{L}^{\mathrm{BM}}_{1:N(x)}(x, x') \leq \frac{\Delta^2}{2T_{N(x)}(x)} + \frac{\Delta}{T_{N(x)}(x)} \sqrt{2(T_{N(x)}(x) + \rho) \log\left(\frac{1}{\delta}\sqrt{\frac{T_{N(x)}(x) + \rho}{\rho}}\right)}$$
$$= \psi^M_\rho(T_{N(x)}(x)),$$

proving that ψ^M_ρ is a valid δ-privacy boundary. Likewise, by Lemma 5.A.1, we have that

$$\mathcal{L}^{\mathrm{BM}}_{1:N(x)}(x, x') \leq \frac{\Delta^2}{2T_{N(x)}(x)} + \frac{\Delta}{T_{N(x)}(x)}(aT_{N(x)}(x) + b) = \frac{\Delta}{T_{N(x)}(x)}\left(\frac{\Delta}{2} + b\right) + \Delta a$$
$$= \psi^L_{a,b}(T_{N(x)}(x)),$$

showing $\psi^L_{a,b}$ is a valid δ-privacy boundary.

■

5.C Proofs From Section 5.5

In this appendix, we provide proofs of the results in Section 5.5. We start by proving the privacy guarantees for $\mathrm{ReducedAboveThreshold}$.

Proof of Theorem 5.5.1. For $\mathrm{ReducedAboveThreshold}$ as described in Algorithm 1, on the event $\{N(x) = n\}$, all information leaked about the underlying private dataset is contained in $\mathrm{Alg}_{1:n}(x)$ and $\alpha_{1:n}(x)$, where $\alpha_n(x)$ is defined to be the nth bit output by $\mathrm{ReducedAboveThreshold}$. For any $y \in \mathcal{X}$, let $q^y_{1:n}$ denote the joint density of $(\mathrm{Alg}_{1:n}(y), \alpha_{1:n}(y))$, $p^y_{1:n}$ the marginal density of $\mathrm{Alg}_{1:n}(y)$, and $p^y_{1:n}(\cdot \mid \cdot)$ the conditional pmf of $\alpha_{1:n}(y)$ given the observed values of

$\mathrm{Alg}_{1:n}(y)$. As such, for any neighboring datasets $x \sim x'$, on the event $\{N(x) = n\}$, the privacy loss of $\mathrm{ReducedAboveThreshold}$, denoted by $\mathcal{L}^{\mathrm{RAT}}(x, x')$, is given by

$$\begin{aligned}
\mathcal{L}^{\mathrm{RAT}}_{1:n}(x, x') &= \log\left(\frac{q^x_{1:n}(\mathrm{Alg}_{1:n}(x), \alpha_{1:n}(x))}{q^{x'}_{1:n}(\mathrm{Alg}_{1:n}(x), \alpha_{1:n}(x))}\right) \\
&= \log\left(\frac{p^x_{1:n}(\mathrm{Alg}_{1:n}(x))}{p^{x'}_{1:n}(\mathrm{Alg}_{1:n}(x))}\right) + \log\left(\frac{p^x_{1:n}(\alpha_{1:n}(x) \mid \mathrm{Alg}_{1:n}(x))}{p^{x'}_{1:n}(\alpha_{1:n}(x) \mid \mathrm{Alg}_{1:n}(x))}\right) \\
&= \log\left(\frac{p^x_{1:n}(\mathrm{Alg}_{1:n}(x))}{p^{x'}_{1:n}(\mathrm{Alg}_{1:n}(x))}\right) + \log\left(\frac{p^x_{1:n}(0^{n-1}1 \mid \mathrm{Alg}_{1:n}(x))}{p^{x'}_{1:n}(0^{n-1}1 \mid \mathrm{Alg}_{1:n}(x))}\right) \\
&= \mathcal{L}^{\mathrm{Alg}}_{1:n}(x, x') + L_n(x, x'),
\end{aligned}$$

where $0^{n-1}1$ denotes the string of $n-1$ 0's followed by a single 1. In the last line we leverage the definition of the privacy loss between $\mathrm{Alg}_{1:n}(x)$ and $\mathrm{Alg}_{1:n}(x')$ and define

$$L_n(x, x') := \log\left(\frac{p^x_{1:n}(0^{n-1}1 \mid \mathrm{Alg}_{1:n}(x))}{p^{x'}_{1:n}(0^{n-1}1 \mid \mathrm{Alg}_{1:n}(x))}\right).$$

Now, to finish the result, it suffices to prove that, for any n, $L_n(x, x') \le \mathcal{E}_n(\mathrm{Alg}_{1:n-1}(x))$. Without loss of generality, we can assume all thresholds take the same value τ across rounds, as we can always define the shifted function $u'_n(\mathrm{Alg}_{1:n}(x), x) := u_n(\mathrm{Alg}_{1:n}(x), x) - \tau_n + \tau$. To prove our desired inequality, we proceed largely in the same way as the proof of $\mathrm{AboveThreshold}$ found in Lyu et al. [114], noting that conditioning on $\mathrm{Alg}_{1:n}(x)$ serves to fix the utility functions $u_1(\mathrm{Alg}_1(x), \cdot), \ldots, u_n(\mathrm{Alg}_{1:n}(x), \cdot)$ and the privacy levels

$$\mathcal{E}_1, \mathcal{E}_2(\mathrm{Alg}_1(x)), \ldots, \mathcal{E}_n(\mathrm{Alg}_{1:n-1}(x)).$$

For simplicity, going forward, we refer to the former quantities as $u_1(\cdot), \ldots, u_n(\cdot)$ and the latter quantities just as $\epsilon_1, \ldots, \epsilon_n$. The only remaining caveat that we must take care in handling variable amount of noise on the threshold introduced by LNR. Going forward, let $\mathbb{P}_{1:n}$ denote the conditional probability $\mathbb{P}(\cdot \mid \mathrm{Alg}_{1:n}(x))$. First, observe that we can write the numerator of $L_n(x, x')$ as

$$p^x\left(0^{n-1}1 \mid \mathrm{Alg}_{1:n}(x)\right) = \int_{\mathbb{R}^n} g^\tau_{1:n}(s_1, \ldots, s_n) \left(\prod_{i=1}^{n-1} \mathbb{P}_{1:n}\left(u_i(x) + \xi_i < s_i\right)\right) \mathbb{P}_{1:n}\left(u_n(x) + \xi_n \ge s_n\right) d\vec{s},$$

where $g^\tau_{1:n}$ represents the density for the joint distribution of $(\tau + Z(2\Delta/\epsilon_m))_{m=1}^n$, where $(Z(t))_{t \ge \eta}$ is as defined in Equation (5.4.2). We now need three inequalities. The first two are standard from the analysis of Lyu et al. [114], so we do not provide a proof. The third inequality is a product of our novel $\mathrm{ReducedAboveThreshold}$ mechanism, and hence we provide a proof. The inequalities are:

1. For $i < n$ and fixed s_i, $\mathbb{P}_{1:n}(u_i(x) + \xi_i < s_i) \le \mathbb{P}_{1:n}(u_i(x') + \xi_i < s_i + \Delta)$,
2. for $i = n$ and any s_n, $\mathbb{P}_{1:n}(u_n(x) + \xi_n \ge s_n) \le e^{\epsilon_n/2}\mathbb{P}_{1:n}(u_n(x') + \xi_n \ge s_n + \Delta)$, and
3. for any $s_{1:n} \in \mathbb{R}^n$, $g^\tau_{1:n}(s_1, \ldots, s_n) \le e^{\epsilon_n/2} g^\tau_{1:n}(s_1 + \Delta, \ldots, s_n + \Delta)$.

We now prove the third inequality. We have that

$$\frac{g^{\tau}_{1:n}(s_1,\ldots,s_n)}{g^{\tau-\Delta}_{1:n}(s_1,\ldots,s_n)} = \frac{g^{\tau}_{n}(s_n)g^{\tau}_{1:n-1}(s_1,\ldots,s_{n-1} \mid s_n)}{g^{\tau-\Delta}_{n}(s_n)g^{\tau-\Delta}_{1:n-1}(s_1,\ldots,s_{n-1} \mid s_n)} = \frac{g^{\tau}_{n}(s_n)}{g^{\tau-\Delta}_{n}(s_n)} \leq e^{\epsilon_n/2},$$

where the first equality follows from applying Bayes rule to the joint densities of the noisy thresholds, and the second equality follows from the fact that $(Z(t))$ forms a Markov process. This in particular implies that the density conditional density given the nth threshold satisfies $g^{a}_{1:n-1}(s_1,\ldots,s_{n-1} \mid s_n) = g^{b}_{1:n-1}(s_1,\ldots,s_{n-1} \mid s_n)$ for all $a, b \in \mathbb{R}$. The last inequality follows from examining the ratio of densities of $\mathrm{Lap}(\tau, 2\Delta/\epsilon_n)$ and $\mathrm{Lap}(\tau-\Delta, 2\Delta/\epsilon_n)$ random variables. Now, observe that by a simple shift of parameters we have

$$g^{\tau-\Delta}_{1:n}(s_1,\ldots,s_n) = g^{\tau}_{1:n}(s_1+\Delta,\ldots,s_n+\Delta).$$

Plugging this in, we have

$$\begin{aligned}
&p^x\left(0^{n-1}1 \mid \mathrm{Alg}_{1:n}(x)\right) \\
&= \int_{\mathbb{R}^n} g^{\tau}_{1:n}(s_1,\ldots,s_n)\left(\prod_{i=1}^{n-1}\mathbb{P}_{1:n}(u_i(x)+\xi_i < s_i)\right)\mathbb{P}_{1:n}(u_n(x)+\xi_n \geq s_n)d\vec{s} \\
&\leq e^{\epsilon_n/2}\int_{\mathbb{R}^n} g^{T-\Delta}_{1:n}(s_1,\ldots,s_n)\left(\prod_{i=1}^{n-1}\mathbb{P}_{1:n}(u_i(x)+\xi_i < s_i)\right)\mathbb{P}_{1:n}(u_n(x)+\xi_n \geq s_n)d\vec{s} \\
&\leq e^{\epsilon_n}\int_{\mathbb{R}^n} g^{\tau-\Delta}_{1:n}(s_1,\ldots,s_n)\left(\prod_{i=1}^{n-1}\mathbb{P}_{1:n}(u_i(x')+\xi_i < s_i+\Delta)\right)\mathbb{P}(u_n(x')+\xi_n \geq s_n+\Delta)d\vec{s} \\
&= e^{\epsilon_n}\int_{\mathbb{R}^n} g^{\tau}_{1:n}(s_1,\ldots,s_n)\left(\prod_{i=1}^{n-1}\mathbb{P}_{1:n}(u_i(x')+\xi_i < s_i)\right)\mathbb{P}_{1:n}(u_n(x')+\xi_n \geq s_n)d\vec{s} \\
&= e^{\epsilon_n}p^{x'}\left(0^{n-1}1 \mid \mathrm{Alg}_{1:n}(x)\right).
\end{aligned}$$

Rearranging furnishes the desired result. ■

We can also prove a corresponding utility guarantee for $\mathrm{ReducedAboveThreshold}$. As mentioned earlier, this utility guarantee is naive in the sense that it is derived from a union bound. Thus, instead of plotting the utility guarantee in our experiments in Section 5.6, we instead plot empirically observed loss/accuracy. Additionally, for the utility guarantee to hold, the sequence of privacy functions $(\mathcal{E}_n)_{n\geq 1}$ must be constant functions, i.e. $\mathcal{E}_n = \epsilon_n$ for each n. We now state the formal, high-probability utility guarantee in the following proposition.

Proposition 5.C.1. *Let $(p_n)_{n\geq 1}$ be a sequence of non-negative numbers such that $\sum_{i=1}^{\infty} p_i = 1$, and let $\gamma \in (0,1)$ be a confidence parameter. Define the sequence of parameters $(\eta_n)_{n\geq 1}$ by*

$$\eta_n := \frac{4\Delta}{\epsilon_n}\left(\log\left(\frac{2}{\gamma}\right) - \log(p_n)\right).$$

Then, if $N(x)$ is the time defined in Theorem 5.5.1, with probability at least $1-\gamma$, we have

$$u_{N(x)}(x) \geq \tau_{N(x)} - \eta_{N(x)}.$$

Proof. The above utility guarantee follows from applying two simple union bounds. First, we have

$$\mathbb{P}\left(\bigcup_{n\geq 1}\{|\xi_n| \geq \eta_n/2\}\right) \leq \sum_{n\geq 1}\mathbb{P}(|\xi_n| \geq \eta_n/2) = \sum_{n\geq 1}\exp\left(\frac{-\epsilon_n\eta_n}{4\Delta}\right) = \frac{\gamma}{2}\sum_{n\geq 1} p_n = 1.$$

Second, we have that

$$\mathbb{P}\left(\bigcup_{n\geq 1}\{|\zeta_n| \geq \eta_n/2\}\right) \leq \sum_{n\geq 1}\mathbb{P}(|\zeta_n| \geq \eta_n/2) = \sum_{n\geq 1}\exp\left(\frac{-\epsilon_n\eta_n}{2\Delta}\right) \leq \frac{\gamma}{2}\sum_{n\geq 1} p_n = 1.$$

Thus, with probability at least $1-\gamma$, we have simultaneously for all $n \geq 1$ that $|\xi_n| \leq \eta_n/2$ and $|\zeta_n| \leq \eta_n/2$. Thus, with the same probability, on round $N(x)$, we have

$$u_{N(x)}(x) \geq \tau_{N(x)} - \eta_{N(x)}.$$

■

5.D Proofs From Section 5.4

We first prove that the process defined in Equation (5.4.2) has Laplace marginal distributions.

Theorem 5.D.1. *Let $(Z_t)_{t\geq\eta}$ be the process defined in Equation* (5.4.2)*. Then, for any $t \geq \eta$, we have*

$$Z_t \sim \mathrm{Lap}(t).$$

In what follows, we sometimes use the notation $Z(t)$ interchangeably with Z_t for convenience.

Proof. Recall that if $X \sim \mathrm{Lap}(s)$, then X has characteristic function φ_s given by

$$\varphi_s(\lambda) = \frac{1}{1+\lambda^2 s^2}.$$

Let ϕ denote the characteristic function of $Z_t - Z_\eta$. Since Z_η and $Z_t - Z_\eta$ are independent, to show $Z_t \sim \mathrm{Lap}(t)$, it suffices to show that

$$\phi(\lambda) = \frac{\varphi_t(\lambda)}{\varphi_\eta(\lambda)} = \frac{1+\lambda^2\eta^2}{1+\lambda^2 t^2}.$$

Now, observe that the inhomogenous Poisson process $(P_t)_{t\geq\eta}$ can be written as $(\widetilde{P}(e^{t/2}))_{t\geq\log(\eta^2)}$ where $\widetilde{P}$ is a homogeneous Poisson process with rate $\lambda = 1$ on $[\log(\eta^2), \infty)$. In terms of the process $\widetilde{P}$, we can consider the process $(\widetilde{Z}_t)_{t\geq\log(\eta^2)}$ given by

$$\widetilde{Z}_t = \sum_{n\leq\widetilde{P}_t} \mathrm{Lap}\left(e^{\widetilde{\mathcal{T}}_n/2}\right),$$

where $\widetilde{\mathcal{T}}_n := \inf\{t \geq \log(\eta^2) : \widetilde{P}_t \geq n\}$ and $\widetilde{\mathcal{T}}_0 = \log(\eta^2)$. It is easy to see that

$$\widetilde{Z}(\log(t^2)) - \widetilde{Z}(\log(\eta^2)) =_d Z_t - Z_\eta.$$

Leveraging this identity, it follows that we have

$$\begin{aligned}
\phi(\lambda) &= \mathbb{E}\left[e^{i\lambda(Z_t - Z_\eta)}\right] = \mathbb{E}\left[e^{i\lambda(\widetilde{Z}(\log(t^2)) - \widetilde{Z}(\log(\eta^2)))}\right] \\
&= \sum_{n=0}^{\infty} \frac{\eta^2}{t^2} \frac{[\log(t^2/\eta^2)]^n}{n!} \int_{\log(\eta^2) \leq u_1 < u_2 < \cdots < u_n \leq \log(t^2)} f^{(n)}(u_1, \ldots, u_n) \prod_{i=1}^{n} \mathbb{E}\left[e^{i\lambda \mathrm{Lap}(e^{u_i/2})}\right] d\mathbf{u} \\
&= \frac{\eta^2}{t^2} \sum_{n=0}^{\infty} \int_{\log(\eta^2) \leq u_1 < u_2 < \cdots < u_n \leq \log(t^2)} \prod_{i=1}^{n} \frac{1}{1 + \lambda^2 e^{u_i}} d\mathbf{u}. \qquad \text{(5.D.1)}
\end{aligned}$$

In the above, $f^{(n)}(u_1, \ldots, u_n) := \frac{n!}{[\log(t^2/\eta^2)]^n}$ is the distribution of the order statistics $(U_{(1)}, \ldots, U_{(n)})$ of n i.i.d. random variables that are uniform on $[\log(\eta^2), \log(t^2)]$. Essentially, what we have done is *first* conditioned of the number of Poisson arrivals that occur in the interval $[\log(\eta^2), \log(t^2)]$. Then, on the event $\{N(t) = n\}$, we condition again on the location of the n arrivals, which we know to be uniformly distributed across the time interval. Once the arrival locations are known, we can compute the conditional characteristic function, which is the the product of characteristic functions as illustrated in the integral above.

Now, we show inductively that

$$\int_{\log(\eta^2) \leq u_1 < u_2 < \cdots < u_n \leq \log(t^2)} \prod_{i=1}^{n} \frac{1}{1 + \lambda^2 e^{u_i}} d\mathbf{u} = \frac{1}{n!} \left[\log\left(\frac{t^2}{\eta^2} \frac{1 + \lambda^2 \eta^2}{1 + \lambda^2 t^2}\right)\right]^n.$$

The base case of $n = 1$ is trivially true. Now, we have that

$$\begin{aligned}
&\int_{\log(\eta^2) \leq u_1 < u_2 < \cdots < u_n \leq \log(t^2)} \prod_{i=1}^{n} \frac{1}{1 + \lambda^2 e^{u_i}} d\mathbf{u} \\
&= \int_{u_1 = \log(\eta^2)}^{\log(t^2)} \frac{1}{1 + \lambda^2 e^{u_1}} \int_{u_1 < u_2 < \cdots < u_n} \prod_{i=2}^{n} \frac{1}{1 + \lambda^2 e^{u_i}} d\mathbf{u}_{-1} du_1 \\
&= \frac{1}{(n-1)!} \int_{u = \log(\eta^2)}^{\log(t^2)} \frac{1}{1 + \lambda^2 e^u} \left[\log\left(\frac{t^2}{e^u} \frac{1 + \lambda^2 e^u}{1 + \lambda^2 t^2}\right)\right]^{n-1} du \\
&= \frac{1}{n!} \int_{\log(\eta^2)}^{\log(t^2)} \frac{d}{du} \left[-\log\left(\frac{t^2}{e^u} \frac{1 + \lambda^2 e^u}{1 + \lambda^2 t^2}\right)\right]^n du \; = \frac{1}{n!} \left[\log\left(\frac{t^2}{\eta^2} \frac{1 + \lambda^2 \eta^2}{1 + \lambda^2 t^2}\right)\right]^n.
\end{aligned}$$

Leveraging this identity and picking up from the expression for $\phi(\lambda)$ in Equation (5.D.1), we have that

$$\begin{aligned}
\phi(\lambda) &= \frac{\eta^2}{t^2} \sum_{n=0}^{\infty} \frac{1}{n!} \left[\log\left(\frac{t^2}{\eta^2} \frac{1 + \lambda^2 \eta^2}{1 + \lambda^2 t^2}\right)\right]^n \\
&= \frac{\eta^2}{t^2} \exp\left(\log\left(\frac{t^2}{\eta^2} \frac{1 + \lambda^2 \eta^2}{1 + \lambda^2 t^2}\right)\right) \; = \; \frac{1 + \lambda^2 \eta^2}{1 + \lambda^2 t^2}.
\end{aligned}$$

This proves the desired result. ■

The above proof can also be leveraged to show that, for any finite fixed sequence of times $(t_n)_{n\in[K]}$, $(Z(t_1),\ldots,Z(t_K))$ has the same distribution as $(\zeta_1,\ldots,\zeta_K)$, where $(\zeta_n)_{n\in[K]}$ is the Laplace process associated with times $(t_n)_{n\in[K]}$ as outlined in Equation (5.4.1). This justifies that the process $(Z(t))_{t\geq\eta}$ is in fact a continuous time generalization of the aforementioned discrete time process.

5.E Additional Experimental Details

Parameter settings: We set the regularization parameter to be $\lambda = 0.05$ and note that the ℓ_2 and ℓ_1-sensitivity for the output perturbation of logistic regression are respectively $\frac{2}{n\lambda}$ and $\frac{2\sqrt{d}}{n\lambda}$. Likewise, for covariance perturbation in ridge regression, the ℓ_2-sensitivities for privately releasing X^TX and X^Ty are both 2.0, and the corresponding ℓ_1-sensitivities for releasing these quantities are $2.0d$ and $2.0\sqrt{d}$ respectively [109, 25]. We set the failure probability for BM to be $\delta = 10^{-6}$, and in each task map privacy parameters (ϵ_n) to times (t_n) using the linear privacy boundary $\psi^L_{a,b}$ optimized for tightness at $\epsilon = 0.3$.

Optimizing privacy boundaries: We provide a high level description of how one may set the parameters associated with the privacy boundaries discussed in Theorem 5.3.6. Let us consider the case of the mixture boundary ψ^M_ρ for illustrative purposes.

Suppose a data analyst desires that the final level of privacy loss obtained by interacting with the Brownian mechanism should be approximately ϵ. Then, intuitively, the analyst should want to add the variance of the Gaussian noise added to be as small as possible when the privacy boundary takes value ϵ. In mathematical notation, the analyst wants to find a parameter ρ^* satisfying

$$\rho^* = \arg\min_\rho (\psi^M_\rho)^{-1}(\epsilon),$$

where we note that the inverse function $(\psi^M_\rho)^{-1}$ exists as ψ^M_ρ is strictly increasing. While this inverse has no closed form in general, the parameter ρ^* can be efficiently computed using a few lines of code. A similar, even more straightforward computation can be conducted for the linear privacy boundary.

Simulating Noise Reduction Mechanisms: We briefly describe how a data analyst can produce samples from the Brownian mechanism and the Laplace noise reduction mechanism. First, since $T_1(x)$ is a constant, we have $\mathrm{BM}_1(x) \sim \mathcal{N}(f(x), T_1(x))$. Then, given $\mathrm{BM}_{1:m-1}(x)$, we have $\mathrm{BM}_m(x) \sim \mathcal{N}\left(f(x) + \frac{T_m(x)}{T_{m-1}(x)}(B_{T_{m-1}}(x) - f(x)), \frac{(T_{m-1}(x)-T_m(x))T_m(x)}{T_{m-1}(x)}\right)$. Since simulating the Brownian mechnaism only requires normal samples, it can be efficiently computed.

Second, to sample from LNR, one can first generate the the points of arrival of the inhomogeneous Poisson process $(P_t)_{t\geq\eta}$ up to time $T_1(x)$. Let $\mathcal{T}_1,\ldots,\mathcal{T}_N$ denote these arrival times, where we note that N, the number of arrivals up to time $T_1(x)$, is a random variable. Then, one can generate $Y_m \sim \mathrm{Lap}(\mathcal{T}_m)$ for $m \leq N$. From this information, the process $(Z_t)_{\eta\leq t\leq T_1(x)}$ can be readily computed, as in Equation (5.4.2).

Chapter 6

On the Sublinear Regret of GP-UCB

In the kernelized bandit problem, a learner aims to sequentially compute the optimum of a function lying in a reproducing kernel Hilbert space given only noisy evaluations at sequentially chosen points. In particular, the learner aims to minimize regret, which is a measure of the suboptimality of the choices made. Arguably the most popular algorithm is the Gaussian Process Upper Confidence Bound (GP-UCB) algorithm, which involves acting based on a simple linear estimator of the unknown function. Despite its popularity, existing analyses of GP-UCB give a suboptimal regret rate, which fails to be sublinear for many commonly used kernels such as the Matérn kernel. This has led to a longstanding open question: are existing regret analyses for GP-UCB tight, or can bounds be improved by using more sophisticated analytical techniques? In this work, we resolve this open question and show that GP-UCB enjoys nearly optimal regret. In particular, our results yield sublinear regret rates for the Matérn kernel, improving over the state-of-the-art analyses and partially resolving a COLT open problem posed by Vakili et al. Our improvements rely on a key technical contribution — regularizing kernel ridge estimators in proportion to the smoothness of the underlying kernel k. Applying this key idea together with a largely overlooked concentration result in separable Hilbert spaces (for which we provide an independent, simplified derivation), we are able to provide a tighter analysis of the GP-UCB algorithm.

6.1 Introduction

An essential problem in areas such as econometrics [60, 71], medicine [117, 119], optimal control [12, 7], and advertising [107] is to optimize an unknown function given *bandit feedback*, in which algorithms only get to observe the outcomes for the chosen actions. Due to the bandit feedback, there is a fundamental tradeoff between *exploiting* what has been observed about the local behavior of the function and *exploring* to learn more about the function's global behavior. There has been a long line of work on bandit learning that investigates this tradeoff across different settings, including multi-armed bandits [144, 102, 168], linear bandits [3, 147], and kernelized bandits [30, 139, 154].

In this work, we focus on the kernelized bandit framework, which can be viewed as an exten-

sion of the well-studied linear bandit setting to an infinite-dimensional reproducing kernel Hilbert space (or RKHS) $(H, \langle \cdot, \cdot \rangle_H)$. In this problem, there is some unknown function $f^* : \mathcal{X} \to \mathbb{R}$ of bounded norm in H, where $\mathcal{X} \subset \mathbb{R}^d$ is a bounded set. In each round $n \in [N]$, the learner uses previous observations to select an action $X_n \in \mathcal{X}$, and then observes feedback $Y_n := f^*(X_n) + \epsilon_n$, where ϵ_n is a zero-mean noise variable. The learner aims to minimize (with high probability) the regret at time T, which is defined as

$$R_T := \sum_{n=1}^{T} f^*(x^*) - f^*(X_n)$$

where $x^* := \arg\max_{x \in \mathcal{X}} f^*(x)$. The goal is to develop simple, efficient algorithms for the kernelized bandit problem that minimize regret R_T. We make the following standard assumption. We also make assumptions on the underlying kernel k, which we discuss in Section 6.2.

Assumption 4. *We assume that (a) there is some constant $D > 0$ known to the learner such that $\|f^*\|_H \leq D$ and (b) for every $n \geq 1$, ϵ_n is σ-subGaussian conditioned on $\sigma(Y_{1:n-1}, X_{1:n})$.*

Arguably the simplest algorithm for the kernelized bandit problem is GP-UCB (Gaussian process upper confidence bound) [150, 30]. GP-UCB works by maintaining a kernel ridge regression estimator of the unknown function f^* alongside a confidence ellipsoid, optimistically selecting in each round the action that provides the maximal payoff over all feasible functions. Not only is GP-UCB efficiently computable thanks to the kernel trick, but it also offers strong empirical guarantees [30]. The only seeming deficit of GP-UCB is its regret guarantee, as existing analyses only show that, with high probability, $R_T = \widetilde{O}(\gamma_T \sqrt{T})$, where γ_T is a kernel-dependent measure of complexity known as the maximum information gain [150, 35]. In contrast, more complicated, less computationally efficient algorithms such as SupKernelUCB [156, 135] have been shown to obtain regret bounds of $\widetilde{O}(\sqrt{\gamma_T T})$, improving over the analysis of GP-UCB by a multiplicative factor of $\sqrt{\gamma_T}$. This gap is stark as the bound $\widetilde{O}(\gamma_T \sqrt{T})$ fails, in general, to be sub-linear for the practically relevant Matérn kernel, whereas $\widetilde{O}(\sqrt{\gamma_T T})$ is sublinear for *any* kernel experiencing polynomial eigendecay [154].

This discrepancy has prompted the development of many variants of GP-UCB that, while less computationally efficient, offer better regret guarantees in some situations [80, 140, 141]. (See a detailed discussion of these algorithms along with other related work in Appendix 6.A.) However, the following question remains an open problem in online learning [155]: are existing analyses of vanilla GP-UCB tight, or can an improved analysis show GP-UCB enjoys sublinear regret?

6.1.1 Contributions

In this work, we show that GP-UCB obtains almost optimal, sublinear regret for any kernel experiencing polynomial eigendecay. This, in particular, implies that GP-UCB obtains sublinear regret for the commonly used Matérn family of kernels. We provide a brief roadmap of our paper below.

1. In Section 6.3, we provide background into self-normalized concentration in Hilbert spaces. In particular, in Theorem 6.3.1, we provide an independent, simplified derivation of a bound

due to Abbasi-Yadkori [2], which concerns to self-normalized concentration of certain process in separable Hilbert spaces. This bound has been largely overlooked in the kernel bandit literature, so we draw attention to it in hopes it can be leveraged in solving further kernel-based learning problems. As opposed to the existing bound of Chowdhury and Gopalan [30], which involves employing a complicated "double mixture" argument, the bound we present follows directly from applying the well-studied finite-dimensional method of mixtures alongside a simple truncation argument [40, 41, 44, 3]. These bounds are clean and show simple dependence on the regularization parameter.

2. In Section 6.4, we use leverage the self-normalized concentration detailed in Theorem 6.3.1 to provide an improved regret analysis for GP-UCB. By carefully choosing regularization parameters based on the smoothness of the underlying kernel, we demonstrate that GP-UCB enjoys sublinear regret of $\widetilde{O}\left(T^{\frac{3+\beta}{2+2\beta}}\right)$ for any kernel experiencing (C, β)-polynomial eigendecay. As a special case of this result, we obtain regret bounds of $\widetilde{O}\left(T^{\frac{\nu+2d}{2\nu+2d}}\right)$ for the commonly used Matérn kernel with smoothness ν in dimension d. Our new analysis improves over existing state-of-the-art analysis for GP-UCB, which fails to guarantee sublinear regret in general for the Matérn kernel family [30], and thus partially resolves an open problem posed by [155] on the suboptimality of GP-UCB.

In sum, our results show that GP-UCB, the go-to algorithm for the kernelized bandit problem, is nearly optimal, coming close to the algorithm-independent lower bounds of Scarlett et al. [135]. Our work thus can be seen as providing theoretical justification for the strong empirical performance of GP-UCB [150]. Perhaps the most important message of our work is the importance of careful regularization in online learning problems. While many existing bandit works treat the regularization parameter as a small, kernel-independent constant, we are able to obtain significant improvements by carefully selecting the regularization parameter. We hope our work will encourage others to pay close attention to the selection of regularization parameters in future works.

6.2 Background and Problem Statement

Notation. We briefly touch on basic definitions and notational conveniences that will be used throughout our work. If $a_1, \ldots, a_n \in \mathbb{R}$, we let $a_{1:n} := (a_1, \ldots, a_n)^\top$. Let $(H, \langle \cdot, \cdot \rangle_H)$ be a reproducing kernel Hilbert space associated with a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We refer to the identity operator on H as id_H. This is distinct from the identity mapping on $\mathbb{R}^d$, which we will refer to as I_d. For elements $f, g \in H$, we define their outer product as $fg^\top := f\langle g, \cdot \rangle_H$ and inner product as $f^\top g := \langle f, g \rangle_H$. For any $n \geq 1$ and sequence of points $x_1, \ldots, x_n \in \mathcal{X}$ (which will typically be understood from context), let $\Phi_n := (k(\cdot, x_1), \ldots, k(\cdot, x_n))^\top$. We can respectively define the Gram matrix $K_n : \mathbb{R}^n \to \mathbb{R}^n$ and covariance operator $V_n : H \to H$ as $K_n := (k(x_i, x_j))_{i,j \in [n]} = \Phi_n \Phi_n^\top$ and $V_n := \sum_{m=1}^n k(\cdot, x_m) k(\cdot, x_m)^\top = \Phi_n^\top \Phi_n$. These two operators essentially encode the same information about the observed data points, the former being easier to work with when actually performing computations (by use of the well known kernel trick) and latter being easier to algebraically manipulate.

Suppose $A : H \to H$ is a Hermitian operator of finite rank; enumerate its non-zero eigenvalues as $\lambda_1(A), \ldots, \lambda_k(A)$. We can define the Fredholm determinant of $I + A$ as $\det(I + A) := \prod_{m=1}^{k}(1 + \lambda_i(A))$ [103]. For any $n \geq 1, \rho > 0$, and $x_1, \ldots, x_n \in \mathcal{X}$, one can check via a straightforward computation that $\det(I_n + \rho^{-1}K_n) = \det(\mathrm{id}_H + \rho^{-1}V_n)$, where K_n and V_n are the Gram matrix and covariance operator defined above. We, again, will use these two quantities interchangeably in the sequel, but will typically prefer the latter in our proofs.

If $(H, \langle \cdot, \cdot \rangle_H)$ is a (now general) separable Hilbert space and $(\varphi_i)_{i \geq 1}$ is an orthonormal basis for H, for any $N \geq 1$ we can define the orthogonal projection operator $\pi_N : H \to \mathrm{span}\{\varphi_1, \ldots, \varphi_N\} \subset H$ by $\pi_N f := \sum_{i=1}^{N} \langle f, \varphi_i \rangle_H \varphi_i$. We can correspondingly the define the projection onto the remaining basis functions to be the map $\pi_N^{\perp} : H \to \mathrm{span}\{\varphi_1, \ldots, \varphi_N\}^{\perp}$ given by $\pi_N^{\perp} f := f - \pi_N f$. Lastly, if $A : H \to H$ is a symmetric, bounded linear operator, we let $\lambda_{\max}(A)$ denote the maximal eigenvalue of A, when such a value exists. In particular, $\lambda_{\max}(A)$ will exist whenever A has a finite rank, as will typically be the case considered in this paper.

Basics on RKHSs. Let $\mathcal{X} \subset \mathbb{R}^d$ be some domain. A *kernel* is a positive semidefinite map $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that is square-integrable, i.e. $\int_{\mathcal{X}} \int_{\mathcal{X}} |k(x, y)|^2 dx dy < \infty$. Any kernel k has an associated *reproducing kernel Hilbert space* or *RKHS* $(H, \langle \cdot, \cdot \rangle_H)$ containing the closed span of all partial kernel evaluations $k(\cdot, x), x \in \mathcal{X}$. In particular, the inner product $\langle \cdot, \cdot \rangle_H$ on H satisfies the reproducing relationship $f(x) = \langle f, k(\cdot, x) \rangle_H$ for all $x \in \mathcal{X}$.

A kernel k can be associated with a corresponding *Hilbert-Schmidt operator*, which is the Hermitian operator $T_k : L^2(\mathcal{X}) \to L^2(\mathcal{X})$ given by $(T_k f)(x) := \int_{\mathcal{X}} f(y) k(x, y) dy$ for any $x \in \mathcal{X}$. In short, T_k can be thought of as "smoothing out" or "mollifying" a function f according to the similarity metric induced by k. T_k plays a key role in kernelized learning through *Mercer's Theorem*, which gives an explicit representation for H in terms of the eigenvalues and eigenfunctions of T_k.

Fact 6.2.1 (Mercer's Theorem). *Let $(H, \langle \cdot, \cdot \rangle_H)$ be the RKHS associated with kernel k, and let $(\mu_i)_{i \geq 1}$ and $(\phi_i)_{i \geq 1}$ be the sequence of non-increasing eigenvalues and corresponding eigenfunctions for T_k. Let $(\varphi_i)_{i \geq 1}$ be the sequence of rescaled functions $\varphi_i := \sqrt{\mu_i}\phi_i$. Then,*

$$H = \left\{ \sum_{i=1}^{\infty} \theta_i \varphi_i : \sum_{i=1}^{\infty} \theta_i^2 < \infty \right\},$$

and $(\varphi_i)_{i \geq 1}$ forms an orthonormal basis for $(H, \langle \cdot, \cdot \rangle_H)$.

We make the following assumption throughout the remainder of our work, which is standard and comes from Vakili et al. [154].

Assumption 5 (Assumption on kernel k). *The kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfies (a) $|k(x, y)| \leq L$ for all $x, y \in \mathcal{X}$, for some constant $L > 0$ and (b) $|\phi_n(x)| \leq B$ for all $x \in \mathcal{X}$, for some $B > 0$.*

"Complexity" of RKHS's. By the eigendecay of a kernel k, we really mean the rate of decay of the sequence of eigenvalues $(\mu_i)_{i \geq 1}$. In the literature, there are two common paradigms for studying the eigendecay of k: (C_1, C_2, β)-exponential eigendecay, under which $\forall i \geq 1, \mu_i \leq$

$C_1 \exp\left(-C_2 i^\beta\right)$, and (C, β)-polynomial eigendecay, under which $\forall i \geq 1, \mu_i \leq C i^{-\beta}$. For kernels experiencing exponential eigendecay, of which the squared exponential is the most important example, GP-UCB is known to be optimal up to poly-logarithmic factors. However, for kernels experiencing polynomial eigendecay, of which the Matérn family is a common example, existing analyses of GP-UCB fail to yield sublinear regret. It is this latter case we focus on in this work.

Given the above representation in Fact 6.2.1, it is clear that the eigendecay of the kernel k governs the "complexity" or "size" of the RKHS H. We make this notion of complexity precise by discussing *maximum information gain*, a sequential, kernel-dependent quantity governing concentration and hardness of learning in RKHS's [35, 150, 154].

Let $n \geq 1$ and $\rho > 0$ be arbitrary. The maximum information gain at time n with regularization ρ is the scalar $\gamma_n(\rho)$ given by

$$\gamma_n(\rho) := \sup_{x_1,\ldots,x_n \in \mathcal{X}} \frac{1}{2} \log\det\left(\mathrm{id}_H + \rho^{-1} V_n\right) = \sup_{x_1,\ldots,x_n \in \mathcal{X}} \frac{1}{2} \log\det\left(I_n + \rho^{-1} K_n\right).$$

Our presentation of maximum information gain differs from some previous works in that we encode the regularization parameter ρ into our notation. This inclusion is key for our results, as we obtain improvements by carefully selecting ρ. Vakili et al. [154] bound the rate of growth of $\gamma_n(\rho)$ in terms of the rate of eigendecay of the kernel k. We leverage the following fact in our main results.

Fact 6.2.2 (Corollary 1 in Vakili et al. [154]). *Suppose that kernel k satisfies Assumption 5 and experiences (C, β)-polynomial eigendecay. Then, for any $n \geq 1$, we have*

$$\gamma_n(\rho) \leq \left(\left(\frac{CB^2 n}{\rho} \right)^{1/\beta} \log^{-1/\beta}\left(1 + \frac{Ln}{\rho}\right) + 1 \right) \log\left(1 + \frac{Ln}{\rho}\right).$$

We last define the practically relevant Matérn kernel and discuss its eigendecay.

Definition/Fact 6.2.3. *The Matérn kernel with bandwidth $\sigma > 0$ and smoothness $\nu > 1/2$ is given by*

$$k_{\nu,\sigma}(x, y) := \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left(\frac{\sqrt{2\nu}\|x - y\|_2}{\sigma} \right)^\nu B_\nu \left(\frac{\sqrt{2\nu}\|x - y\|_2}{\sigma} \right),$$

where Γ is the gamma function and B_ν is the modified Bessel function of the second kind. It is known that there is some constant $C > 0$ that may depend on σ but not on d or ν such that $k_{\nu,\sigma}$ experiences $\left(C, \frac{2\nu+d}{d}\right)$-eigendecay [134, 154].

Basics on martingale concentration: If $\mathcal{F}$ is a σ-algebra, and ϵ is an $\mathbb{R}$-valued random variable, we say ϵ is σ-subGaussian conditioned on $\mathcal{F}$ if, for any $\lambda \in \mathbb{R}$, we have $\log \mathbb{E}\left(e^{\lambda \epsilon} \mid \mathcal{F}\right) \leq \frac{\lambda^2 \sigma^2}{2}$; in particular this condition implies that ϵ is mean zero. With this, we state the following result on self-normalized processes. To our understanding, the following result was first presented in some form as Example 4.2 of de la Peña et al. [41] (in the setting of continuous local martingales), and can be derived leveraging the argument of Theorem 1 in de la Peña et al. [44]. The exact form below was established (in the setting of discrete-time processes) in Theorem 1 of Abbasi-Yadkori et al. [3], which is commonly leveraged to construct confidence ellipsoids in the linear bandit setting.

Fact 6.2.4 (Example 4.2 from [41], Theorem 1 from [3]). *Let $(\mathcal{F}_n)_{n\geq 0}$ be a filtration, let $(X_n)_{n\geq 1}$ be an $(\mathcal{F}_n)_{n\geq 0}$-predictable sequence in $\mathbb{R}^d$, and let $(\epsilon_n)_{n\geq 1}$ be a real-valued $(\mathcal{F}_n)_{n\geq 1}$-adapted sequence such that conditional on $\mathcal{F}_{n-1}$, ϵ_n is mean zero and σ-subGaussian. Then, for any $\rho > 0$, the process $(M_n)_{n\geq 0}$ given by*

$$M_n := \frac{1}{\sqrt{\det(I_d + \rho^{-1}V_n)}} \exp\left\{\frac{1}{2}\left\|(\rho I_d + V_n)^{-1/2} S_n/\sigma\right\|_2^2\right\}$$

is a non-negative supermartingale with respect to $(\mathcal{F}_n)_{n\geq 0}$, where $S_n := \sum_{m=1}^n \epsilon_m X_m$ and $V_n := \sum_{m=1}^n X_m X_m^\top$. Consequently, by Theorem 1.0.2, for any confidence $\delta \in (0,1)$, the following holds: with probability at least $1-\delta$, simultaneously for all $n \geq 1$, we have

$$\left\|(V_n + \rho I_d)^{-1/2} S_n\right\|_2 \leq \sigma\sqrt{2\log\left(\frac{1}{\delta}\sqrt{\det(I_d + \rho^{-1}V_n)}\right)}.$$

Note the simple dependence on the regularization parameter $\rho > 0$ in the above bound. While the regularization parameter ρ doesn't prove important in regret analysis for linear bandits (where ρ is treated as constant), the choice for ρ will be critical in our setting. In the following section, we will discuss how Fact 6.2.4 can be extended to the setting of separable Hilbert spaces essentially verbatim (an observation first noticed by Abbasi-Yadkori [2]).

6.3 A Remark on Self-Normalized Concentration in Hillbert Spaces

We begin by discussing a key, self-normalized concentration inequality for martingales. We use this bound in the sequel to construct simpler, more flexible confidence ellipsoids than currently exist for GP-UCB. The bound we present (in Theorem 6.3.1 below) is, more or less, equivalent to Corollary 3.5 in the thesis of Abbasi-Yadkori [2]. Our result is mildly more general in the sense that it directly argues that a target mixture process is a nonnegative supermartingale. The result in Abbasi-Yadkori [2] is more general in the sense it allows the regularization (or shift) matrix to be non-diagonal. Either concentration result is sufficient for the regret bounds obtained in the sequel.

The aforementioned corollary in [2], quite surprisingly, has not been referenced in central works on the kernelized bandit problem, namely Chowdhury and Gopalan [30] and Vakili et al. [154, 155]. In fact, strictly weaker versions of the conclusion have been independently rediscovered in the context of kernel regression [48]. We emphasize that this result of Abbasi-Yadkori [2] (and the surrounding technical conclusions) are very general and may allow for further improvements in problems related to kernelized learning.

We now present Theorem 6.3.1, providing a brief sketch and a full proof in Appendix 6.B. We believe our proof, which directly shows a target process is a nonnegative supermartingale, is of independent interest when compared to that of Abbasi-Yadkori [2] due to its simplicity. In particular, our proof follows from first principles, avoiding advanced topological notions of convergence (e.g. in the weak operator topology) and existence of certain Gaussian measures on

separable Hilbert spaces, which were heavily utilized in the proof of Corollary 3.5 in Abbasi-Yadkori [2].

Theorem 6.3.1 (Self-normalized concentration in Hilbert spaces). *Let $(\mathcal{F}_n)_{n\geq 0}$ be a filtration, $(f_n)_{n\geq 1}$ be an $(\mathcal{F}_n)_{n\geq 0}$-predictable sequence in a separable Hilbert space[1] H such that $\|f_n\|_H < \infty$ a.s. for all $n \geq 0$, and $(\epsilon_n)_{n\geq 1}$ be an $(\mathcal{F}_n)_{n\geq 1}$-adapted sequence in $\mathbb{R}$ such that conditioned on $\mathcal{F}_{n-1}$, ϵ_n is mean zero and σ-subGaussian. Defining $S_n := \sum_{m=1}^n \epsilon_m f_m$ and $V_n := \sum_{m=1}^n f_m f_m^\top$, we have that for any $\rho > 0$, the process $(M_n)_{n\geq 0}$ defined by*

$$M_n := \frac{1}{\sqrt{\det(\mathrm{id}_H + \rho^{-1}V_n)}} \exp\left\{\frac{1}{2}\left\|(\rho \mathrm{id}_H + V_n)^{-1/2} S_n/\sigma\right\|_H^2\right\}$$

*is a nonnegative supermartingale with respect to $(\mathcal{F}_n)_{n\geq 0}$. Consequently, by Fact **??**, for any $\delta \in (0,1)$, with probability at least $1-\delta$, simultaneously for all $n \geq 1$, we have*

$$\left\|(V_n + \rho I_d)^{-1/2} S_n\right\|_H \leq \sigma\sqrt{2\log\left(\frac{1}{\delta}\sqrt{\det(\mathrm{id}_H + \rho^{-1}V_n)}\right)}.$$

We can summarize our independent proof in two simple steps. First, following from Fact 6.2.4, the bound in Theorem 6.3.1 holds when we project S_n and V_n onto a finite number N of coordinates, defining a "truncated" nonnegative supermartingale $M_n^{(N)}$. Secondly, we can make a limiting arugment, showing $M_n^{(N)}$ is "essentially" M_n for large values of N.

Proof Sketch for Theorem 6.3.1. Let $(\varphi_n)_{n\geq 1}$ be an orthonormal basis for H, and, for any $N \geq 1$, let π_N denote the projection operator onto $H_N := \mathrm{span}\{\varphi_1, \ldots, \varphi_N\}$. Note that the projected process $(\pi_N S_n)_{n\geq 1}$ is an H-valued martingale with respect to $(\mathcal{F}_n)_{n\geq 0}$. Further, note that the projected variance process $(\pi_N V_n \pi_N^\top)_{n\geq 0}$ satisfies

$$\pi_N V_n \pi_N^\top = \sum_{m=1}^n (\pi_N f_m)(\pi_N f_m)^\top.$$

Since, for any $N \geq 1$, H_N is a finite-dimensional Hilbert space, it follows from Lemma 6.B.1 that the process $(M_n^{(N)})_{n\geq 0}$ given by

$$M_n^{(N)} := \frac{1}{\sqrt{\det\big(\mathrm{id}_H + \rho^{-1}\pi_N V_n \pi_N^\top\big)}} \exp\left\{\frac{1}{2}\left\|(\rho \mathrm{id}_H + \pi_N V_n \pi_N^\top)^{-1/2} \pi_N S_n\right\|_H^2\right\},$$

is a nonnegative supermartingale with respect to $(\mathcal{F}_n)_{n\geq 0}$. One can check that, for any $n \geq 0$, $M_n^{(N)} \xrightarrow[N\to\infty]{} M_n$. Thus, Fatou's Lemma implies

$$\begin{aligned}\mathbb{E}\left(M_n \mid \mathcal{F}_{n-1}\right) &= \mathbb{E}\left(\liminf_{N\to\infty} M_n^{(N)} \mid \mathcal{F}_{n-1}\right) \\ &\leq \liminf_{N\to\infty} \mathbb{E}\left(M_n^{(N)} \mid \mathcal{F}_{n-1}\right)\end{aligned}$$

[1] A space is separable if it has a countable, dense set. Separability is key, because it means we have a countable basis, whose first N elements we project onto.

$$\leq \liminf_{N\to\infty} M_{n-1}^{(N)}$$
$$= M_{n-1},$$

which proves the first part of the claim. The second part of the claim follows from applying Ville's inequality (Theorem 1.0.2) to the defined nonnegative supermartingale and rearranging. See Appendix 6.B for details. ■

The following corollary specializes Theorem 6.3.1 (and thus Corollary 3.5 of Abbasi-Yadkori [2]) to the case where H is a RKHS and $f_n = k(\cdot, X_n)$, for all $n \geq 1$. In this special case, we can reframe the above theorem in terms familiar Gram matrix K_n, assuming the quantity is invertible. While we prefer the simplicity and elegance of working directly in the RKHS H in the sequel, the follow corollary allows us to present Theorem 6.3.1 in a way that is computationally tractable.

Corollary 6.3.2. *Let us assume the same setup as Theorem 6.3.1, and additionally assume that (a) $(H, \langle\cdot,\cdot\rangle_H)$ is a RKHS associated with some kernel k, and (b) there is some $\mathcal{X}$-valued $(\mathcal{F}_n)_{n\geq 0}$-predictable process $(X_n)_{n\geq 1}$ such that $(f_n)_{n\geq 1} = (k(\cdot, X_n))_{n\geq 1}$. Then, for any $\rho > 0$ and $\delta \in (0,1)$, we have that, with probability at least $1-\delta$, simultaneously for all $n \geq 0$,*

$$\left\|(V_n + \rho \mathrm{id}_H)^{-1/2} S_n\right\|_H \leq \sigma \sqrt{2\log\left(\sqrt{\frac{1}{\delta}\det(I_n + \rho^{-1}K_n)}\right)}.$$

If, in addition, the Gram matrix $K_n = (k(X_i, X_j))_{i,j\in[n]}$ is invertible, we have the equality

$$\|(I_n + \rho K_n^{-1})^{-1/2}\epsilon_{1:n}\|_2 = \|(\rho \mathrm{id}_H + V_n)^{-1/2} S_n\|_H.$$

We prove Corollary 6.3.2 in Appendix 6.B. With this reframing of Theorem 6.3.1, we compare the concentration results of Theorem 6.3.1 (and thus Abbasi-Yadkori [2]) to the following, commonly leveraged result from Chowdhury and Gopalan [30].

Fact 6.3.3 (**Theorem 1 from Chowdhury and Gopalan [30]**). *Assume the same setup as Fact 6.2.4. Let $\eta > 0$ be arbitrary, and let $K_n := (k(X_i, X_j))_{i,j\in[n]}$ be the Gram matrix corresponding to observations made by time $n \geq 1$. Then, with probability at least $1-\delta$, simultaneously for all $n \geq 1$, we have*

$$\left\|\left((K_n + \eta I_n)^{-1} + I_n\right)^{-1/2}\epsilon_{1:n}\right\|_2 \leq \sigma\sqrt{2\log\left(\frac{1}{\delta}\sqrt{\det\left((1+\eta)I_n + K_n\right)}\right)}.$$

To make comparison with this bound clear, we parameterize the bounds in the above fact in terms of $\eta > 0$ instead of $\rho > 0$ to emphasize the following difference: both sides of the bound presented in Theorem 6.3.1 shrink as ρ is increased, whereas both sides of the bound in Fact 6.3.3 increase as η grows. Thus, increasing ρ in Theorem 6.3.1 should be seen as decreasing η in the bound of Chowdhury and Gopalan [30]. The bounds in Corollary 6.3.2 and Fact 6.3.3 coincide when $\rho = 1$ and $\eta \downarrow 0$ (per Lemma 1 in Chowdhury and Gopalan [30]), but are otherwise not equivalent for other choices of ρ and η.

We believe Theorem 6.3.1 and Corollary 3.5 of Abbasi-Yadkori [2] to be signficantly more usable than the result of Chowdhury and Gopalan [30] for several reasons. First, the aforementioned bounds *directly* extend the method of mixtures (in particular, Fact 6.2.4) to potentially infinite-dimensional Hilbert spaces. This similarity in form allows us to leverage existing analysis of Abbasi-Yadkori et al. [3] to prove our regret bounds, with only slight modifications. This is in contrast to the more cumbersome regret analysis that leverages Fact 6.3.3, which is not only more difficult to follow, but also obtains inferior, sometimes super-linear regret guarantees.

Second, we note that Theorem 6.3.1 provides a bound that has a simple dependence on $\rho > 0$. In more detail, directly as a byproduct of the simplified bounds, Theorem 6.4.1 offers a regret bound that can readily be tuned in terms of ρ. Due to their use of a "double mixture" technique in proving Fact 6.3.3, Chowdhury and Gopalan [30] essentially wind up with a nested, doubly-regularized matrix $((K_n + \eta I_n)^{-1} + I_n)^{-1/2}$ with which they normalize the residuals $\epsilon_{1:n}$. In particular, this more complicated normalization make it difficult to understand how varying η impacts regret guarantees, which we find to be essential for proving improved regret guarantees.

We note that the central bound discussed in this section *does not* provide an improvement in dependence on maximum information gain in the sense hypothesized by Vakili et al. [155]. In particular, the authors hypothesized the possibility of shaving a $\sqrt{\gamma_n}$ multiplicative factor off of self-normalized concentration inequalities in RKHS's. This was shown in a recent work (see Lattimore [101]) to be impossible in general. Instead, Theorem 6.3.1 and Corollary 3.5 of Abbasi-Yadkori [2] give one access to a family of bounds parameterized by the regularization parameter $\rho > 0$. As will be seen in the sequel, by optimizing over this parameter, one can obtain significant improvements in regret.

6.4 An Improved Regret Analysis of GP-UCB

In this section, we provide the second of our main contributions, which is an improved regret analysis for the GP-UCB algorithm. We provide a description of GP-UCB in Algorithm 2. While we state the algorithm directly in terms of quantities in the RKHS H, these quantities can be readily converted to those involving Gram matrices or Gaussian processes for those who prefer that perspective [30, 169].

As seen in Section 6.3, by carefully extending the "method of mixtures" technique (originally by Robbins) of Abbasi-Yadkori et al. [3], Abbasi-Yadkori [2] and de la Peña et al. [40, 41] to Hilbert spaces, we can construct self-normalized concentration inequalities that have simple dependence on the regularization parameter ρ. These simplified bounds, in conjunction with information about the eigendecay of the kernel k [154], can be combined to carefully choose ρ to obtain improved regret. We now present our main result.

Theorem 6.4.1. *Let $T > 0$ be a fixed time horizon, $\rho > 0$ a regularization parameter, and assume Assumptions 5 and 4 hold. Let $\delta \in (0, 1)$, and for $n \geq 1$ define*

$$U_n := \sigma \sqrt{2 \log \left(\frac{1}{\delta} \sqrt{\det(\mathrm{id}_H + \rho^{-1} V_n)} \right)} + \rho^{1/2} D.$$

Then, with probability at least $1 - \delta$, the regret of Algorithm 2 run with parameters $\rho, (U_n)_{n \geq 1}, D$

Algorithm 2 Gaussian Process Upper Confidence Bound (GP-UCB)

Input: Regularization parameter $\rho > 0$, norm bound D, confidence bounds $(U_n)_{n\geq 1}$, and time horizon T.
Set $V_0 := 0$, $f_0 := 0$, $\mathcal{E}_0 := \{f \in H : \|f\|_H \leq D\}$
for $n = 1, \ldots, T$ **do**
Let $(X_n, \widetilde{f}_n) := \arg\max_{x\in\mathcal{X}, f\in\mathcal{E}_{n-1}} \langle f, k(\cdot, x)\rangle_H$
Play action X_n and observe reward $Y_n := f^*(X_n) + \epsilon_n$
Set $V_n := V_{n-1} + k(\cdot, X_n)k(\cdot, X_n)^\top$ and $f_n := (V_n + \rho \mathrm{id}_H)^{-1}\Phi_n^\top Y_{1:n}$
Set $\mathcal{E}_n := \left\{f \in H : \left\|(V_n + \rho\mathrm{id}_H)^{1/2}(f_n - f)\right\|_H \leq U_n\right\}$

satisfies

$$R_T = O\left(\gamma_n(\rho)\sqrt{T} + \sqrt{\rho\gamma_n(\rho)T}\right),$$

where in the big-Oh notation above we treat δ, D, σ, B, and L as being held constant. If the kernel k experiences (C, β)-polynomial eigendecay for some $C > 0$ and $\beta > 1$, taking $\rho = O(T^{\frac{1}{1+\beta}})$ yields $R_n = \widetilde{O}\left(T^{\frac{3+\beta}{2+2\beta}}\right)$[2], which is always sub-linear in T.

While we present the above bound with a fixed time-horizon, it can be made anytime by carefully applying a standard doubling argument (see Lattimore and Szepesvári [102], for instance). We specialize the above theorem to the case of the Matérn kernel in the following corollary.

Corollary 6.4.2. *Definition 6.2.3 states that the Matérn kernel with smoothness $\nu > 1/2$ in dimension d experiences $(C, \frac{2\nu+d}{d})$-eigendecay, for some constnat $C > 0$. Thus, GP-UCB obtains a regret rate of $R_n = \widetilde{O}\left(T^{\frac{\nu+2d}{2\nu+2d}}\right)$.*

We note that our regret analysis is the first to show that GP-UCB attains sublinear regret for general kernels experiencing polynomial eigendecay. Of particular import is that Corollary 6.4.2 of Theorem 6.4.1 yields the first analysis of GP-UCB that implies sublinear regret for the Matérn kernel under general settings of ambient dimension d and smoothness ν. A recent result by Janz [79], using a uniform lengthscale argument, demonstrates that GP-UCB obtains sublinear regret for the specific case of the Matérn family when the parameter ν and dimension d satisfy a uniform boundedness condition independent of scale. Our results are (a) more general, holding for *any* kernel exhibiting polynomial eigendecay, (b) don't require checking uniform boundedness independent of scale condition, and (c) follow from a simple regularization based argument. In particular, the arguments of Janz [79] require advanced functional analytic and Fourier analytic machinery.

We note that our analysis does not obtain optimal regret, as the theoretically interesting but computationally cumbersome SupKernelUCB algorithm [135, 156] obtains a slightly improved regret bound of $\widetilde{O}\left(T^{\frac{\beta+1}{2\beta}}\right)$ for (C, β)-polynomial eigendecay and $\widetilde{O}\left(T^{\frac{\nu+d}{2\nu+d}}\right)$ for the Matérn kernel with smoothness ν in dimension d. Due to the aforementioned result of Lattimore [101], which shows that improved dependence on maximum information gain cannot be generally obtained in Hilbert space concentration, we believe further improvements on regret analysis for

[2]The notation $\widetilde{O}$ suppresses multiplicative, poly-logarithmic factors in T

GP-UCB may not possible.

To wrap up this section, we provide a proof sketch for Theorem 6.4.1. The entire proof, along with full statements and proofs of the technical lemmas, can be found in Appendix 6.C.

Proof Sketch for Theorem 6.4.1. Letting, for any $n \in [T]$, the "instantaneous regret" be defined as $r_n := f^*(x^*) - f^*(X_n)$, a standard argument yields that, with probability at least $1-\delta$, simultaneously for all $n \in [T]$,

$$r_n \leq 2U_{n-1} \left\| (\rho \mathrm{id}_H + V_{n-1})^{-1/2} k(\cdot, X_n) \right\|_H .$$

A further standard argument using Cauchy-Schwarz and an elliptical potential argument yields

$$\begin{aligned} R_n &= \sum_{n=1}^{T} r_n \leq U_T \sqrt{2T \log\det(\mathrm{id}_H + \rho^{-1} V_T)} \\ &= \left(\sigma \sqrt{2 \log\left(\frac{1}{\delta} \sqrt{\det(\mathrm{id}_H + \rho^{-1} V_T)} \right)} + \rho^{1/2} D \right) \sqrt{2T \log\det(\mathrm{id}_H + \rho^{-1} V_T)} \\ &\leq \left(\sigma \sqrt{2\log(1/\delta)} + \sigma \sqrt{2\gamma_T(\rho)} + \rho^{1/2} D \right) \sqrt{4T\gamma_T(\rho)} = O\left(\gamma_T(\rho)\sqrt{T} + \sqrt{\rho \gamma_T(\rho) T} \right), \end{aligned}$$

which proves the first part of the claim. If, additionally, k experiences (C, β)-polynomial eigendecay, we know that $\gamma_T(\rho) = \widetilde{O}\left(\left(\frac{T}{\rho} \right)^{1/\beta} \right)$ by Fact 6.2.2. Setting $\rho := O(T^{\frac{1}{1+\beta}})$ thus yields

$$R_T = O\left(\gamma_T(\rho)\sqrt{T} + \sqrt{\rho \gamma_T(\rho) T} \right) = \widetilde{O}\left(T^{\frac{3+\beta}{2+2\beta}} \right),$$

proving the second part of the claim.

■

6.5 Conclusion

In this work, we present an improved analysis for the GP-UCB algorithm in the kernelized bandit problem. We provide the first analysis showing that GP-UCB obtains sublinear regret when the underlying kernel k experiences polynomial eigendecay, which in particular implies sublinear regret rates for the practically relevant Matérn kernel. In particular, we show GP-UCB obtains regret $\widetilde{O}\left(T^{\frac{3+\beta}{2+2\beta}} \right)$ when k experiences (C, β)-polynomial eigendecay, and regret $\widetilde{O}\left(T^{\frac{\nu+2d}{2\nu+2d}} \right)$ for the Matérn kernel with smoothness ν in dimension d.

Our contributions are twofold. First, we show the importance of finding the "right" concentration inequality for tackling problems in online learning — in this case the correct bound being a self-normalized inequality originally due to Abbasi-Yadkori [2]. We provide an independent proof of a result equivalent to Corollary 3.5 of Abbasi-Yadkori [2] in Theorem 6.3.1, and hope that our simplified, truncation-based analysis will make the result more accessible to researchers working on problems in kernelized learning. Second, we demonstrate the importance of regularization in the kernelized bandit problem. In particular, since the smoothness of the kernel k

governs the hardness of learning, by regularizing in proportion to the rate of eigendecay of k, one can obtain significantly improved regret bounds.

A shortcoming of our work is that, despite obtaining the first generally sublinear regret bounds for GP-UCB, our rates are not optimal. In particular, there are discretization-based algorithms, such as SupKernelUCB [156], which obtain slightly better regret bounds of $\widetilde{O}\left(T^{\frac{1+\beta}{2\beta}}\right)$ for (C, β)-polynomial eigendecay. We hypothesize that the vanilla GP-UCB algorithm, which involves constructing confidence ellipsoids directly in the RKHS H, cannot obtain this rate.

The common line of reasoning [155] is that because the Lin-UCB (the equivalent algorithm in $\mathbb{R}^d$) obtains the optimal regret rate of $\widetilde{O}(d\sqrt{T})$ in the linear bandit problem setting, then GP-UCB should attain optimal regret as well. In the linear bandit setting, there is no subtlety between estimating the optimal action and unknown slope vector, as these are one and the same. In the kernel bandit setting, estimating the function and optimal action are not equivalent tasks. In particular, the former serves in essence as a nuisance parameter in estimating the latter: tight estimation of unknown function under the Hilbert space norm implies tight estimation of the optimal action, but not the other way around. Existing optimal algorithms are successful because they discretize the input domain, which has finite metric dimension [139], and make no attempts to estimate the unknown function in RKHS norm. Since compact sets in RKHS's do not, in general, have finite metric dimension [158], this makes estimation of the unknown function a strictly more difficult task. In fact, recent work by Lattimore [101] demonstrate that self-normalized concentration in RKHS's, in general, cannot exhibit improved dependence on maximum information gain. This further supports our hypothesis on the further unimprovability of the regret analysis of GP-UCB past the improvements made in this paper.

6.A Related Work

The kernelized bandit problem was first studied by Srinivas et al. [150], who introduce the GP-UCB algorithm and characterize its regret in both the Bayesian and Frequentist setting. While the authors demonstrate that GP-UCB obtains sublinear regret in the Bayesian setting for the commonly used kernels, their bounds fail to be sublinear in general in the frequentist setting for the Matérn kernel, one of the most popular kernel choices in practice. Chowdhury and Gopalan [30] further study the performance of GP-UCB in the frequentist setting. In particular, by leveraging a martingale-based "double mixture" argument, the authors are able to significantly simplify the confidence bounds presented in Srinivas et al. [150]. Unfortunately, the arguments introduced by Chowdhury and Gopalan [30] did not improve regret bounds beyond logarithmic factors, and thus GP-UCB continued to fail to obtain sublinear regret for certain kernels in their work. Lastly, Janz [79] are able to obtain sublinear regret guarantees for certain parameter settings of the Matérn kernel — in particular in settings where the eigenfunctions of the Hilbert-Schmidt operator associated with the kernel are uniformly bounded independent of scale (Definition 28 in the cited work).

There are many other algorithms that have been created for kernelized bandits. Janz et al. [80] introduce an algorithm specific to the Matérn kernel that obtains significantly improved regret over GP-UCB. This algorithm adaptively partitions the input domain into small hypercubes and running an instance of GP-UCB in each element of the discretized domain. Shekhar and Javidi [141] introduce an algorithm called LP-GP-UCB, which augments the GP-UCB estimator with local polynomial corrections. While in the worst case this algorithm recovers the regret bound of Chowdhury and Gopalan [30], if additional information is known about the unknown function f^* (e.g. it is Holder continuous), it can provide improved regret guarantees. Perhaps the most important non-GP-UCB algorithm in the literature is the SupKernel algorithm introduced by Valko et al. [156], which discretizes the input domain and successively eliminates actions from play. This algorithm is signficant because, despite its complicated nature, it obtains regret rates that match known lower bounds provided by Scarlett et al. [135] up to logarithmic factors.

Intimately tied to the kernelized bandit problem is the information-theoretic quantity of maximum information gain [35, 150], which is a sequential, kernel-specific measure of hardness of learning. Almost all preceding algorithms provide regret bounds in terms of the max information gain. Of particular import for our paper is the work of Vakili et al. [154]. In this work, the authors use a truncation argument to upper bound the maximum information gain of kernels in terms of their eigendecay. We directly employ these bounds in our improved analysis of GP-UCB. The max-information gain bounds presented in Vakili et al. [154] can be coupled with the regret analysis in Chowdhury and Gopalan [30] to yield a regret bound of $\widetilde{O}\left(T^{\frac{\nu+3d/2}{2\nu+d}}\right)$ in the case of the Matérn kernel with smoothness ν in dimension d. In particular, when $\nu \leq \frac{d}{2}$, this regret bound fails to be sublinear. In practical setting, d is viewed as large and ν is taken to be $3/2$ or $5/2$, making these bounds vacuous [139, 169] The regret bounds in this paper are sublinear for *any* selection of smoothness $\nu > \frac{1}{2}$ and $d \geq 1$. Moreover, a simple computation yields that our regret bounds strictly improve over (in terms of d and ν) those implied by Vakili et al. [154].

Last, we touch upon the topic of self-normalized concentration, which is an integral tool for constructing confidence bounds in UCB-like algorithms. Heuristically, self-normalized aims to

sequentially control the growth of processes that have been rescaled by their variance to look, roughly speaking, normally (or subGaussian) distributed. The prototypical example of self-normalized concentration in the bandit literature comes from Abbasi-Yadkori et al. [3], wherein the authors use a well known technique called the "method of mixtures" to construct confidence ellipsoids for finite dimensional online regression estimates. The concentration result in the aforementioned work is a specialization of results in de la Peña et al. [40], which provide self-normalized concentration for a wide variety of martingale-related processes, several of which have been recently improved [74]. In a work that is largely overlooked in the kernel bandit community, Abbasi-Yadkori [2] extend their concentration result from Abbasi-Yadkori et al. [3] to separable Hilbert spaces by using advanced functional analytic machinery. The bound we present in this work is equivalent to the aforementioned bound in separable Hilbert spaces — we provide an independent, simpler proof that avoids needing advanced tools from functional analysis. Perhaps the best-known result on concentration in Hilbert spaces is that of Chowdhury and Gopalan [30], who extend the results of Abbasi-Yadkori et al. [3] to the kernel setting using a "double mixture" technique, allowing them to construct self-normalized concentration inequalities for infinite-dimensional processes in RKHS's. This bound has historically been used in analyzing kernel bandit algorithms, although as we show in this work the bound of Abbasi-Yadkori [2] (which we independently derive in Theorem 6.3.1) is perhaps better suited for online kernelized learning problems.

6.B Technical Lemmas for Theorem 6.3.1

In this appendix, prove Theorem 6.3.1 along with several corresponding technical lemmas. While many of the following results are intuitively true, we provide their proofs in full rigor, as there can be subtleties when working in infinite-dimensional spaces. Throughout, we assume that the subGaussian noise parameter is $\sigma = 1$. The general case can readily be recovered by considering the rescaled process $(S_n/\sigma)_{n\geq 0}$.

The first lemma we present is a restriction of Theorem 6.3.1 to the case where the underlying Hilbert space $(H, \langle\cdot,\cdot\rangle_H)$ is finite dimensional, say of dimension N. In this setting, the result essentially follows immediately from Fact 6.2.4. All we need to do is construct a natural isometric isomorphism between the spaces H and $\mathbb{R}^N$, and then argue that applying such a mapping doesn't alter the norm of the self-normalized process.

Lemma 6.B.1. *Theorem 6.3.1 holds if we additionally assume that H is finite dimensional, i.e. if there exists $N \geq 1$ and orthonormal functions $\varphi_1, \dots, \varphi_N$ such that*

$$H := \operatorname{span}\{\varphi_1, \dots, \varphi_N\}.$$

Proof. Let $\tau : H \to \mathbb{R}^N$ be the map that takes a function $f = \sum_{n=1}^N \theta_n \varphi_n \in H$ to its natural embedding $\tau f := (\theta_1, \dots, \theta_N)^\top \in \mathbb{R}^N$. Not only is the map τ an isomorphism between H and $\mathbb{R}^N$, but it is also an isometry, i.e. $\|f\|_H = \|\tau f\|_2$ for all $f \in H$. Further, τ satisfies the relation $\tau^\top = \tau^{-1}$.

Define the "hatted" processes $(\widehat{S}_n)_{n\geq 1}$ and $(\widehat{V}_n)_{n\geq 1}$, which take values in $\mathbb{R}^N$ and $\mathbb{R}^{N\times N}$

respectively as

$$\widehat{S}_n = \sum_{m=1}^{n} \epsilon_m \tau k(\cdot, X_m) \qquad \text{and} \qquad \widehat{V}_n = \sum_{m=1}^{n} (\tau k(\cdot, X_m))(\tau k(\cdot, X_m))^\top.$$

It is not hard to see that, by the linearity of τ, that for any $n \geq 1$, we have $\widehat{S}_n = \tau S_n$ and $\widehat{V}_n = \tau V_n \tau^\top$. We observe that (a) $(\widehat{V}_n + \rho I_N)^{-1/2} = \tau(V_n + \rho \mathrm{id}_H)^{-1/2}\tau^\top$ and (b) that the eigenvalues of $\widehat{V}_n$ are exactly those of V_n.

Since the processes $(\widehat{S}_n)_{n\geq 1}$ and $(\widehat{V}_n)_{n\geq 1}$ satisfy the assumptions of Theorem 6.2.4, we see that the process $(M_n)_{n\geq 0}$ given by

$$M_n := \frac{1}{\sqrt{\det\Big(I_N + \rho^{-1}\widehat{V}_n\Big)}} \exp\left\{\frac{1}{2}\left\|(\rho I_N + \widehat{V}_n)^{-1/2}\widehat{S}_n\right\|_2^2\right\}$$

is a non-negative supermartingale with respect to $(\mathcal{F}_n)_{n\geq 0}$. From observation (a), the fact τ is an isometry, and the fact $\tau^\top = \tau^{-1}$, it follows that

$$\begin{aligned}\left\|(\widehat{V}_n + \rho I_N)^{-1/2}\widehat{S}_n\right\|_2 &= \left\|\tau(V_n + \rho \mathrm{id}_H)^{-1/2}\tau^\top \tau S_n\right\|_2 \\ &= \left\|(V_n + \rho \mathrm{id}_H)^{-1/2}\tau^{-1}\tau S_n\right\|_H \\ &= \left\|(V_n + \rho \mathrm{id}_H)^{-1/2} S_n\right\|_H.\end{aligned}$$

Further, observation (b) implies that

$$\det\Big(I_N + \rho\widehat{V}_n\Big) = \det(\mathrm{id}_H + \rho V_n).$$

Substituting these identities into the definition of $(M_n)_{n\geq 0}$ yields the desired result, i.e. that

$$M_n = \frac{1}{\sqrt{\det(\mathrm{id}_H + \rho^{-1}V_n)}} \exp\left\{\frac{1}{2}\left\|(V_n + \rho I_d)^{-1/2} S_n\right\|_H^2\right\}.$$

is a non-negative supermartingale with respect to $(\mathcal{F}_n)_{n\geq 0}$. The remainder of the result follows from applying Ville's Inequality (Theorem 1.0.2) and rearranging.

■

We can prove Theorem 6.3.1 by truncating the Hilbert space H onto the first N components, applying Lemma 6.B.1 to the "truncated" processes $(\pi_N S_n)_{n\geq 0}$ and $(\pi_N V_n \pi_N)_{n\geq 0}$ to construct a relevant, non-negative supermartingale $M_n^{(N)}$, and then show that the error from truncation in this non-negative supermartingale tends towards zero as N grows large. The following two technical lemmas are useful in showing that this latter truncation tends towards zero.

Lemma 6.B.2. *For any $n \geq 1$, let V_n be as in the statement of Theorem 6.3.1, and let π_N be as in Section 6.2. Then, we have*

$$\pi_N V_n \pi_N \xrightarrow[N\to\infty]{} V_n,$$

where the above convergence holds under the operator norm on H.

Proof. Fix $\epsilon > 0$, $n \geq 1$, and for $m \in [n]$, let us write $f_m = \sum_{i=1}^{\infty} \theta_i(m)\varphi_i$. Since we have assumed $\|f_n\|_H < \infty$ for all $n \geq 1$, there exists some $N_n < \infty$ such that, for all $m \in [n]$, $\|\pi_{N_n}^{\perp} f_m\|_H^2 = \sum_{i=N_n+1}^{\infty} \theta_i(m)^2 < \frac{\epsilon}{2n}$. We also have, for any $m \in [n]$ and $N \geq 1$, that f_m is an eigenfunction of $f_m f_m^{\top} \pi_N^{\perp} = f_m \langle f_m, \pi_N^{\perp}(\cdot)\rangle_H$ with corresponding (unique) eigenvalue $\|f_m f_m^{\top} \pi_N^{\perp}\|_{op} = \lambda_{\max}(f_m f_m^{\top} \pi_N^{\perp}) = \|\pi_N^{\perp} f_m\|_H^2 = \sum_{i=N+1}^{\infty} \theta_i(m)^2$. Observe that, as an orthogonal projection operator, π_N is self-adjoint, i.e. $\pi_N = \pi_N^{\top}$. With this information, we see that, for $N \geq N_n$, we have

$$\begin{aligned}
\|\pi_N V_n \pi_N - V_n\|_{op} &\leq \sum_{m=1}^{n} \left\|\pi_N f_m f_m^{\top} \pi_N - f_m f_m^{\top}\right\|_{op} \\
&= \sum_{m=1}^{n} \left\|\pi_N f_m f_m^{\top} \pi_N - \pi_N f_m f_m^{\top} + \pi_N f_m f_m^{\top} - f_m f_m^{\top}\right\|_{op} \\
&\leq \sum_{m=1}^{n} \left\|\pi_N f_m f_m^{\top} \pi_N - \pi_N f_m f_m^{\top}\right\|_{op} + \left\|\pi_N f_m f_m^{\top} - f_m f_m^{\top}\right\|_{op} \\
&\leq \sum_{m=1}^{n} \|\pi_N\|_{op} \left\|f_m f_m^{\top} \pi_N - f_m f_m^{\top}\right\|_{op} + \left\|\pi_N f_m f_m^{\top} - f_m f_m^{\top}\right\|_{op} \\
&= \sum_{m=1}^{n} 2 \left\|f_m f_m^{\top} \pi_N^{\perp}\right\|_{op} = \sum_{m=1}^{n} 2\|\pi_N^{\perp} f_m\|_H^2 < \epsilon.
\end{aligned}$$

Since $\epsilon > 0$ was arbitrary, we have shown the desired result. ■

Lemma 6.B.3. *For any $n \geq 1$, let V_n be as in Theorem 6.3.1, $\rho > 0$ arbitrary, and π_N as in Section 6.2. Then, we have*

$$\det\big(\mathrm{id}_H + \rho^{-1}\pi_N V_n \pi_N\big) \xrightarrow[N\to\infty]{} \det\big(\mathrm{id}_H + \rho^{-1} V_n\big).$$

Proof. We know that the mapping $A \mapsto \det(\mathrm{id}_H + A)$ is continuous under the "trace norm" $\|A\|_1 := \sum_{n=1}^{\infty} |\lambda_n(A)|$ [103]. Thus, to show the desired result, it suffices to show that $\|\pi_N V_n \pi_N - V_n\|_1 \xrightarrow[N\to\infty]{} 0$. Observe that both $\pi_N V_n \pi_N$ and V_n are operators of rank at most n, so so their difference $\pi_N V_n \pi_N - V_n$ has rank at most $2n$. Thus, we know that

$$\|\pi_N V_n \pi_N - V_n\|_1 \leq 2n \|\pi_N V_n \pi_N - V_n\|_{op} \xrightarrow[N\to\infty]{} 0,$$

where the final convergence follows from Lemma 6.B.2. Thus, we have shown the desired result. ■

We now tie together all of these technical (but intuitive) results in the proof of Theorem 6.3.1 below.

Proof of Theorem 6.3.1. Let $(\varphi_i)_{i\geq 1}$ be an orthonormal basis for H, and for $N \geq 1$, let π_N denote the projection operator outlined in Section 6.2. Recall that $\pi_N = \pi_N^{\top}$. Further $H_N := \mathrm{span}\{\varphi_1, \ldots, \varphi_N\} \subset H$ is the image of H under π_N. Since $(S_n)_{n\geq 0}$ is an H-valued martingale

with respect to $(\mathcal{F}_n)_{n\geq 0}$, it follows that the projected process $(\pi_N S_n)_{n\geq 0}$ is an H_N-valued martingale with respect to $(\mathcal{F}_n)_{n\geq 0}$. Further, note that the projected variance process $(\pi_N V_n \pi_N^\top)_{n\geq 0}$ satisfies

$$\pi_N V_n \pi_N^\top = \sum_{m=1}^{n} (\pi_N f_m)(\pi_N f_m)^\top.$$

Since, for any $N \geq 1$, H_N is a finite-dimensional Hilbert space, it follows from Lemma 6.B.1 that the process $(M_n^{(N)})_{n\geq 0}$ given by

$$\begin{aligned} M_n^{(N)} &:= \frac{1}{\sqrt{\widetilde{\det}(\mathrm{id}_{H_N} + \rho^{-1}\pi_N V_n \pi_N^\top)}} \exp\left\{\frac{1}{2}\left\|(\rho \mathrm{id}_{H_N} + \pi_N V_n \pi_N^\top)^{-1/2}\pi_N S_n\right\|_{H_N}^2\right\} \\ &= \frac{1}{\sqrt{\det\big(\mathrm{id}_H + \rho^{-1}\pi_N V_n \pi_N^\top\big)}} \exp\left\{\frac{1}{2}\left\|(\rho \mathrm{id}_H + \pi_N V_n \pi_N^\top)^{-1/2}\pi_N S_n\right\|_{H}^2\right\}, \end{aligned}$$

is a non-negative supermartingale with respect to $(\mathcal{F}_n)_{n\geq 0}$. In the above id_{H_N} denotes the identity id_H restricted to $H_N \subset H$ and $\widetilde{\det}$ denotes the determinant restricted to the subspace H_N. The equivalence of the second and third terms above is trivial.

We now argue that for any $n \geq 1$,

$$\lim_{N\to\infty} M_n^{(N)} = M_n. \tag{6.B.1}$$

If we show this to be true, then we have, for any $n \geq 1$

$$\begin{aligned} \mathbb{E}\left(M_n \mid \mathcal{F}_{n-1}\right) &= \mathbb{E}\left(\liminf_{N\to\infty} M_n^{(N)} \mid \mathcal{F}_{n-1}\right) \\ &\leq \liminf_{N\to\infty} \mathbb{E}\left(M_n^{(N)} \mid \mathcal{F}_{n-1}\right) \\ &\leq \liminf_{N\to\infty} M_{n-1}^{(N)} \\ &= M_{n-1}, \end{aligned}$$

which implies $(M_n)_{n\geq 0}$ is a non-negative supermartingale with respect to $(\mathcal{F}_n)_{n\geq 0}$ thus proving the result. In the above, the first inequality follows from Fatou's lemma for conditional expectations (see Durrett [50], for instance), and the second inequality follows from the supermartingale property.

Lemma 6.B.3 tells us that $\det(\mathrm{id}_H + \rho^{-1}\pi_N V_n \pi_N) \xrightarrow[N\to\infty]{} \det(\mathrm{id}_H + \rho^{-1} V_n)$ for all $n \geq 1$, so to show the desired convergence in (6.B.1), it suffices to show that

$$\|(\rho \mathrm{id}_H + \pi_N V_n \pi_N)^{-1/2}\pi_N S_n\|_H \xrightarrow[N\to\infty]{} \|(\rho \mathrm{id}_H + V_n)^{-1/2} S_n\|_H \text{ for any } n.$$

Let $\mathcal{V}_n := \rho \mathrm{id}_H + V_n$ and $\mathcal{V}_n(N) := \rho \mathrm{id}_H + \pi_N V_n \pi_N$ in the following line of reason for simplicity. We trivially have

$$\begin{aligned} \left|\|\mathcal{V}_n(N)^{-1/2}\pi_N S_n\|_H - \|\mathcal{V}_n^{-1/2} S_n\|_H\right| &\leq \left\|\mathcal{V}_n(N)^{-1/2}\pi_N S_n - \mathcal{V}_n^{-1/2} S_n\right\|_H \\ &= \left\|\mathcal{V}_n(N)^{-1/2}\pi_N S_n - \mathcal{V}_n(N)^{-1/2} S_n + \mathcal{V}_n(N)^{-1/2} S_n - \mathcal{V}_n^{-1/2} S_n\right\|_H \end{aligned}$$

$$\leq \left\|\mathcal{V}_n(N)^{-1/2}\right\|_{op} \left\|\pi_N^{\perp} S_n\right\|_H + \left\|\mathcal{V}_n(N)^{-1/2} - \mathcal{V}_n^{-1/2}\right\|_{op} \|S_n\|_H$$
$$\xrightarrow[N\to\infty]{} 0.$$

as $\lim_{N\to\infty} \|\pi_N^{\perp} f\| = 0$ for any $f \in H$ of finite norm, and Lemma 6.B.2 tells us that $\|V_n - \pi_N V_n \pi_N\|_{op} \xrightarrow[N\to\infty]{} 0$, which in turn implies that $\|\mathcal{V}_n(N)^{-1/2} - \mathcal{V}_n^{-1/2}\|_{op} = \|(\rho \mathrm{id}_H + \pi_N V_n \pi_N)^{-1/2} - (\rho \mathrm{id}_H + V_n)^{-1/2}\|_H \xrightarrow[N\to\infty]{} 0$. Thus, we have shown the desired result.

The second part of the claim follows from a direct application of Theorem 1.0.2 and rearranging.

■

As a final result in this appendix, we provide a proof of Corollary 6.3.2. This corollary allows for a more direct comparison of Theorem 6.3.1 (and thus Corollary 3.5 of Abbasi-Yadkori [2]) with those of Chowdhury and Gopalan [30]. Our proof is a simple generalization Lemma 1 in the aforementioned paper to the case of arbitrary regularization parameters.

Proof of Corollary 6.3.2. The first result is straightforward, and follows from the identity

$$\det\big(\mathrm{id}_H + \rho^{-1} V_n\big) = \det\big(I_n + \rho^{-1} K_n\big),$$

which we bring to attention in Section 6.2.

The second result follows from the following line of reasoning. Before proceeding, recall that $\Phi_n := (k(\cdot, X_1), \ldots, k(\cdot, X_n))^\top$, $V_n = \Phi_n^\top \Phi_n$, $K_n = \Phi_n \Phi_n^\top$ and $S_n = \sum_{m=1}^n \epsilon_m k(\cdot, X_m) = \Phi_n^\top \epsilon_{1:n}$.

$$\begin{aligned}
\left\|(\rho \mathrm{id}_H + V_n)^{-1} S_n\right\|_H^2 &= \epsilon_{1:n}^\top \Phi_n (\rho \mathrm{id}_H + \Phi_n^\top \Phi_n)^{-1} \Phi_n^\top \epsilon_{1:n} \\
&= \epsilon_{1:n}^\top (\rho^{-1/2} \Phi_n) \left(\mathrm{id}_H + (\rho^{-1/2}\Phi_n)^\top (\rho^{-1/2}\Phi_n)\right)^{-1} (\rho^{-1/2}\Phi_n)^\top \epsilon_{1:n} \\
&= \epsilon_{1:n}^T \rho^{-1} \Phi_n \Phi_n^\top \left(I_n + \rho^{-1} \Phi_n \Phi_n^\top\right)^{-1} \epsilon_{1:n} \\
&= \epsilon_{1:n}^\top (\rho^{-1} K_n)(I_n + \rho^{-1} K_n)^{-1} \epsilon_{1:n} \\
&= \epsilon_{1:n}^\top (I_n + \rho K_n^{-1})^{-1} \epsilon_{1:n} \\
&= \left\|(I_n + \rho K_n^{-1})^{-1/2} \epsilon_{1:n}\right\|_2^2.
\end{aligned}$$

In the above, the second equality comes from pulling out a multiplicative factor of ρ form the center operator inverse. The third inequality comes from the famed "push through" identity. Lastly, the second to last equality comes from observing that (a) $\rho^{-1} K_n$ and $(I_n + \rho^{-1} K_n)^{-1}$ are simultaneously diagonalizable matrices and (b) for scalars, we have the identity $(1 + a^{-1})^{-1} = a(1+a)^{-1}$. Thus, we have shown the desired result. ■

6.C Technical Lemmas for Theorem 6.4.1

In this appendix, we provide various technical lemmas needed for the proof of Theorem 6.4.1. We then follow these lemmas with a full proof of Theorem 6.4.1, which extends the sketch

provided in the main body of the paper. Most of the following technical lemmas either already exist in the literature [30] or are extensions of what is known in the case of finite-dimensional, linear bandits [3]. We nonetheless provide self-contained proofs for the sake of completeness.

Lemma 6.C.1. *Let $(f_n)_{n\geq 1}$ be the sequence of functions defined in Algorithm 2, and assume Assumption 4 holds. Let $\delta \in (0,1)$ be an arbitrary confidence parameter. Then, with probability at least $1-\delta$, simultaneously for all $n \geq 1$, we have*

$$\left\|(V_n + \rho \mathrm{id}_H)^{1/2}(f_n - f^*)\right\|_H \leq \sigma\sqrt{2\log\left(\frac{1}{\delta}\sqrt{\det(\mathrm{id}_H + \rho^{-1}V_n)}\right)} + \rho^{1/2}D,$$

where we recall that the right hand side equals U_n.

Proof. First, observe that we have

$$\begin{aligned}
f_n - f^* &= (\rho \mathrm{id}_H + V_n)^{-1}\Phi_n^\top Y_{1:n} - f^* \\
&= (\rho \mathrm{id}_H + V_n)^{-1}\Phi_n^\top(\Phi_n f^* + \epsilon_{1:n}) - f^* \\
&= (\rho \mathrm{id}_H + V_n)^{-1}\Phi_n^\top(\Phi_n f^* + \epsilon_{1:n}) - f^* \pm \rho(\rho \mathrm{id}_H + V_n)^{-1}f^* \\
&= (\rho \mathrm{id}_H + V_n)^{-1}\Phi_n^\top\epsilon_{1:n} - \rho(\rho \mathrm{id}_H + V_n)^{-1}f^*.
\end{aligned}$$

Applying the triangle inequality to the above, we have

$$\begin{aligned}
\left\|(\rho \mathrm{id}_H + V_n)^{1/2}(f_n - f^*)\right\|_H &\leq \left\|(\rho \mathrm{id}_H + V_n)^{-1/2}\Phi_n^\top\epsilon_{1:n}\right\|_H + \rho\left\|(\rho \mathrm{id}_H + V_n)^{-1/2}f^*\right\|_H \\
&\leq \sigma\sqrt{2\log\left(\frac{1}{\delta}\sqrt{\det(\mathrm{id}_H + \rho^{-1}V_n)}\right)} + \rho^{1/2}D.
\end{aligned}$$

To justify the final inequality, we look at each term separately. For the first term, observe that $V_n = \rho \mathrm{id}_H + \sum_{m=1}^n k(\cdot, X_n)k(\cdot, X_n)^\top$ and $S_n := \Phi_n^\top\epsilon_{1:n} = \sum_{m=1}^n \epsilon_m k(\cdot, X_m)$. Thus, we are in the setting of Theorem 6.3.1, and thus have, with probability at least $1-\delta$, simultaneously for all $t \geq 0$,

$$\left\|(\rho \mathrm{id}_H + V_n)^{-1/2}\Phi_n^\top\epsilon_{1:n}\right\|_H \leq \sigma\sqrt{2\log\left(\frac{1}{\delta}\sqrt{\det(\mathrm{id}_H + \rho^{-1}V_n)}\right)}.$$

For the second term, observe that (a) $\lambda_{\min}(\rho \mathrm{id}_H + V_n) \geq \rho$ and (b) by Assumption 4, we have $\|f^*\|_H \leq D$. Thus applying Holder's inequality, we have, deterministically

$$\rho\left\|(\rho \mathrm{id}_H + V_n)^{-1/2}f^*\right\|_H \leq \rho\left\|(\rho \mathrm{id}_H + V_n)^{-1/2}\right\|_{op}\|f^*\|_H \leq \rho^{1/2}\|f^*\|_H \leq \rho^{1/2}D.$$

These together give us the desired result.

■

The following "elliptical potential" lemma, abstractly, aims to control the the growth of the squared, self-normalized norm of the selected actions. We more or less port the argument from Abbasi-Yadkori et al. [3], which provides an analogue in the linear stochastic bandit case. We just need to be mildly careful to work around the fact we are using Fredholm determinants.

Lemma 6.C.2. *For any $n \geq 1$, let V_n be the covariance operator defined in Algorithm 2, and let $\rho > 0$ be arbitrary. We have the identity*

$$\det\big(\mathrm{id}_H + \rho^{-1}V_n\big) = \prod_{m=1}^{n}\Big(1 + \big\|(\rho\mathrm{id}_H + V_{m-1})^{-1/2}k(\cdot, X_m)\big\|_H^2\Big).$$

In particular, if $\rho \geq 1 \vee L$, where L is the bound outlined in Assumption 5, we have

$$\sum_{m=1}^{n}\big\|(\rho\mathrm{id}_H + V_{m-1})^{-1/2}k(\cdot, X_m)\big\|_H^2 \leq 2\log\det\big(\mathrm{id}_H + \rho^{-1}V_n\big).$$

Proof. Let $H_n \subset H$ be the finite-dimensional Hilbert space $H_n := \mathrm{span}\{k(\cdot, X_1), \ldots, k(\cdot, X_n)\}$. Let $\det_{H_n}$ denote the determinant restricted to H_n, i.e. the map that acts on a (symmetric) operator $A : H_n \to H_n$ by $\det_{H_n}(A) := \prod_{m=1}^{n}\lambda_m(A)$, where $\lambda_1(A), \ldots, \lambda_n(A)$ are the enumerated eigenvalues of A. Observe the identity

$$\det\big(\mathrm{id}_H + \rho^{-1}V_n\big) = \det_{H_n}(\mathrm{id}_{H_n} + \rho^{-1}V_n),$$

where we recall the determinant on the lefthand side is the Fredholm determinant, as defined in Section 6.2. Next, following the same line of reasoning as Abbasi-Yadkori et al. [3], we have

$$\begin{aligned}
&\det_{H_n}(\rho\mathrm{id}_{H_n} + V_n)\\
&= \det_{H_n}(\rho\mathrm{id}_{H_n} + V_{n-1})\det_{H_n}\big(\mathrm{id}_{H_n} + (\rho\mathrm{id}_{H_n} + V_{n-1})^{-1/2}k(\cdot, X_n)k(\cdot, X_n)^\top(\rho\mathrm{id}_{H_n} + V_{n-1})^{-1/2}\big)\\
&= \det_{H_n}(\rho\mathrm{id}_{H_n} + V_{n-1})\Big(1 + \big\|(\rho\mathrm{id}_{H_n} + V_{n-1})^{-1/2}k(\cdot, X_n)\big\|_H^2\Big)\\
&= \cdots \text{ (Iterating } n-1 \text{ more times)}\\
&= \det_{H_n}(\rho\mathrm{id}_H)\prod_{m=1}^{n}\Big(1 + \big\|(\rho\mathrm{id}_{H_n} + V_{m-1})^{-1/2}k(\cdot, X_m)\big\|_H^2\Big)\\
&= \det_{H_n}(\rho\mathrm{id}_H)\prod_{m=1}^{n}\Big(1 + \big\|(\rho\mathrm{id}_H + V_{m-1})^{-1/2}k(\cdot, X_m)\big\|_H^2\Big),
\end{aligned}$$

where the last equality comes from realizing, for all $m \in [n]$, $\|(\rho\mathrm{id}_{H_n} + V_{m-1})^{-1/2}k(\cdot, X_m)\|_H = \|(\rho\mathrm{id}_H + V_{m-1})^{-1/2}k(\cdot, X_m)\|_H$. Thus, rearranging yields

$$\det_{H_n}(\mathrm{id}_{H_n} + \rho^{-1}V_n) = \prod_{m=1}^{n}\Big(1 + \big\|(\rho\mathrm{id}_H + V_{m-1})^{-1/2}k(\cdot, X_m)\big\|_H^2\Big),$$

which yields the first part of the claim.

Now, to see the second part of the claim, observe the bound $x \leq 2\log(1+x), \forall x \in [0, 1]$. Observing that, for all $m \in [n]$, $\big\|(\rho\mathrm{id}_H + V_{m-1})^{-1/2}k(\cdot, X_m)\big\|_H \leq 1$ when $\rho \geq 1 \vee L$, we have

$$\sum_{m=1}^{n}\big\|(\rho\mathrm{id}_H + V_{m-1})^{-1/2}k(\cdot, X_m)\big\|_H^2 \leq 2\sum_{m=1}^{n}\log\Big(1 + \big\|(\rho\mathrm{id}_H + V_{m-1})^{-1/2}k(\cdot, X_m)\big\|_H^2\Big)$$

$$= 2\log\left(\prod_{m=1}^{n}\left(1+\left\|(\rho\mathrm{id}_H + V_{m-1})^{-1/2}k(\cdot,X_m)\right\|_H^2\right)\right)$$
$$= 2\log\det\left(\mathrm{id}_H + \rho^{-1}V_n\right),$$

proving the second part of the lemma. ■

With the above lemmas, along with the concentration results provided by Theorem 6.3.1, we can provide a full proof for Theorem 6.4.1.

Proof of Theorem 6.4.1. We take the standard approach of (a) first bounding instantaneous regret and then (b) applying the Cauchy-Schwarz inequality to bound the aggregation of terms. To start, for any $t \in [T]$, define the "instantaneous regret" as $r_n := f^*(x^*) - f^*(X_n)$, where we recall $x^* := \arg\max_{x\in\mathcal{X}} f^*(x)$. By applying Lemma 6.C.1, we have with probability at least $1-\delta$ that

$$\begin{aligned}
r_n &= f^*(x^*) - f^*(X_n)\\
&\le \widetilde{f}_n(X_n) - f^*(X_n)\\
&= \widetilde{f}_n(X_n) - f_{n-1}(X_n) + f_{n-1}(X_n) - f^*(X_n)\\
&= \langle \widetilde{f}_n - f_{n-1}, k(\cdot,X_n)\rangle_H - \langle f_{n-1} - f^*, k(\cdot,X_n)\rangle_H\\
&\le \left\|(\rho\mathrm{id}_H + V_{n-1})^{-1/2}k(\cdot,X_n)\right\|_H \left(\left\|(\rho\mathrm{id}_H + V_{n-1})^{1/2}(\widetilde{f}_n - f_{n-1})\right\|_H + \left\|(\rho\mathrm{id}_H + V_{n-1})^{1/2}(f_{n-1} - f^*)\right\|_H\right)\\
&\le 2U_{n-1}\left\|(\rho\mathrm{id}_H + V_{n-1})^{-1/2}k(\cdot,X_n)\right\|_H,
\end{aligned}$$

where $\widetilde{f}_n$ and f_{n-1} are as in Algorithm 2. Note that, in the above, we apply Lemma 6.C.1 in obtaining the first inequality (which is the "optimism in the face of uncertainty" part of the bound), and additionally in obtaining the last inequality. The second to last inequality follows from applying Cauchy-Schwarz.

With the above bound, we can apply again the Cauchy-Schwarz inequality to see

$$\begin{aligned}
R_T = \sum_{n=1}^{T} r_n &\le \sqrt{T\sum_{n=1}^{T} r_n^2} \le U_T\sqrt{2T\sum_{n=1}^{T}\left\|(\rho\mathrm{id}_H + V_{n-1})^{-1/2}k(\cdot,X_n)\right\|_H^2}\\
&\le U_T\sqrt{2T\log\det(\mathrm{id}_H + \rho^{-1}V_T)}\\
&= \left(\sigma\sqrt{2\log\left(\frac{1}{\delta}\sqrt{\det(\mathrm{id}_H + \rho^{-1}V_T)}\right)} + \rho^{1/2}D\right)\sqrt{2T\log\det(\mathrm{id}_H + \rho^{-1}V_T)}\\
&\le \left(\sigma\sqrt{2\log(1/\delta)} + \sigma\sqrt{2\gamma_T(\rho)} + \rho^{1/2}D\right)\sqrt{4T\gamma_T(\rho)}\\
&= \sigma\gamma_T(\rho)\sqrt{8T} + D\sqrt{4\rho\gamma_T(\rho)T} + \sigma\sqrt{8T\log(1/\delta)}\\
&= O\left(\gamma_T(\rho)\sqrt{T} + \sqrt{\rho\gamma_T(\rho)T}\right).
\end{aligned}$$

In the above, the second inequality follows from the second part of Lemma 6.C.2, the following equality follows from substituting in U_T, and the final inequality follows from the definition of

the maximum information gain $\gamma_T(\rho)$ and the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for all $a, b \geq 0$. The last, big-Oh bound is straightforward. With this, we have proven the first part of the theorem.

Now, suppose the kernel k experiences (C, β)-polynomial eigendecay. Then, by Fact 6.2.2, we know that

$$\begin{aligned}\gamma_T(\rho) &\leq \left(\left(\frac{CB^2T}{\rho}\right)^{1/\beta} \log^{-1/\beta}\left(1 + \frac{LT}{\rho}\right) + 1\right) \log\left(1 + \frac{LT}{\rho}\right) \\ &= \widetilde{O}\left(\left(\frac{T}{\rho}\right)^{1/\beta}\right).\end{aligned}$$

We aim to set $\rho \asymp \left(\frac{T}{\rho}\right)^{1/\beta}$, which occurs when $\rho = O(T^{\frac{1}{1+\beta}})$. When this happens, we have

$$\left(\frac{T}{\rho}\right)^{1/\beta} \sqrt{T} = T^{\frac{1}{1+\beta} + \frac{1}{2}} = T^{\frac{3+\beta}{2+2\beta}}.$$

Applying this, we have that

$$\begin{aligned}R_T &= O\left(\gamma_T(\rho)\sqrt{T} + \sqrt{\rho\gamma_T(\rho)T}\right) \\ &= \widetilde{O}\left(T^{\frac{3+\beta}{2+2\beta}}\right),\end{aligned}$$

which, in particular, is sublinear for any $\beta > 1$. Thus, we are done.

■

Part III

Conclusions and Future Research Directions

Chapter 7

Concluding Remarks and Open Problems

Modern machine learning and data science are inherently sequential. For many tasks, a learner must estimate unknown statistical quantities as data is adaptively collected over time. This data is typically highly-correlated, often being generated through some human-in-the-loop process. For instance, when a learner runs statistical queries in sequence using differentially private algorithms, they must bound the amount of information that is leaked in order to ensure a target privacy level is met. Likewise, in bandit optimization tasks, a learner must estimate unknown reward functions as data is adaptively collected according to some policy. Classical statistical methods, which provide convergence guarantees under i.i.d. assumptions, fail to allow the learner to perform inference in these settings. Instead, to form valid conclusions, researchers must turn to martingale methods.

Martingale methods are often treated as a "trick" or as a "means to an end", with martingale concentration results being applied in many machine learning papers in an ad-hoc, black-box manner. Naive application of martingale concentration inequalities can lead to suboptimal convergence rates, and thus often an inappropriate treatment of the topic at hand. We view martingales and martingale methods as an end in themselves. That is, we believe that they warrant thorough, independent study.

In this thesis, we provided a "start-to-finish" treatment of martingale concentration. In the first part of this work, we focused on the theory underlying martingale concentration. We proved time-uniform bounds on the growth of self-normalized martingales in scalar and multivariate settings, with our focus being on constructing bounds that can be readily tuned to almost any tail assumptions. We also derived novel martingale concentration inequalities in infinite-dimensional spaces and provided intrinsic links between these bounds and heavy-tailed mean estimation. The proofs of our results in these sections were often simple and geometric, offering fundamental insights into how martingales concentrate in a variety of settings.

In the second half of this thesis, we switched our focus to the application of martingale methods in practically relevant data science and machine learning tasks. We focused on two applications of time-uniform martingale concentration to differentially private machine learning. These applications required the careful use martingale concentration inequalities to control the growth of privacy loss martingales, and resulted in significant improvements over the current state-of-the-art. Likewise, we also showed how applying improved self-normalized concentration inequalities in separable Hilbert spaces could lead to sublinear regret rates in online kernel-

ized learning tasks where previously only vacuous rates existed.

We now conclude by discussing some interesting open problems related to the work covered in this thesis. We enumerate several interesting directions in the following paragraphs.

7.1 Open Questions

While this thesis thoroughly covered both the foundations and applications of martingale methods, there are still many interesting open questions that remain to be answered. The questions enumerated below aim to greatly extend the generality of the results contained within this document. For instance, this includes removing dimension dependence in self-normalized bounds to extend finite-dimensional results to infinite-dimensional settings. We are unsure if the questions below can be answered in the affirmative, but nonetheless find it important to include them.

Time-Uniform Martingale Concentration under Dual Norms In Chapter 3, we showed that a simple, truncation-based estimator could estimate a unknown, infinite-dimensional mean at a rate that matched that of geometric median-of-medians estimator due to Minsker [121]. To prove our results, we generalized classical results on the concentration of bounded observations in *smooth Banach spaces* originally due to Pinelis [128, 129, 74]. The assumption of smoothness is quite general, and examples of smooth Banach spaces include all separable Hilbert spaces, ℓ^α sequence spaces for $\alpha \geq 2$, and L^α function spaces for $\alpha \geq 2$. Unfortunately, our results did not apply for ℓ^α and L^α spaces when $\alpha < 2$.

In principal, one would expect a certain amount symmetry for ℓ^α and L^α norms. In particular, if we let η denote the Holder conjugate of α (i.e. the value η defined through $1/\alpha + 1/\eta = 1$), one would hope that rates of concentration under the ℓ^η and L^η norms would be similar to that of ℓ^α and L^α spaces. More broadly, if we assume that some reflexive Banach spaces $(\mathbb{B}, \|\cdot\|)$ is smooth, can we expect to obtain similar rates of concentration in the continuous dual space $(\mathbb{B}^*, \|\cdot\|_*)$? We make the following hypothesis, which may or may not be true.

Hypothesis 1. *Let $(S_n)_{n\geq 0}$ be a martingale taking values in the continuous dual $(\mathbb{B}^*, \|\cdot\|_*)$ of a reflexive, smooth Banach space $(\mathbb{B}, \|\cdot\|)$. Further, suppose $\mathbb{E}_{n-1}\|\Delta S_n\|_* \leq 1$ almost surely. Let, for a confidence parameter $\delta \in (0,1)$, $(U_n^\delta)_{n\geq 1}$ be a sequence satisfying*

$$\mathbb{P}(\exists n \geq 0 : \|T_n\| \geq U_n^\delta) \leq \delta$$

for all martingales/filtration pairs $(T_n, \mathcal{G}_n)_{n\geq 0}$ in $(\mathbb{B}, \|\cdot\|)$ with $\mathbb{E}_{n-1}\|\Delta T_n\| \leq 1$ almost surely. Then, one has

$$\mathbb{P}(\exists n \geq 0 : \|S_n\|_* \geq U_n^\delta) \leq \delta.$$

Dimension-Independent Self-Normalized Concentration Next, we spend some time discussing open directions related to self-normalized concentration. The results presented in Chapter 2 applied to general classes of self-normalized processes and served as a significant generalization of existing results due to de la Peña et al. [40, 41]. While the latter set of bounds could only be applied under sub-Gaussian tail conditions, our bounds could be applied in both light-tailed (sub-Gaussian, sub-Gamma, sub-Exponential, sub-Poisson) and heavy-tailed (symmetric

increments, finite variance) settings. Even in settings where both types of bounds could be applied, the results were incomparable. The self-normalized results due de la Peña, which relied upon the method of mixtures [98, 99], depended on the accumulated variance $(V_n)_{n\geq 0}$ through a term that looked like the log-determinant of V_n. On the other hand, our bound depended on the logarithm of the condition number of V_n. In some settings, our bounds would be preferable, and in others, those of de la Peña would be tighter.

One advantage of the sub-Gaussian method of mixtures bounds is that they don't explicitly depend on the ambient dimension of the space. Rather, the ambient dimension is implicitly captured through the log-determinant term. Due to our derivation relying on a sequence of covers of the unit ball, our bounds have explicit dependence on the ambient dimension d. We hypothesize that, for general sub-ψ processes, it is possible to construct time-uniform concentration results that do not explicitly depend on the dimension of the space. In particular, these bounds would be applicable in general separable Hilbert spaces, including ℓ^2 sequence spaces, L^2 function spaces, and reproducing kernel Hilbert spaces (RKHS's). We formalize this hypothesis as follows, wherein $u_\psi(V_n, \delta)$ should be viewed as a time-uniform bound that depends on the underlying CGF-like function ψ, the accumulated variance process $(V_n)_{n\geq 0}$, and the chosen failure probability $\delta \in (0, 1)$, but *not* the dimension $\dim(\mathbb{H})$ of the separable Hilbert space $(\mathbb{H}, \langle \cdot, \cdot \rangle)$.

Hypothesis 2. *Let $(S_n)_{n\geq 0}$ be a process in some separable Hilbert space $(\mathbb{H}, \langle \cdot, \cdot \rangle)$ and assume (S_n) is sub-ψ with variance proxy $(V_n)_{n\geq 0}$. Then, there exists some function $u_\psi(V_n, \delta)$ such that, for any $\delta \in (0, 1)$, we have*

$$\mathbb{P}\left(\exists n \geq 0 : \left\|V_n^{-1/2} S_n\right\| \geq u_\psi(V_n, \delta)\right) \leq \delta.$$

Noise Reduction for Sequential Differentially Private Algorithms Finally, we return to the problem of developing noise reduction algorithms for differentially private empirical risk minimization. In Chapter 5, we described *the Brownian mechanism* BM, which added correlated Gaussian noise to a risk minimizing parameter, and then slowly stripped away this noise until a target accuracy was met. We similarly described analogues of the Brownian mechanisms for other additive noise distributions such as Skellam noise and Laplace noise. While our algorithm was naturally applicable to lightweight statistical models, such as ridge and logistic regression, it was not readily applicable to more complicated models such as deep neural networks.

Why was this the case? While is it straightforward to measure the sensitivity of the outputs of simple regression algorithms, it is extremely difficult if not impossible to measure how much the training trajectory of a neural network changes if a single data point is removed from the training set. To get around this, researchers train neural networks not by using additive noise mechanisms on the final trained model, but rather by injecting appropriately scaled Gaussian noise during model training via stochastic gradient descent. The Brownian mechanism, as outlined in Chapter 5, is just an additive noise mechanism. It is not clear how to couple such a mechanism with iterative training methods such as private SGD. We leave it as interesting and high impact future work to figure out how to combine the two algorithms.

7.2 Future Directions

Over the course of my PhD, I have worked on problems in a variety of fields. I started my PhD by working on problems in scheduling and queueing theory. Halfway through my second year, I switched to working on problems related to differential privacy and differentially private machine learning. I felt like I first really hit my stride in my fourth year, when I started working on more abstract problems related to martingale concentration. I enjoyed the generality of the problems and the broad applicability of the results. Luckily, it seems I was able to sew together the various projects I've worked on into a relatively coherent thesis.

So, what's next? In the final year of my PhD, I began studying the connections between causal inference and model calibration [167]. I enjoyed the theoretical depth, room for experimentation, and practical relevance of this new research direction. Motivated by this, I'll soon begin a postdoc focused on the intersection of causal inference and machine learning. Once again, this marks a substantial departure from my previous line of work. While changes can always be a little frightening, the excitement of learning something new is what drew me to the PhD in the first place.

Bibliography

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

[2] Yasin Abbasi-Yadkori. *Online learning for linearly parametrized control problems*. PhD thesis, University of Alberta, 2013.

[3] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24, 2011.

[4] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Online least squares estimation with self-normalized processes: An application to bandit problems. *arXiv preprint arXiv:1102.2670*, 2011.

[5] Anish Agarwal, Muhammad Jehangir Amjad, Devavrat Shah, and Dennis Shen. Model agnostic time series analysis via matrix estimation. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2(3):1–39, 2018.

[6] Naman Agarwal, Peter Kairouz, and Ziyu Liu. The skellam mechanism for differentially private federated learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[7] Rajeev Agrawal. The continuum-armed bandit problem. *SIAM Journal on Control and Optimization*, 33(6):1926–1951, 1995.

[8] Rudolf Ahlswede and Andreas Winter. Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579, 2002.

[9] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*, pages 20–29, 1996.

[10] Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

[11] Borja Balle and Yu-Xiang Wang. Improving the gaussian mechanism for differential pri-

vacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning*, pages 394–403. PMLR, 2018.

[12] Andrew G Barto. Reinforcement learning control. *Current Opinion in Neurobiology*, 4 (6):888–893, 1994.

[13] George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

[14] Bernard Bercu and Abderrahmen Touati. Exponential inequalities for self-normalized martingales with applications. *The Annals of Applied Probability*, 18(5):1848–1869, 2008.

[15] Bernard Bercu and Taieb Touati. New insights on concentration inequalities for self-normalized martingales. *Electronic Communications in Probability*, 24:1 – 12, 2019.

[16] D Blackwell. Large deviations for martingales. *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 89–91, 1997.

[17] David Blackwell. Equivalent comparisons of experiments. *The annals of mathematical statistics*, pages 265–272, 1953.

[18] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: a nonasymptotic theory of independence*. Oxford university press, 2013.

[19] Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

[20] Mark Bun and Thomas Steinke. Concentrated differential privacy: simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

[21] Olivier Catoni. PAC-Bayesian supervised classification: the thermodynamics of statistical learning. *arXiv preprint arXiv:0712.0248*, 2007.

[22] Olivier Catoni and Ilaria Giulini. Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv preprint arXiv:1712.02747*, 2017.

[23] Olivier Catoni and Ilaria Giulini. Dimension-free PAC-Bayesian bounds for the estimation of the mean of a random vector. *arXiv preprint arXiv:1802.04308*, 2018.

[24] Mark Cesar and Ryan Rogers. Bounding, concentrating, and truncating: Unifying privacy loss composition for data analytics. In *Algorithmic Learning Theory*, pages 421–457. PMLR, 2021.

[25] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

[26] Peng Chen, Xinghu Jin, Xiang Li, and Lihu Xu. A generalized Catoni's M-estimator under finite α-th moment assumption with $\alpha \in (1, 2)$. *Electronic Journal of Statistics*, 15 (2):5523–5544, 2021.

[27] Shanshan Chen, Zhenping Wang, Wenfei Xu, and Yu Miao. Exponential inequalities for self-normalized martingales. *Journal of Inequalities and Applications*, 2014(289):1–12, 2014.

[28] Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L Bartlett. Fast mean estimation with sub-Gaussian rates. In *Conference on Learning Theory*, pages 786–806. PMLR, 2019.

[29] Alejandro Cholaquidis, Emilien Joly, and Leonardo Moreno. GROS: A general robust aggregation strategy. *arXiv preprint arXiv:2402.15442*, 2024.

[30] Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pages 844–853. PMLR, 2017.

[31] Demetres Christofides and Klas Markström. Expansion properties of random Cayley graphs and vertex transitive graphs via matrix martingales. *Random Structures & Algorithms*, 32(1):88–100, 2008.

[32] Ben Chugg, Hongjian Wang, and Aaditya Ramdas. Time-uniform confidence spheres for means of random vectors. *arXiv preprint arXiv:2311.08168*, 2023.

[33] Ben Chugg, Hongjian Wang, and Aaditya Ramdas. A unified recipe for deriving (time-uniform) PAC-Bayes bounds. *arXiv preprint arXiv:2302.03421*, 2023.

[34] Michael B Cohen, Ben Cousins, Yin Tat Lee, and Xin Yang. A near-optimal algorithm for approximating the John ellipsoid. In *Conference on Learning Theory*, pages 849–873. PMLR, 2019.

[35] Thomas M Cover and Joy A Thomas. Information theory and statistics. *Elements of Information Theory*, 1(1):279–335, 1991.

[36] Ashok Cutkosky. Combining online learning guarantees. In *Conference on Learning Theory*, pages 895–913. PMLR, 2019.

[37] DA Darling and Herbert Robbins. Iterated logarithm inequalities. *Proceedings of the National Academy of Sciences*, 57(5):1188–1192, 1967.

[38] DA Darling and Herbert Robbins. Some further remarks on inequalities for sample sums. *Proceedings of the National Academy of Sciences*, 60(4):1175–1182, 1968.

[39] Serge Darolles, Christian Gourieroux, and Joann Jasiak. Structural Laplace transform and compound autoregressive models. *Journal of Time Series Analysis*, 27(4):477–503, 2006.

[40] Victor de la Peña, Michael J Klass, and Tze Leung Lai. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *The Annals of Probability*, 32(3):1902 – 1933, 2004.

[41] Victor de la Peña, Michael J Klass, and Tze Leung Lai. Pseudo-maximization and self-normalized processes. *Probability Surveys*, 4:172 – 192, 2007.

[42] Victor de la Peña, Michael J Klass, and Tze Leung Lai. Theory and applications of multivariate self-normalized processes. *Stochastic Processes and their Applications*, 119(12): 4210–4227, 2009.

[43] Victor de la Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized Processes: Limit Theory and Statistical Applications*. Springer, 2009.

[44] Victor H de la Peña, Michael J Klass, and Tze Leung Lai. Theory and applications of multivariate self-normalized processes. *Stochastic Processes and their Applications*, 119 (12):4210–4227, 2009.

[45] Victor H. de la Peña, Michael J. Klass, and Tze Leung Lai. Pseudo-maximization and self-normalized processes. *Probability Surveys*, 4:172 – 192, 2007. doi: 10.1214/07-PS119.

[46] Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. In *Journal of the Royal Statistical Society: Series B*, pages 1–35, 2021.

[47] Joseph L Doob. Regularity properties of certain families of chance variables. *Transactions of the American Mathematical Society*, 47(3):455–486, 1940.

[48] Audrey Durand, Odalric-Ambrym Maillard, and Joelle Pineau. Streaming kernel regression with provably adaptive mean, variance, and regularization. *The Journal of Machine Learning Research*, 19(1):650–683, 2018.

[49] Richard Durrett. *Probability: theory and examples*. Duxbury Press, Belmont, CA, second edition, 1996. ISBN 0-534-24318-5.

[50] Rick Durrett. *Probability: Theory and Examples*, volume 49. Cambridge university press, 2019.

[51] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the Forty-first Annual ACM Symposium on Theory of Computing*, pages 371–380, 2009.

[52] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[53] Cynthia Dwork and Guy N. Rothblum. Concentrated differential privacy. *CoRR*, abs/1603.01887, 2016.

[54] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 486–503. Springer, 2006.

[55] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[56] Cynthia Dwork, Moni Naor, Omer Reingold, Guy N Rothblum, and Salil Vadhan. On the complexity of differentially private data release: efficient algorithms and hardness results. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 381–390, 2009.

[57] Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 51–60. IEEE, 2010.

[58] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pages 1054–1067, 2014.

[59] Xiequan Fan, Ion Grama, and Quansheng Liu. Exponential inequalities for martingales with applications. *Electronic Journal of Probability*, 20(none):1 – 22, 2015.

[60] Vivek Farias, Ciamac Moallemi, Tianyi Peng, and Andrew Zheng. Synthetically controlled bandits. *arXiv preprint arXiv:2202.07079*, 2022.

[61] Vitaly Feldman and Tijana Zrnic. Individual privacy accounting via a Rényi filter. *Advances in Neural Information Processing Systems*, 2021.

[62] Vitaly Feldman and Tijana Zrnic. Individual privacy accounting via a Rényi filter. *Advances in Neural Information Processing Systems*, 34, 2021.

[63] Stephen E Fienberg, Alessandro Rinaldo, and Xiaolin Yang. Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables. In *International Conference on Privacy in Statistical Databases*, pages 187–199. Springer, 2010.

[64] David A Freedman. On tail probabilities for martingales. *The Annals of Probability*, pages 100–118, 1975.

[65] Andy Greenberg. Apple's 'differential privacy' is about collecting your data—but not your data. *Wired Magazine*, 2016.

[66] Vincent Guingona, Alexei Kolesnikov, Julianne Nierwinski, and Avery Schweitzer. Comparing approximate and probabilistic differential privacy parameters. *Information Pro-*

cessing Letters, page 106380, 2023.

[67] Shivam Gupta, Samuel Hopkins, and Eric Price. Beyond Catoni: Sharper rates for heavy-tailed and robust mean estimation. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 2232–2269. PMLR, 2024.

[68] James D Hamilton. *Time Series Analysis*. Princeton university press, 2020.

[69] Philip Hartman and Aurel Wintner. On the law of the iterated logarithm. *American Journal of Mathematics*, 63(1):169–176, 1941.

[70] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[71] Matthew Hoffman, Eric Brochu, and Nando De Freitas. Portfolio allocation for Bayesian optimization. In *UAI*, pages 327–336, 2011.

[72] Samuel B Hopkins. Mean estimation with sub-Gaussian rates in polynomial time. *The Annals of Statistics*, 48(2):1193–1213, 2020.

[73] Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257 – 317, 2020. doi: 10.1214/18-PS321.

[74] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.

[75] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2): 1055–1080, 2021.

[76] Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055 – 1080, 2021. doi: 10.1214/20-AOS1991.

[77] Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, 17(18):1–40, 2016.

[78] Peter J Huber. Robust statistics. *Wiley Series in Probability and Mathematical Statistics*, 1981.

[79] David Janz. *Sequential decision making with feature-linear models*. PhD thesis, 2022.

[80] David Janz, David Burt, and Javier González. Bandit optimisation of functions in the matérn kernel RKHS. In *International Conference on Artificial Intelligence and Statistics*,

pages 2486–2495. PMLR, 2020.

[81] Mark R Jerrum, Leslie G Valiant, and Vijay V Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169–188, 1986.

[82] Fritz John. Extremum problems with inequalities as subsidiary conditions. *Traces and Emergence of Nonlinear Programming*, pages 197–215, 2014.

[83] Kwang-Sung Jun and Francesco Orabona. Parameter-free online convex optimization with sub-exponential noise. In *Conference on Learning Theory*, pages 1802–1823. PMLR, 2019.

[84] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International Conference on Machine Learning*, pages 1376–1385. PMLR, 2015.

[85] Shiva P Kasiviswanathan and Adam Smith. On the semantics of differential privacy: A Bayesian formulation. *Journal of Privacy and Confidentiality*, 6(1), 2014.

[86] Emilie Kaufmann and Wouter M Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22 (246):1–44, 2021.

[87] Emilie Kaufmann and Wouter M Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22 (246):1–44, 2021.

[88] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17: 1–42, 2016.

[89] François Kawala, Ahlame Douzal-Chouakria, Eric Gaussier, and Eustache Dimert. Prédictions d'activité dans les réseaux sociaux en ligne. In *4ième Conférence Sur les Modèles et L'analyse des Réseaux: Approches Mathématiques et Informatiques*, page 16, 2013.

[90] KDD. KDD cup 1999 data, 1999.

[91] Robert W Keener. *Theoretical Statistics: Topics for a Core Course*. Springer, 2010.

[92] Harry Kesten. The 1971 rietz lecture sums of independent random variables–without moment conditions. *The Annals of Mathematical Statistics*, pages 701–732, 1972.

[93] Antti Koskela, Marlon Tobaben, and Antti Honkela. Individual privacy accounting with gaussian differential privacy. *CoRR*, abs/2209.15596, 2022. doi: 10.48550/arXiv.2209.

15596. URL https://doi.org/10.48550/arXiv.2209.15596.

[94] Fragkiskos Koufogiannis, Shuo Han, and George J Pappas. Gradual release of sensitive data under differential privacy. *Journal of Privacy and Confidentiality*, 7(2), 2017.

[95] Akshay Krishnamurthy, Zhiwei Steven Wu, and Vasilis Syrgkanis. Semiparametric contextual bandits. In *International Conference on Machine Learning*, pages 2776–2785. PMLR, 2018.

[96] Arun K Kuchibhotla and Rohit K Patra. On least squares estimation under heteroscedastic and heavy-tailed errors. *The Annals of Statistics*, 50(1):277–302, 2022.

[97] Arun Kumar Kuchibhotla and Abhishek Chakrabortty. Moving beyond sub-Gaussianity in high-dimensional statistics: applications in covariance estimation and linear regression. *Information and Inference: A Journal of the IMA*, 11(4):1389–1456, 2022.

[98] Tze Leung Lai and Herbert Robbins. Consistency and asymptotic efficiency of slope estimates in stochastic approximation schemes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 56(3):329–360, 1981.

[99] Tze Leung Lai and Ching Zong Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1):154–166, 1982.

[100] John Langford and John Shawe-Taylor. PAC-Bayes & margins. *Advances in Neural Information Processing Systems*, 15, 2002.

[101] Tor Lattimore. A lower bound for linear and kernel regression with adaptive covariates. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2095–2113. PMLR, 2023.

[102] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[103] Peter D Lax. *Functional Analysis*, volume 55. John Wiley & Sons, 2002.

[104] Jean-François Le Gall. *Brownian motion, martingales, and stochastic calculus*. Springer, 2016.

[105] Mathias Lécuyer. Practical privacy filters and odometers with Rényi differential privacy and applications to differentially private deep learning. *arXiv Preprint arXiv:2103.01379*, 2021.

[106] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and processes*. Springer Science & Business Media, 2013.

[107] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide web*, pages 661–670, 2010.

[108] Xiaoyu Li and Francesco Orabona. A high probability analysis of adaptive SGD with momentum. *arXiv preprint arXiv:2007.14294*, 2020.

[109] Katrina Ligett, Seth Neel, Aaron Roth, Bo Waggoner, and Steven Z Wu. Accuracy first: Selecting a differential privacy level for accuracy constrained erm. *Advances in Neural Information Processing Systems*, 30, 2017.

[110] Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

[111] Gábor Lugosi and Shahar Mendelson. Near-optimal mean estimators with respect to general norms. *Probability theory and related fields*, 175(3):957–973, 2019.

[112] Gábor Lugosi and Shahar Mendelson. Sub-Gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783–794, 2019.

[113] Gabor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: the optimality of trimmed mean. *Annals of Statistics*, 2021.

[114] Min Lyu, Dong Su, and Ninghui Li. Understanding the sparse vector technique for differential privacy. *Proc. VLDB Endow.*, 10(6):637–648, Feb 2017.

[115] RA Maller. On the law of the iterated logarithm in the infinite variance case. *Journal of the Australian Mathematical Society*, 30(1):5–14, 1980.

[116] Tudor Manole and Aaditya Ramdas. Martingale methods for sequential estimation of convex functionals and divergences. *IEEE Transactions on Information Theory*, 2023.

[117] Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. *Operations Research*, 68(4):1132–1161, 2020.

[118] Diego Martinez-Taboada and Aaditya Ramdas. Empirical Bernstein in smooth Banach spaces. *arXiv preprint arXiv:2409.06060*, 2024.

[119] Aditya Mate, Jackson Killian, Haifeng Xu, Andrew Perrault, and Milind Tambe. Collapsing bandits and their application to public health intervention. *Advances in Neural Information Processing Systems*, 33:15639–15650, 2020.

[120] Timothée Mathieu. Concentration study of M-estimators using the influence function. *Electronic Journal of Statistics*, 16(1):3695–3750, 2022.

[121] Stanislav Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, pages 2308–2335, 2015.

[122] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

[123] Jack Murtagh and Salil Vadhan. The complexity of computing the optimal composition of differential privacy. In *Theory of Cryptography Conference*, pages 157–175. Springer, 2016.

[124] Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

[125] Roberto I Oliveira and Paulo Orenstein. The sub-Gaussian property of trimmed means estimators. *Technical Report, IMPA*, 2019.

[126] Maria Pacurar. Autoregressive conditional duration models in finance: a survey of the theoretical and empirical literature. *Journal of Economic Surveys*, 22(4):711–751, 2008.

[127] Nicolas Papernot and Thomas Steinke. Hyperparameter tuning with renyi differential privacy. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL `https://openreview.net/forum?id=-70L8lpp9DF`.

[128] Iosif Pinelis. An approach to inequalities for the distributions of infinite-dimensional martingales. *Probability in Banach Spaces*, 8:128–134, 1992.

[129] Iosif Pinelis. Optimum bounds for the distributions of martingales in Banach paces. *The Annals of Probability*, pages 1679–1706, 1994.

[130] Aleksandr Podkopaev, Patrick Blöbaum, Shiva Kasiviswanathan, and Aaditya Ramdas. Sequential kernelized independence testing. In *International Conference on Machine Learning*, pages 27957–27993. PMLR, 2023.

[131] Aaditya Ramdas and Ruodu Wang. Hypothesis testing with e-values. *arXiv preprint arXiv:2410.23614*, 2024.

[132] Herbert Robbins. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409, 1970.

[133] Ryan M Rogers, Aaron Roth, Jonathan Ullman, and Salil Vadhan. Privacy odometers and filters: pay-as-you-go composition. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[134] Gabriele Santin and Robert Schaback. Approximation of eigenfunctions in kernel-based spaces. *Advances in Computational Mathematics*, 42(4):973–993, 2016.

[135] Jonathan Scarlett, Ilija Bogunovic, and Volkan Cevher. Lower bounds on regret for noisy Gaussian process bandit optimization. In *Conference on Learning Theory*, pages 1723–1742. PMLR, 2017.

[136] Henry Scheffe. *The Analysis of Variance*, volume 72. John Wiley & Sons, 1999.

[137] Xiaofeng Shao. Self-normalization for time series: a review of recent developments. *Journal of the American Statistical Association*, 110(512):1797–1817, 2015.

[138] Or Sheffet. Differentially private ordinary least squares. In *International Conference on Machine Learning*, pages 3105–3114. PMLR, 2017.

[139] Shubhanshu Shekhar and Tara Javidi. Gaussian process bandits with adaptive discretization. *Electronic Journal of Statistics*, 12(2):3829 – 3874, 2018.

[140] Shubhanshu Shekhar and Tara Javidi. Multi-scale zero-order optimization of smooth functions in an RKHS. *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 288–293, 2020.

[141] Shubhanshu Shekhar and Tara Javidi. Instance dependent regret analysis of kernelized bandits. In *International Conference on Machine Learning*, pages 19747–19772. PMLR, 2022.

[142] Shubhanshu Shekhar and Aaditya Ramdas. Nonparametric two-sample testing by betting. *IEEE Transactions on Information Theory*, 2023.

[143] Shubhanshu Shekhar, Ilmun Kim, and Aaditya Ramdas. A permutation-free kernel two-sample test. *Advances in Neural Information Processing Systems*, 35:18168–18180, 2022.

[144] Aleksandrs Slivkins et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

[145] Adam Smith, Abhradeep Thakurta, and Jalaj Upadhyay. Is interaction necessary for distributed private learning? In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 58–77. IEEE, 2017.

[146] Adam D. Smith and Abhradeep Thakurta. Fully adaptive composition for gaussian differential privacy. *CoRR*, abs/2210.17520, 2022. doi: 10.48550/arXiv.2210.17520. URL `https://doi.org/10.48550/arXiv.2210.17520`.

[147] Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 27, 2014.

[148] Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 245–248. IEEE, 2013.

[149] Zhao Song, Xin Yang, Yuanyuan Yang, and Tianyi Zhou. Faster algorithm for structured John ellipsoid computation. *arXiv preprint arXiv:2211.14407*, 2022.

[150] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *International Conference on Machine Learning*, page 1015–1022. PMLR, 2010.

[151] Joel Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16, 2011.

[152] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12:389–434, 2012.

[153] John W Tukey and Donald H McLaughlin. Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/winsorization 1. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 331–352, 1963.

[154] Sattar Vakili, Kia Khezeli, and Victor Picheny. On information gain and regret bounds in gaussian process bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 82–90. PMLR, 2021.

[155] Sattar Vakili, Jonathan Scarlett, and Tara Javidi. Open problem: Tight online confidence intervals for RKHS elements. In *Conference on Learning Theory*, pages 4647–4652. PMLR, 2021.

[156] Michal Valko, Nathaniel Korda, Rémi Munos, Ilias Flaounas, and Nelo Cristianini. Finite-time analysis of kernelised contextual bandits. *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, 2013.

[157] Jean Ville. Etude critique de la notion de collectif. *Bull. Amer. Math. Soc*, 45(11):824, 1939.

[158] Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge university press, 2019.

[159] Hongjian Wang and Aaditya Ramdas. Catoni-style confidence sequences for heavy-tailed mean estimation. *Stochastic Processes and Their Applications*, 163:168–202, 2023.

[160] Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled Rényi differential privacy and analytical moments accountant. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1226–1235. PMLR, 2019.

[161] Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2023.

[162] Justin Whitehouse, Aaditya Ramdas, Steven Z Wu, and Ryan M Rogers. Brownian noise reduction: Maximizing privacy subject to accuracy constraints. *Advances in Neural Information Processing Systems*, 35:11217–11228, 2022.

[163] Justin Whitehouse, Aaditya Ramdas, Ryan Rogers, and Steven Wu. Fully-adaptive composition in differential privacy. In *International Conference on Machine Learning*, pages 36990–37007. PMLR, 2023.

[164] Justin Whitehouse, Aaditya Ramdas, and Steven Z Wu. On the sublinear regret of GP-UCB. *Advances in Neural Information Processing Systems*, 36:35266–35276, 2023.

[165] Justin Whitehouse, Zhiwei Steven Wu, and Aaditya Ramdas. On the sublinear regret of GP-UCB. *Advances in Neural Information Processing Systems*, 2023.

[166] Justin Whitehouse, Zhiwei Steven Wu, and Aaditya Ramdas. Time-uniform self-normalized concentration for vector-valued processes. *arXiv preprint arXiv:2310.09100*, 2023.

[167] Justin Whitehouse, Christopher Jung, Vasilis Syrgkanis, Bryan Wilder, and Zhiwei Steven Wu. Orthogonal causal calibration. *arXiv preprint arXiv:2406.01933*, 2024.

[168] Peter Whittle. Multi-armed bandits and the Gittins index. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):143–149, 1980.

[169] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[170] Yuqing Zhu, Jinshuo Dong, and Yu-Xiang Wang. Optimal accounting of differential privacy via characteristic function. In *International Conference on Artificial Intelligence and Statistics*, pages 4782–4817. PMLR, 2022.