

# A variational approach to dimension-free self-normalized concentration

Ben Chugg<sup>1</sup> and Aaditya Ramdas<sup>1</sup>

<sup>1</sup>Departments of Machine Learning and Statistics, CMU  
`{benchugg, aramdass}@cmu.edu`

August 11, 2025

## Abstract

We study the self-normalized concentration of vector-valued stochastic processes. We focus on bounds for sub- $\psi$  processes, a tail condition that encompasses a wide variety of well-known distributions (including sub-exponential, sub-Gaussian, sub-gamma, and sub-Poisson distributions). Our results recover and generalize the influential bound of Abbasi-Yadkori et al. [1] and fill a gap in the literature between determinant-based bounds and those based on condition numbers. As applications we prove a Bernstein inequality for random vectors satisfying a moment condition (which is more general than boundedness), and also provide the first dimension-free, self-normalized *empirical* Bernstein inequality. Our techniques are based on the variational (PAC-Bayes) approach to concentration.

## Contents

|   |   |    |
|---|---|----|
| 1 | Introduction                                  | 1  |
| 2 | Preliminaries                                 | 5  |
| 3 | Warmup: Sub-Gaussian Processes                | 8  |
| 4 | General Sub- $\psi$ Processes                 | 9  |
| 5 | Bernstein and Bennett Inequalities            | 15 |
| 6 | An Empirical Bernstein Inequality             | 16 |
| 7 | Summary                                       | 18 |
| A | Method of Mixtures for sub-Gaussian Processes | 23 |
| B | Omitted Proofs                                | 24 |
| C | Simulation Details                            | 36 |

## 1 Introduction

The modern theory of finite-sample, self-normalized concentration originates primarily in the work of de la Peña, Klass, Lai, Shao and co-authors in the 2000s [18, 15, 19, 17, 16]. Much of this work was focused on scalar-valued processes until de la Peña et al. [16] began studying the concentration of *vector-valued* self-normalized processes in Euclidean spaces. These results take the form of bounds on  $\|S_t\|_{V_t^{-1}}$  where  $(S_t)_{t \geq 0}$  is a stochastic process in

$\mathbb{R}^d$  and  $(V_t)_{t \geq 0}$ ,  $V_t \in \mathbb{R}^{d \times d}$ , is some associated positive-definite variance process. In the sub-Gaussian case, de la Peña et al. [18] provide bounds—much earlier than the oft-cited bound of Abbasi-Yadkori et al. [1]—depending on the log-determinant of  $V_t$  using the method of mixtures with a multivariate Gaussian mixing distribution. Under slightly more general tail assumptions on  $(S_t)$ , de la Peña et al. [16] provide bounds that  $\|S_t\|_{V_t^{-1}}$  lies in a convex set, though these are not computable in closed-form.

What about more general classes of processes? Recently, Whitehouse et al. [60] gave self-normalized concentration bounds for vector-valued sub- $\psi$  processes, a powerful and general abstraction first introduced by Howard et al. [33] which captures many distributions of interest, including sub-Poisson, sub-exponential, sub-gamma, among others. While comprehensive, the bounds of Whitehouse et al. [60] deviate from the aforementioned log-determinant bounds in two ways. First, they are dimension-dependent and second, they scale with the condition number of  $V_t$  instead of the determinant. Neither dependency is strictly better than the other; both can be tighter in different regimes. However, the discrepancy invites the following question:

*Do dimension-free, determinant-based bounds exist for general sub- $\psi$  processes?*

This paper answers in the affirmative, providing dimension-free, determinant-based self-normalized concentration inequalities for vector-valued sub- $\psi$  processes. To demonstrate the scope of our technique, in Section 3 we will recover the bound of Abbasi-Yadkori et al. [1] exactly and show that it holds without modification for general sub-Gaussian processes (i.e., outside of their particular bandit setting). We then turn our attention to general sub- $\psi$  processes. Here our bounds have a similar functional form as the sub-Gaussian bound but are modified by multiplicative and additive factors that depend on  $\psi$ . Nevertheless, they remain closed-form and easily computable.

Our techniques are based on the variational approach to concentration (sometimes called the PAC-Bayesian approach [10, 11]), which has been successfully leveraged in recent years to study the concentration of random vectors and matrices [63, 49, 48, 65, 14]. As we will explain below, the variational approach is comparable to a data-dependent method of mixtures, in which we choose the mixture distribution to be data-dependent but pay the price of this dependency with an  $f$ -divergence.

While closing the gap between determinant-based and condition number-based bounds is mathematically interesting, it is also practically relevant. Beyond their application in bandit problems [1, 2, 56, 12, 39, 59], self-normalized bounds are used in many areas, such as system identification [44], control of dynamic systems [38], estimation in autoregressive processes and diffusion-limited aggregation processes [6], sequential inference in time-series [53], and Markov decision-processes [62, 56]. Moreover, in many of these applications, matrices are often ill-conditioned due to numerical instability or anisotropy [31]. Determinant-based bounds which omit the dependence on condition numbers can thus improve results in these areas.

## 1.1 Related Work

The canonical example of a self-normalized process is the  $t$ -statistic, introduced by William Gosset (aka Student) in 1908 [54]. The behavior of the  $t$ -statistic can be understood by studying objects of the form  $\sum_{i \leq n} X_i / \sqrt{\sum_{i \leq n} X_i^2}$ , where  $X_1, \dots, X_n$  are iid scalar observations.

The asymptotic behavior of such “self-normalized” sums was the focus of a significant body of working beginning in the late 60s and early 70s [20, 40] and gaining significant momentum in the 90s [29, 28, 35, 26, 25, 52].

De la Peña and others then built off of this work to provide the foundations of a general theory of self-normalized concentration, best summarized in the 2008 book of de la Peña, Lai, and Shao [19]. In the scalar setting, subsequent work by Bercu and Touati [5, 6] extended these results to heavy-tailed increments, developing self-normalized bounds when the normalization process  $(V_t)$  is the predictable or total quadratic variation. In 2020, Howard et al. [33] further generalized this framework by the studying time-uniform concentration of sub- $\psi$  processes—processes characterized by specific tail conditions on the cumulant generating functions—thereby providing a general, unifying framework for both self-normalized and non-self-normalized bounds.

Beyond the scalar setting, de la Peña et al. [16] (see also de la Peña et al. [19, Chapter 14]) considered the self-normalized concentration of vector-valued processes, which take the form of bounds on  $\|S_t\|_{V_t^{-1}}$  for a vector-valued process  $(S_t)$  and a matrix-valued process  $(V_t)$ . Hotelling’s  $T^2$ -statistic [32]—the generalization of the Student t-statistic to the multivariate setting—is an example of such a process. De la Peña [16] give bounds on the probability that  $\|S_t\|_{V_t^{-1}}$  belongs to a convex set, when  $S$  and  $V$  obey a “generalized canonical” assumption (see [19, Equation (14.5)]). This assumption is similar to, but weaker than, the sub- $\psi$  assumption of Howard et al. [33] (extended to  $\mathbb{R}^d$ ) that we employ (Definition 2.1). Unfortunately, outside of the sub-Gaussian case, explicit expressions for these sets appear unavailable.

In the contextual bandit setting with sub-Gaussian noise, Abbasi-Yadkori et al. [1] apply results of de la Peña et al. [18] and give what is now a famous self-normalized inequality. They show that, if  $S_t = \sum_{k \leq t} X_k \epsilon_k$  and  $V_t = U_0 + \sum_{k \leq t} X_k X_k^\top$  where  $\epsilon_k$  is scalar sub-Gaussian and  $X_k \in \mathbb{R}^d$  is predictable (i.e., measurable at time  $k - 1$ ) then  $\|S_\tau\|_{V_\tau^{-1}} \lesssim \sqrt{\log(\det V_\tau)}$  at all stopping times  $\tau$ . This bound was extended to infinite-dimensional spaces by Chowdhury and Gopalan [12].

Recently, Whitehouse et al. [60] extended the sub- $\psi$  condition of Howard et al. [33] to  $\mathbb{R}^d$  and showed that, if  $(S_t)$  and  $(V_t)$  obey such a condition, then

$$\|S_\tau\|_{V_\tau^{-1}} \lesssim \gamma_{\min}^{1/2}(V_\tau)(\psi^*)^{-1}(\gamma_{\min}^{-1}(V_\tau)(\log \log(\gamma_{\max}(V_\tau)) + d \log(\kappa(V_\tau))), \quad (1.1)$$

where  $\kappa(V_t) = \gamma_{\max}(V_t)/\gamma_{\min}(V_t)$  is the ratio of the largest to smallest eigenvalue of  $V_t$  (the condition number) and  $\psi^*$  is the convex conjugate of  $\psi$ . In the sub-Gaussian case we have  $(\psi^*)^{-1}(x) \asymp \sqrt{x}$ , so the bound of Whitehouse et al. [60] scales as

$$\|S_\tau\|_{V_\tau^{-1}} \lesssim \sqrt{\log \log(\gamma_{\max}(V_\tau)) + d \log \kappa(V_t)}. \quad (1.2)$$

Neither  $\det V_\tau$  (the product of all eigenvalues) or  $\kappa(V_\tau)$  dominates the other, making the bound of Whitehouse et al. [60] and Abbasi-Yadkori et al. [1] incomparable in general. If the matrices are well-conditioned then the former may be tighter, depending on the dimension. If there are many small eigenvalues (e.g., perhaps  $V_t$  is the result of many rank  $k$  updates for  $k \ll d$ ) then the determinant-based bound may be tighter. In any case, the gap between condition number-based bounds and determinant-based bounds is what motivates this work.

Very recently, both Akhavan et al. [3] Metelli et al. [46] gave dimension-free, self-normalized Bernstein inequalities, building off the earlier work of Faury et al. [22] and Zhou et al. [64],

|                            | <i>Condition</i> | <i>Dimension-free</i> | <i>Dependence</i> |
|----------------------------|------------------|-----------------------|-------------------|
| Faury et al. [22]          | Bounded          |                       | $\det V_t$        |
| Zhou et al. [64]           | Bounded          |                       | UB on $\ S_t\ $   |
| Ziemann [65]               | Bounded          |                       | $\det V_t$        |
| Akhavan et al. [3]**       | Bounded          | ✓                     | $\det V_t$        |
| Metelli et al. [46]**      | Bounded          | ✓                     | $\det V_t$        |
| Kirschner et al. [36]      | Bounded          |                       | $\det V_t$        |
| Kirschner et al. [36]      | Gaussian         | ✓                     | $\det V_t$        |
| de la Peña et al. [18]     | sub-Gaussian     | ✓                     | $\det V_t$        |
| Abbasi-Yadkori et al. [1]  | sub-Gaussian     | ✓                     | $\det V_t$        |
| Chowdhury and Gopalan [12] | sub-Gaussian     | ✓                     | $\det V_t$        |
| de la Peña et al. [19]*    | Gen. canonical   | ✓                     | $\det V_t$        |
| Whitehouse et al. [60]     | sub- $\psi$      |                       | $\kappa(V_t)$     |
| <b>This work</b>           | sub- $\psi$      | ✓                     | $\det V_t$        |

Table 1: A brief overview of previous vector-valued self-normalized concentration inequalities. “Condition” refers to the distributional assumptions made on the underlying process. “UB on  $\|S_t\|$ ” means the bound is a function of the upper bound of the norm of the observations (i.e., it is not adaptive to  $V_t$ ). We note that Kirschner et al. [36] provide two bounds: one under a Gaussian assumption and one under a bounded assumption. “Gen. canonical” refers to the generalized canonical assumption of de la Peña et al. [19, Section 14.1.2]. (\*) indicates that the results are not in closed-form. (\*\*) indicates that these works were concurrent to this one.

both of which provide dimension-dependent Bernstein-type bounds. Akhavan et al. [3] study the case in which the normalization process ( $V_t$ ) is the quadratic variation of ( $S_t$ ) and Metelli et al. [46] study the bandit setting (as do Faury et al. [22] and Zhou et al. [64]). All of these results apply to bounded processes and do not extend to general sub- $\psi$  processes. They are closest to the Bennett and Bernstein results we give in Section 5, though we note that our Bernstein bound applies to random vectors satisfying a more general condition than boundedness. See Table 1 for an overview of known self-normalized concentration results in the multivariate setting.

Finally, let us discuss previous work on the variational approach to concentration, which undergirds our work here. This technique originated in PAC-Bayesian approach to statistical learning theory as a method to bound the risk of randomized predictors. As far as we are aware, the first to employ the variational approach explicitly for the purposes of multivariate concentration was Oliveira [49] in 2016, who was interested in bounding the tails of quadratic forms. Since then, a number of authors have employed it to various problems.

In 2018, Giulini [27] used the variational approach to estimate the Gram operator in Hilbert spaces. Around the same time, Catoni and Giulini [10, 11] used it to estimate the mean of random vectors and the operator norm of random matrices in finite-dimensional Euclidean spaces. More recently, both Nakakita et al. [48] and Zhivotovskiy [63] return to the question of matrix concentration, each giving dimension-free bounds on sums of random matrices under various conditions. Chugg et al. [14] study the concentration of random vectors using the variational method, giving bounds under a variety of distributional assumptions

(including sub- $\psi$  condition which we study here).

All of the above works using the variational approach are in non-self-normalized settings. By contrast, Kirschner et al. [36] and Ziemann [65] recently noted that variational ideas can be used to obtain bounds similar to (or identical to) the ones presented by Abbasi-Yadkori et al. [1]. Ziemann [65] also gives a (dimension-dependent) Bernstein inequality for bounded observations in a contextual bandit setting. Overall, one can view our work here as extending the ideas of Kirschner et al. [36] and Ziemann [65] to more general stochastic processes.

## 1.2 Contributions and outline

The overarching goal of this work is to provide self-normalized bounds for sub- $\psi$  processes which depend on  $\log \det V_\tau$  instead of  $d \log \kappa(V_\tau)$ . After defining sub- $\psi$  processes and giving an overview of the variational approach in Section 2, Section 3 provides a gentle introduction to our method, showing that we recover the bound of Abbasi-Yadkori et al. [1] exactly when the process is sub-Gaussian.

Section 4 then provides concentration results for general sub- $\psi$  processes evolving in  $\mathbb{R}^d$ . In particular, Theorem 4.1 provides a dimension-free, self-normalized “line-crossing inequality,” which gives the probability that  $\|S_t\|_{V_t^{-1}}$  ever crosses a threshold which depends on a parameter  $\lambda > 0$ . Different values of  $\lambda$  optimize the threshold at different (intrinsic) times.

Theorems 4.3 and 4.7 employ a method known as *stitching* [34] which iteratively applies Theorem 4.1 with different parameters in order to give a bound which remains tight at all times. This procedure costs only an iterated logarithm term. Theorem 4.3 applies this technique to sub-gamma processes specifically and obtains precise constants, while Theorem 4.7 applies to more general processes but is asymptotic only.

Section 5 applies our results to obtain self-normalized Bennett and Bernstein inequalities, and Section 6 provides a self-normalized, *empirical* Bernstein inequality for bounded random vectors. As far as we are aware, this is the first result of its kind that is dimension-free. Section 7 concludes.

## 2 Preliminaries

Fix a filtered probability space  $(\Omega, \mathcal{A}, \mathcal{F} \equiv (\mathcal{F}_t)_{t \geq 0}, P)$ . We assume that we are working in discrete time, so  $\mathcal{F}$  is indexed by  $\mathbb{N} \cup \{0\}$ .<sup>1</sup> We consider two stochastic processes  $(S_t)_{t \geq 1}$  and  $(V_t)_{t \geq 0}$  on  $(\Omega, \mathcal{A}, \mathcal{F}, P)$  which are adapted to  $\mathcal{F}$  (i.e.,  $S_t$  and  $V_t$  are  $\mathcal{F}_t$ -measurable). Throughout this work we assume that  $S_t \in \mathbb{R}^d$  and  $V_t \in \mathbb{R}^{d \times d}$  for some finite positive integer  $d$ . We assume that  $V_t$  is positive-definite for each  $t \geq 0$ , meaning that  $\langle \theta, V_t \theta \rangle \geq 0$  for all  $\theta \in \mathbb{R}^d$ .

We are interested obtaining bounds on  $\|S_\tau\|_{V_\tau^{-1}}$  for any  $\mathcal{F}$ -adapted stopping time  $\tau$  under various assumptions on the relationship between  $S_t$  and  $V_t$ . More specifically, we think of  $(V_t)$  as the accumulated variance process of  $(S_t)$  or, in the words of Blackwell and Freedman [7], a measure of the intrinsic time-scale of  $(S_t)$ . This is formalized with the notion of a sub- $\psi$  process, which generalizes many distributions of interest. Let  $\mathbb{S}^{d-1} := \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$  denote the unit sphere in  $\mathbb{R}^d$ . Recall that a process  $(A_t)$  is a supermartingale with respect to

---

<sup>1</sup>Though we expect that our techniques are general enough to extend to continuous time as well.

$\mathcal{F}$  if it is  $\mathcal{F}$ -adapted and  $\mathbb{E}[A_t|\mathcal{F}_{t-1}] \leq A_{t-1}$  for all  $t \geq 1$ .<sup>2</sup>

**Definition 2.1** (Sub- $\psi$  process in  $\mathbb{R}^d$ ). Let  $\psi : [0, \lambda_{\max}) \rightarrow \mathbb{R}_{\geq 0}$ . Let  $(S_t)$  be an  $\mathbb{R}^d$ -valued process and  $(V_t)$ ,  $V_t \succ 0$  an  $\mathbb{R}^{d \times d}$ -valued process, both adapted to a filtration  $\mathcal{F}$ . We call  $(S_t)$  a sub- $\psi$  process with variance proxy  $(V_t)$  if, for all  $\theta \in \mathbb{S}^{d-1}$ , all  $\lambda \in [0, \lambda_{\max})$ , and all  $t \geq 0$ ,

$$\exp \{ \lambda \langle \theta, S_t \rangle - \psi(\lambda) \langle \theta, V_t \theta \rangle \} \leq L_t^\lambda(\theta), \quad (2.1)$$

where  $(L_t(\theta))_{t \geq 0}$  is a nonnegative supermartingale adapted to  $(\mathcal{F}_t)$ . Equivalently, we say that  $(S_t, V_t)$  is a sub- $\psi$  process.

For scalar  $S_t$  and  $V_t$ , sub- $\psi$  processes were originally proposed and studied by Howard et al. [33], building off of earlier work of Freedman [24] and de la Peña et al. [18]. The definition of a sub- $\psi$  process was then generalized to  $\mathbb{R}^d$  by Whitehouse et al. [60]. Like both Howard et al. [33] and Whitehouse et al. [60], we assume that  $\psi$  is *CGF-like*, meaning that it is strictly convex and twice continuously differentiable, with  $\psi(0) = \lim_{\lambda \rightarrow 0^+} d\psi(\lambda)/d\lambda = 0$ .

While the definition of a sub- $\psi$  process may appear abstract at first glance, let us reassure the reader that the definition captures many familiar examples.

1.  $\psi_N = \psi_N(\lambda) := \lambda^2/2$  for  $\lambda \in [0, \lambda_{\max})$  results in a sub-exponential process distribution (sub-Gaussian if  $\lambda_{\max} = \infty$ ). If  $X_t$  is  $\Sigma_t$ -sub-Gaussian<sup>3</sup> then  $S_t = \sum_{k \leq t} X_k$  is a sub- $\psi_N$  process with  $\lambda_{\max} = \infty$  and  $V_t = \sum_{k \leq t} \Sigma_k$ . More exotically, consider any sequence of random vectors  $(X_t)$  that are conditionally symmetric:  $X_t \sim -X_t | \mathcal{F}_{t-1}$  for all  $t$ . Lemma 3 of de la Peña et al. [15] implies that  $S_t = \sum_{k \leq t} X_k$  and  $V_t = \sum_{k \leq t} X_k X_k^\top$  define a sub- $\psi_N$  process with  $\lambda_{\max} = \infty$ . As in the scalar case, bounded random vectors can also be shown to be sub- $\psi_N$ .
2.  $\psi_{G,c}(\lambda) = \frac{\lambda^2}{2(1-c\lambda)}$  for  $c \in \mathbb{R}$  and  $\lambda_{\max} = 1/\max\{c, 0\}$  results in a sub-gamma process.<sup>4</sup> As in the scalar case, random vectors whose moments in each direction obey a Bernstein-type inequality can be shown to be sub- $\psi_{G,c}$ ; see Lemma 5.2. Whitehouse et al. [60] further show that if  $\mathbb{E}|\theta, X_t|^3$  is finite for all  $\theta \in \mathbb{S}^{d-1}$ , then  $S_t = \sum_{j \leq t} X_j$  is sub- $\psi_{G,c}$  for  $c = 1/6$  and  $V_t = \sum_{j \leq t} (X_j X_j^\top + \mathbb{E}[\|X_j\|^3 | \mathcal{F}_{j-1}] I_d)$ . Moreover, Howard et al. [33, Proposition 1] show that any twice differentiable, CFG-like  $\psi$  can be bounded by  $a\psi_{G,c}$  for some  $a, c > 0$ . Thus, any sub- $\psi$  process is sub- $\psi_{G,c}$  after scaling.
3.  $\psi_{E,c}(\lambda) = (-\log(1 - c\lambda) - c\lambda)/c^2$  for  $c \in \mathbb{R}$  where  $\lambda_{\max} = 1/\max\{c, 0\}$  results in a sub-negative-exponential process.<sup>5</sup> Bounded random vectors can be shown to be sub-exponential with a predictable variance proxy; see Section 6. This observation undergirds our empirical Bernstein bound. As in the sub-gamma case, Howard et al. [33, Proposition 1] also shows that any CGF-like  $\psi$  can be upper bounded by  $a\psi_{E,c}$  for some  $a, c \geq 0$ .

<sup>2</sup>One would typically specify the distribution  $P$  under which  $(A_t)$  is a supermartingale, and refer to  $(A_t)$  as a  $P$ -supermartingale. However, in our case we consider only one probability measure  $P$ .

<sup>3</sup>That is,  $\mathbb{E}[\exp(\langle \theta, X_t \rangle) | \mathcal{F}_{t-1}] \leq \exp(\langle \theta, \Sigma_t \theta \rangle)$  for all  $\theta \in \mathbb{S}^{d-1}$ .

<sup>4</sup> $\psi_{G,c}$  is an *upper bound* on the CGF of a gamma random variable [9, Section 2.4]

<sup>5</sup> $\psi_{E,c}$  is the CGF of a centered unit-rate negative-exponential random variable. We note that many authors call  $\psi_{E,c}$  the CGF of an exponential distribution [33, 34, 60], but we want to reserve the term sub-exponential for  $\psi_N$  with finite  $\lambda_{\max}$ . Hence we use the term negative-exponential instead.



4.  $\psi_{P,c}(\lambda) = (e^{c\lambda} - c\lambda - 1)/c^2$  for  $c \in \mathbb{R}$  and  $\lambda_{\max} = \infty$  results in a sub-Poisson process.<sup>6</sup> Most proofs of Bennett’s inequality (e.g., Boucheron et al. [9, Theorem 2.9]) involve showing that a bounded random variable is sub-Poisson. In the multivariate setting, Lemma 5.1 shows that vectors  $(X_t)$  with  $\|X_t\| \leq b$  are sub- $\psi_{P,b}$  with  $V_t = \sum_{j \leq t} \mathbb{E}[X_j X_j^\top | \mathcal{F}_{j-1}]$

For any  $\psi$  such that  $\psi(\lambda)/\lambda^2$  is nondecreasing in  $\lambda$  (a property known as super-Gaussianity which all of the above examples satisfy; see also Remark 4.2), we can lift any scalar sub- $\psi$  process to a multivariate sub- $\psi$  process as follows. Consider  $S_t = \sum_{k \leq t} \eta_k X_k$  where  $(X_t)_{t \geq 1}$  is predictable,  $\|X_t\| \leq 1$ , and  $(\eta_t)$  is a scalar-valued sub- $\psi$  process with variance proxy  $(U_t) \subseteq \mathbb{R}_{\geq 0}$ . Then  $(S_t)$  is sub- $\psi$  with variance proxy  $V_t = \sum_{k \leq t} U_k X_k X_k^\top$ . This is the contextual bandit setting described in Section 1.1.

Let us now turn to the techniques we use to prove our results. The bounds of de la Peña et al. [19] and Abbasi-Yadkori et al. [1] are proved using the so-called “method of mixtures” (sometimes also called the pseudo-maximization technique [15]). At its core, the method of mixtures involves (i) finding a family of supermartingales  $Z(\theta)$  indexed by some parameter space  $\Theta$ , (ii) constructing a new supermartingale by integrating this family with respect to some measure  $\rho$  over  $\Theta$ , say  $Z_t = \int_{\Theta} Z_t(\theta) d\rho$ , and (iii) applying Markov’s inequality (or, more precisely, Ville’s inequality) to  $Z_t$ . For instance, one might let  $Z_t(\theta)$  be the left hand side of (2.1) and  $\rho$  be a measure over  $\Theta = \mathbb{S}^{d-1}$ .

To prove our results, we instead turn to a related but distinct technique which we call the *variational approach to concentration*. The variational approach also begins with a mixture but after step (2) applies the Donsker-Varadhan change of measure formula in order to bound  $Z_t$  in terms of the KL divergence between  $\rho$  and some prior distribution. The benefit of this technique is that  $\rho$  can be data-dependent. This approach lies at the heart of (and originated in) PAC-Bayesian learning theory [30, 4]. A generic template for such a bound, which applies to general stochastic processes and handles stopping times, was given by Chugg et al. [13, Theorem 4]. See also Flynn et al. [23] for a similar statement.

**Proposition 2.2** (Variational Template). *Let  $\Theta$  be a measurable parameter space. For each  $\theta \in \Theta$ , let  $Z(\theta) \equiv (Z_t(\theta))_{t \geq 0}$  be a stochastic process upper bounded by a nonnegative supermartingale  $Y(\theta) \equiv (Y_t(\theta))_{t \geq 0}$ . Assume that all processes are adapted to the same filtration  $\mathcal{F}$  and that  $Y_0(\theta) \leq 1$  for each  $\theta \in \Theta$ . Let  $\nu$  be a data-free distribution over  $\Theta$ . Then, for all  $\delta \in (0, 1)$ , with probability  $1 - \delta$ , for any  $\mathcal{F}$ -adapted stopping time  $\tau$  and any  $\mathcal{F}_\tau$ -measurable distribution  $\rho_\tau$  over  $\Theta$ ,*

$$\int \log Z_\tau(\theta) \rho_\tau(d\theta) \leq D_{\text{KL}}(\rho_\tau \| \nu) + \log(1/\delta). \quad (2.2)$$

Unlike approaches to multivariate concentration based on covering arguments or chaining, the variational approach can lead to dimension-free bounds if one is careful in their choice of  $\rho_\tau$  and  $\nu$ . In our case the KL-divergence will typically act as  $\log \det(V_\tau)$ , which is how we obtain dimension-free, determinant-based bounds.

**Notation.** We denote the eigenvalues of a matrix  $M$  as  $\gamma_{\min}(M) = \gamma_1(M) \leq \dots \leq \gamma_d(M) = \gamma_{\max}(M)$ . We let  $\|M\|_{\text{op}} = \gamma_{\max}(M)$  be the operator norm of  $M$ . For a convex function

<sup>6</sup> $\psi_{P,c}$  is the CGF of a centered unit-rate Poisson random variable.

$f : I \rightarrow \mathbb{R}$ ,  $I \subseteq \mathbb{R}$ , we denote its convex conjugate by  $f^*(y) = \sup_{x \in I} (xy - f(x))$ , which is also a convex function. The convex conjugate is also called the Legendre–Fenchel transform. We let  $\mathbb{N}_0 = \{0, 1, 2, \dots\}$  denote the natural numbers including zero and  $\mathbb{R}_{>0}$  denote the positive real numbers. We use  $A \succeq B$  to signify that  $A$  is greater than  $B$  in the Loewner order, i.e.,  $A - B$  is positive semidefinite. For a vector  $v$  and positive semidefinite matrix  $A$ ,  $\|v\|_A$  denotes the norm defined as  $\|v\|_A^2 = \langle v, Av \rangle$ , where  $\langle \cdot, \cdot \rangle$  is the usual dot product in  $\mathbb{R}^d$  unless otherwise specified. We let  $a \vee b = \max\{a, b\}$  and use  $A^\top$  to refer to the transpose of the matrix  $A$ . Throughout, the indices on the sum  $\sum_{j \leq t}$  should be assumed to run from 1 to  $t$ .

### 3 Warmup: Sub-Gaussian Processes

Let us begin by considering the special case of sub-Gaussian processes<sup>7</sup>. That is, we assume that  $\psi(\lambda) = \psi_N(\lambda) = \lambda^2/2$ . This allows for a particularly clean application of the variational template, since we may take  $\Theta = \mathbb{R}^d$  as our parameter space. (For  $\psi = \psi_N$ , one can replace  $\mathbb{S}^{d-1}$  with  $\mathbb{R}^d$ .) This in turn enables the use of normal distributions in Proposition 2.2.

We emphasize that the following result is only a very modest generalization of known bounds. More specifically, it is the same bound of Abbasi-Yadkori et al. [1], though in a slightly more general setting. In fact, the techniques of Abbasi-Yadkori et al. (i.e., the method of mixtures) can also be used to obtain the following bound. This is done in Appendix A.

The result is of interest not for its novelty, but because it demonstrates that the variational approach can recover the same bound with the same constants as the method of mixtures. Ziemann [65] recently made a similar observation, though he focuses on the bandit setting in particular. The proof can be found in Appendix B.1.

**Theorem 3.1.** *Let  $(S_t, V_t)$  be a sub-Gaussian process. Let  $U_0$  be a fixed positive-definite matrix. Then, with probability  $1 - \delta$ , for any stopping time  $\tau$ ,*

$$\|S_\tau\|_{(V_\tau + U_0)^{-1}}^2 \leq \log \left( \frac{\det(V_\tau + U_0)}{\det U_0} \right) + 2 \log(1/\delta). \quad (3.1)$$

Theorem 3.1 recovers the bound of Abbasi-Yadkori et al. [1] if  $(X_t)_{t \geq 1} \subseteq \mathbb{R}^d$  is an  $\mathcal{F}$ -predictable process ( $X_t$  is the  $t$ -th action taken by a bandit algorithm in their case) and  $S_t = \sum_{k=1}^t \eta_k X_k$  where  $\eta_k$  is scalar  $\sigma$ -sub-Gaussian noise.<sup>8</sup> Then  $(S_t)$  is sub- $\psi_N$  with variance proxy  $(V_t)$  defined as  $V_t = \sigma^2 \sum_{k=1}^t X_k X_k^\top$ . Taking  $U_0 = \sigma^2 \rho V_0$  for some  $V_0$  and rearranging (3.1) gives Theorem 1 of Abbasi-Yadkori et al. [1]. More generally, Theorem 3.1 applies to any sub-Gaussian process described in Section 2, such as  $S_t = \sum_{k \leq t} X_k$  and  $V_t = \sum_{k \leq t} X_k X_k^\top$  when  $X_t$  is conditionally symmetric.

While Theorem 4.1 is stated in terms of stopping times, it can equivalently be stated as a time-uniform bound of the form:  $P(\forall t \geq 1 : (3.1) \text{ holds}) \geq 1 - \delta$ . In this paper we state our bounds in terms of stopping times to conform to previous work in this area.

<sup>7</sup>It’s perhaps worth mentioning “sub-Gaussian process” has a double meaning in the literature. In this paper, a sub-Gaussian process refers to Definition 2.1 with  $\psi = \psi_N$ . This is different than the sub-Gaussian process studied by Talagrand [55], which is a collection of zero-mean random variable  $\{Y_\theta : \theta \in \Theta\}$  such that  $\mathbb{E}[\exp(\lambda(X_\theta - X_\phi))] \leq \exp(\psi_N(\lambda)\rho^2(\theta, \phi))$  for some metric  $\rho$  and all  $\theta, \phi \in \Theta$ .

<sup>8</sup>That is,  $\mathbb{E}[\exp(\lambda \eta_k) | \mathcal{F}_{k-1}] \leq \exp(\lambda^2 \sigma^2 / 2)$



## 4 General Sub- $\psi$ Processes

Theorem 3.1 relied on transforming the sub- $\psi$  condition from one which holds over  $\mathbb{S}^{d-1}$  to one which holds over  $\mathbb{R}^d$ , thus enabling us to take  $\Theta = \mathbb{R}^d$  and use normal distributions in Proposition 2.2. This trick cannot be performed for general sub- $\psi$  processes because  $\lambda_{\max}$  may be finite. However, for general  $\psi$  functions we may replace  $\mathbb{S}^{d-1}$  with  $\mathbb{B}^d$ , the unit ball in  $\mathbb{R}^d$ .<sup>9</sup> This allows us to use nested uniform distributions over  $\Theta = \mathbb{B}^d$  in Proposition 2.2.

The machinery we use to handle general sub- $\psi$  processes does have drawbacks over what was used in Section 3 to study the specific case of sub-Gaussian processes. For one, when we apply the results in this section to sub-Gaussian processes, the resulting bound is looser than Theorem 3.1 (see Figure 1). Second, the parameter  $\lambda$  of the sub-Gaussian process in Theorem 3.1 can be optimized independently of other terms in the bound. Such closed-form optimization of  $\lambda$  is not possible in the more general case. We instead end up with *line-crossing inequalities*: For any given  $\lambda$  we obtain a bound on the probability that  $\|S_t\|_{V_t^{-1}}$  ever crosses a threshold parameterized by  $\lambda$ . Different values of  $\lambda$  result in a bound which is tighter at different values of  $\det V_t$ . Theorem 4.1 gives the line-crossing inequality for general sub- $\psi$  processes.

Line-crossing inequalities are common in works studying the concentration of sub- $\psi$  processes [33, 34]. To go from line-crossing inequalities to bounds which are tight at all (intrinsic) times, we deploy a technique which has come to be known as “stitching,” though see Section 4.1 for more background. We provide two kinds of stitched bounds: One for general sub- $\psi$  processes, and one specifically tailored for sub-Gamma processes.

We must define two quantities before stating our first result. First, for a positive semidefinite matrices  $M_1$  and  $M_2$ , define their *Rayleigh-Ritz maximum*<sup>10</sup>

$$\alpha(M_1, M_2) \equiv \sup_{\theta \in \mathbb{R}^d} \frac{\langle \theta, M_1 \theta \rangle}{\langle \theta, M_2 \theta \rangle}. \quad (4.1)$$

When using the Rayleigh-Ritz maximum hereafter, we will always consider matrices  $M_1, M_2$  such that  $M_2 \succeq M_1$ , thus implying that  $0 \leq \alpha(M_1, M_2) \leq 1$ . Next, given a sub- $\psi$  process  $(S_t, V_t)$  and a matrix  $U_0$  such that  $V_t \succeq U_0$  for all  $t \geq 1$ , let

$$\alpha_t = \alpha(U_0, V_t), \quad (4.2)$$

and define the sequence of functions  $(g_{\psi,t} : \text{Im}(\psi) \rightarrow [0, 1])_{t \geq 1}$  by

$$g_{\psi,t}(z) = \sqrt{\alpha_t \psi^{-1}(z) + 1 - \alpha_t} - \sqrt{\alpha_t \psi^{-1}(z)}. \quad (4.3)$$

These functions may appear unwieldy at first glance but they behave relatively simply. Indeed, as  $\alpha_t \rightarrow 0$  (which is the case if  $\gamma_{\min}(V_t) \rightarrow \infty$ ),  $g_{\psi,t}(z) \rightarrow 1$  with a rate dictated by  $\psi^{-1}$ . For most common sub- $\psi$  processes and matrices  $V_t$  which have at least some spread in all directions,  $g_{\psi,t}(z)$  is above 0.5 after 100 samples for all  $\lambda \in \text{Im}(\psi)$ . See Figure 1.

We now present the line-crossing inequality for sub- $\psi$  processes. The proof is in Appendix B.2.

<sup>9</sup>More specifically, we can upper bound any  $\psi$  with some  $\hat{\psi}$  that allows us to replace  $\mathbb{S}^{d-1}$  with  $\mathbb{B}^d$ . Such  $\hat{\psi}$  are called *super-Gaussian*. See Remark 4.2 for further discussion.

<sup>10</sup>This is also the maximum over the generalized Rayleigh coefficient between  $M_1$  and  $M_2$ , and the solution to the generalized eigenvalue problem  $M_1 x = \lambda M_2 x$ .

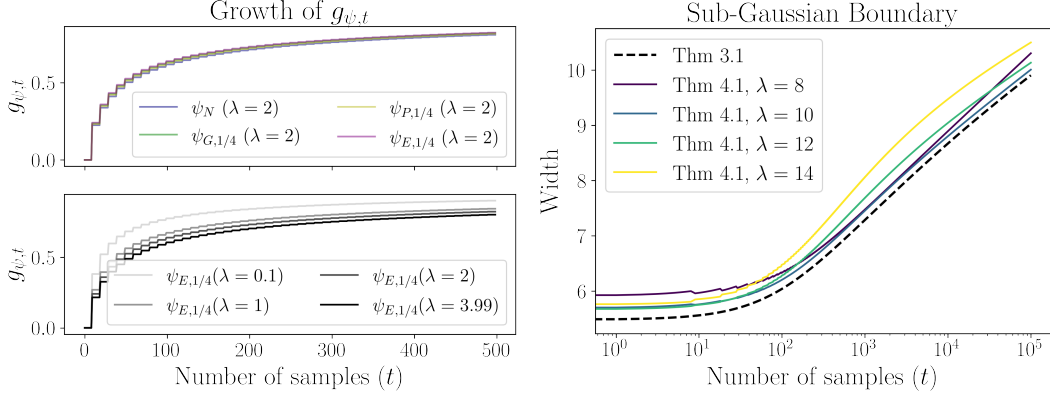


Figure 1: *Left*: The growth of  $g_t$  for various  $\psi$  functions and across various values of  $\lambda$ . For sub-gamma, sub-Poisson, and sub-exponential we fix  $c = 1/4$ . *Right*: A comparison of Theorem 3.1 and Theorem 4.1 instantiated for  $\psi = \psi_N$ . In both figures we use  $U_0 = I_d$  and  $V_t = U_0 + \sum_{k \leq t} X_s X_s^\top$  where the vectors  $X_t$  are chosen based on a bandit algorithm. Details may be found in Appendix C.

**Theorem 4.1.** *Let  $(S_t, V_t)$  be a sub- $\psi$  process such that  $\lambda \mapsto \psi(\lambda)/\lambda^2$  is nondecreasing. Let  $U_0$  be a positive definite matrix such that  $V_t \succeq U_0$  for all  $t \geq 1$ . Let  $\alpha_t = \alpha_t(U_0, V_t)$  be the Rayleigh-Ritz maximum between  $U_0$  and  $V_t$ . Then, for any  $\delta \in (0, 1)$  and any  $\lambda \in [\psi^{-1}(1/\gamma_{\min}(U_0)), \lambda_{\max}) \cap \text{Im}(\psi)$ , with probability  $1 - \delta$ , for any stopping time  $\tau$ ,*

$$\|S_\tau\|_{V_\tau^{-1}} \leq \frac{\|U_0\|_{\text{op}}^{1/2}}{\lambda g_{\psi,\tau}(\lambda)} \left( \frac{1}{2} \log \left( \frac{\det V_\tau}{\det U_0} \right) + 1 + \log(1/\delta) \right) + \frac{g_{\psi,\tau}(\lambda) \psi(\lambda)}{\lambda \|U_0\|_{\text{op}}^{1/2}}. \quad (4.4)$$

**Remark 4.2.** Theorem 4.1 assumes that  $\psi(\lambda)/\lambda^2$  is a nondecreasing function, a condition that is sometimes referred to as  $\psi$  being *super-Gaussian*. This assumption is without loss of generality in the following sense: As we detailed in Section 2, any  $\psi$  function can be upper bounded by a constant times  $\psi_{G,c}$ , which is super-Gaussian. (In fact, all sub- $\psi$  processes mentioned in Section 2 are super-Gaussian). Thus, Theorem 4.1 may be applied to any sub- $\psi$  process, at the cost of some looseness in constants if  $\psi$  is strictly sub-Gaussian (e.g., sub-Bernoulli).

Several remarks are in order. First, the requirement that  $\lambda \geq \psi^{-1}(1/\gamma_{\min}(U_0))$  comes from our proof technique. More specifically, we use a prior which is uniform over the ellipsoid  $\{\theta : \psi(\lambda)\theta^\top U_0 \theta \leq 1\}$ , which we require to be a subset of  $\mathbb{B}^d$ . This holds if  $\psi(\lambda)U_0 \succeq I_d$ , i.e.,  $\lambda \geq \psi^{-1}(\gamma_{\min}^{-1}(U_0))$ . The requirement that  $\lambda \leq \lambda_{\max}$  comes from Definition 2.1, and the requirement that  $\lambda \in \text{Im}(\psi)$  comes from the definition of  $g_{\psi,t}$ . Note, however, that for common  $\psi$  functions including  $\psi_E$ ,  $\psi_P$ ,  $\psi_G$  and  $\psi_N$ ,  $\text{Im}(\psi) = \mathbb{R}_{\geq 0}$ , so requiring that  $\lambda \in \text{Im}(\psi)$  is often a vacuous constraint.

Second, Theorem 4.1 may appear as though it has the wrong order of magnitude. Indeed, Theorem 3.1 is a bound on  $\|S_\tau\|_{V_\tau^{-1}}^2$ , whereas (4.4) is a bound on  $\|S_\tau\|_{V_\tau^{-1}}$  (without the square). To set yourself at ease, let  $D_t(\delta) = \frac{1}{2} \log(\det V_t / \det U_0) + 1 + \log(1/\delta)$  and assume

for simplicity that  $\|U_0\| = 1$ . Then, in the sub-Gaussian case, (4.4) reads

$$\|S_\tau\|_{V_\tau^{-1}} \leq \frac{D_\tau(\delta)}{\lambda g_\tau(\lambda)} + \frac{\lambda g_\tau(\lambda)}{2}.$$

For a fixed  $t$ , if  $\lambda$  is such that  $\lambda g_t(\lambda) = \sqrt{2D_t(\delta)}$  (note that the map  $x \mapsto xg_{\psi_N,t}(x)$  is surjective on  $\mathbb{R}_{\geq 0}$ ), in which case we obtain  $\|S_t\|_{V_t^{-1}}^2 \leq D_t(\delta)$ . This matches Theorem 3.1 up to an additive factor of 2. However, this only demonstrates that there *exists* a  $\lambda$  which makes the widths roughly equal. We cannot in fact choose  $\lambda$  such that  $\lambda g_t(\lambda) = \sqrt{2D_t(\delta)}$  since  $\lambda$  cannot be data-dependent. Moreover, such a choice of  $\lambda$  would only ensure the widths are comparable at a particular fixed time  $t$ .

If  $\lambda$  is not optimized, how does Theorem 4.1 with  $\psi = \psi_N$  compare to Theorem 3.1? The right hand panel of Figure 1 plots the comparison for various values of  $\lambda$ . As one would expect, Theorem 4.1 is looser since it handles a wider class of distributions and is not explicitly optimized for sub-Gaussian processes. That said, Theorem 4.1 loses surprisingly little over Theorem 3.1, even for fixed values of  $\lambda$ . As we can see, one faces a tradeoff when choosing  $\lambda$ , with smaller values being tighter for larger  $t$  (though sometimes this tradeoff is minimal compared to the size of the bound; see Figure 2 for instance). Next we explore how to optimize  $\lambda$  over time.

## 4.1 Optimizing $\lambda$ via Stitching

In order to overcome the hurdle of optimizing  $\lambda$  in Theorem 4.1, will apply the bound of Theorem 4.1 iteratively over geometrically spaced epochs, choosing a different  $\lambda$  in each epoch. This broad technique goes by many names such as doubling, discrete mixtures, chaining, etc., but here we follow a particularly sharp variant called *stitching* by Howard et al. [34]. (The extra sharpness of stitching comes from employing tighter line crossing inequalities due to [33], which results in optimizing  $\lambda$  differently; see the above paper for references to other variants of this technique.) Stitching has since been applied in number of works on time-uniform bounds [13, 14], including Whitehouse et al. [60]. In this section we provide two stitching results: One for sub-gamma processes (Theorem 4.3) and one for more general processes (Theorem 4.7).

We begin with sub-gamma processes. Fortunately, the arithmetic is sufficiently navigable in this case so as to allow us to give a bound with small and precise constants. The result is Theorem 4.3 below. As we mentioned in Section 2, all sub- $\psi$  functions are sub-gamma, in the sense that for all twice-differentiable  $\psi : [0, \lambda_{\max}) \rightarrow \mathbb{R}$  with  $\psi(0) = \lim_{x \rightarrow 0+} \psi'(x) = 0$ , there exist constants  $a, c > 0$  such that  $\psi(\lambda) \leq a\psi_{G,c}(\lambda)$  [33, Proposition 1]. Sub-gamma bounds can thus be applied to all sub- $\psi$  processes after appropriate scaling.

As far as stitching is concerned, in the proof of Theorem 4.3 the epochs are defined relative to the intrinsic time  $\det V_t$ . More specifically, the  $k$ -th epoch is  $E_k = \{t : \eta^k \leq \det V_t < \eta^{k+1}\}$  for some parameter  $\eta > 1$ . The weight assigned to the  $k$ -th epoch is determined by a “stitching function”  $\ell(k)$ , which obeys  $\sum_k \ell^{-1}(k) \leq 1$ . In the result that follows we make the simplifying choices  $\eta = 2$  and  $\ell(k) = (k+1)^2 \zeta(2)$  where  $\zeta(2) = \pi^2/6$  is the Riemann-zeta function. A more general theorem which keeps all hyperparameters unspecified—Theorem B.1—from which this result follows can be found in Appendix B.3.

**Theorem 4.3.** Let  $(S_t, V_t)$  be a sub- $\psi_{G,c}$  process for any  $c > 0$  and suppose that  $V_t \succeq U_0 \succeq I_d$  for all  $t$  and some positive-definite  $U_0$  with  $\rho = \gamma_{\min}(U_0)$ . Fix  $\delta \in (0, 1/\sqrt{2}]$ . Then, with probability  $1 - \delta$ , for all stopping times  $\tau$ ,

$$\|S_\tau\|_{V_\tau^{-1}} \leq \frac{cD_\tau + 1.60\sqrt{D_\tau} + \max\left\{\frac{c+\sqrt{c^2+2\rho}}{2\rho}, \sqrt{\frac{D_\tau}{2}}\right\}}{H_{c,\tau}}, \quad (4.5)$$

where

$$D_\tau = \frac{1}{2} \log\left(\frac{\det V_\tau}{\det U_0}\right) + 1.5 + 2 \log(\log_2(\det V_\tau) + 1) + \log(1/\delta), \quad (4.6)$$

and

$$H_{c,\tau} = 0 \vee \left( \sqrt{1 - \frac{\gamma_{\max}(U_0)}{\gamma_{\min}(V_t)}} - \sqrt{\frac{\gamma_{\max}(U_0)}{\gamma_{\min}(V_t)} \frac{2}{(c + \sqrt{c^2 + 2c})}} \right). \quad (4.7)$$

If the maximum in (4.5) is achieved by  $\sqrt{D_t/2}$ , then we can replace the numerator with  $cD_\tau + 2.3\sqrt{D_\tau}$ . The denominator should be viewed as a penalty we pay for the appearance of  $g_{\psi,t}$  in Theorem 4.1. If  $\gamma_{\min}(V_t) \rightarrow \infty$  then  $H_{c,t} \rightarrow 1$ , just as  $g_{\psi,t}$ . We suspect that the appearance of  $H_{c,\tau}$  in (4.5) is sub-optimal, but it's unclear how to remove this term with our current proof techniques.

Theorem 4.3 assumes that  $V_t \succeq I_d$  to ensure that the union of the epochs  $E_k = \{t : \eta^k \leq \det V_t < \eta^{k+1}\}$ ,  $k \geq 0$  used in the stitching argument constitutes a partitioning of the sample space. If this assumption is not met, however, we can rescale the result as follows. If  $V_t \succeq \beta I_d$  for some  $\beta < 1$  and  $(S_t, V_t)$  is a sub- $\psi$  process then Whitehouse et al. [60, Proposition 2.4] shows that  $(S_t/\sqrt{\beta}, V_t/\beta)$  is a sub- $\psi_\beta$  process for  $\psi_\beta(\lambda) = \beta\psi(\lambda/\sqrt{\beta})$ . We can thus apply Theorem 4.3 to this rescaled process.

**Remark 4.4.** The appearance of both the logarithmic term  $\log \det V_\tau$  and the iterated logarithmic term  $\log \log \det V_\tau$  in Theorem 4.3 may seem strange. They have different origins: the latter is a result of stitching while the former comes from the variational proof technique and, in particular, is the KL-divergence between the prior and posterior in Proposition 2.2. In short, the term  $\log \det V_\tau$  captures the complexity of the process. It replaces  $d \log \kappa(V_\tau)$  in the bounds of Whitehouse et al. [60] and is comparable to the dependence on variance-like terms in non self-normalized bounds (see, e.g., [13, Theorem 2.3]).

**Remark 4.5.** Related to the previous remark, it's worth noting that the appearance of  $\log \det V_\tau$  has unfortunate consequences when the results are applied in  $d = 1$ . In this case,  $\log \kappa(V_\tau) = 0$  and the bounds of Whitehouse et al. [60] scale as  $\|S_\tau\|_2 \lesssim \sqrt{V_\tau}(\psi^*)^{-1}(\log \log(V_\tau) + \log(1/\delta))$ . In our case, however, we retain a  $\log(V_\tau)$  term, which is unusual for time-uniform bounds in the scalar setting (cf. [34]). In particular, one might expect the term  $\log \log(V_\tau)$  instead of  $\log(V_\tau)$ . In the multivariate setting, however, an iterated logarithm dependence on  $\det V_\tau$  is impossible, as shown by the following example.

**Example 4.6.** We borrow the conclusion of an elegant example from de la Peña et al. [15], also studied by Whitehouse et al. [60] (see their Example 4.6). In particular, de la Peña et al. [15] construct a sub-Gaussian process  $(S_t, V_t)$  in  $\mathbb{R}^2$  such that  $\|S_t\|_{V_t^{-1}} \sim \log(t)$ ,  $\gamma_{\max}(V_t) \sim t(1+a)$ , and  $\gamma_{\min}(V_t) \sim \log(t)/(1+)$  for a constant  $a$ . Hence  $\det V_t \sim t \log t$  implying that  $\|S_t\|_{V_t^{-1}} \sim \log(\det V_t)$ .

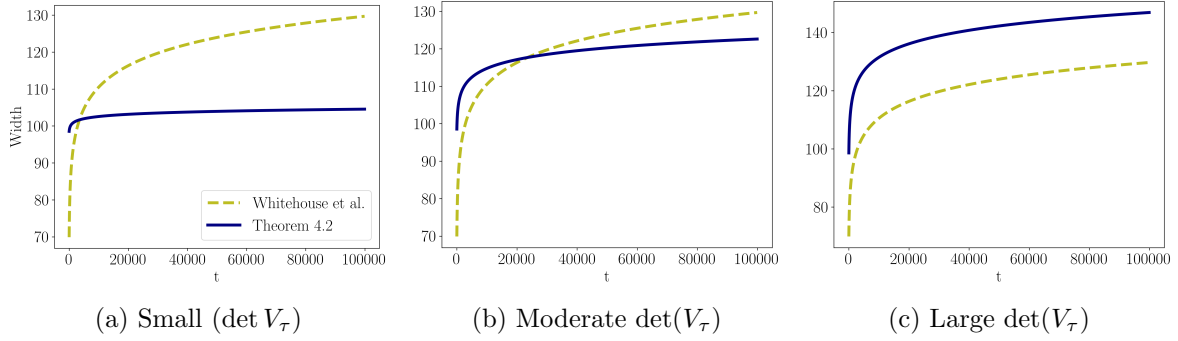


Figure 2: A comparison of Theorem 4.3 with the bound of Whitehouse et al. [60], which we recall is based on the condition number of  $V_\tau$ . We control the growth of the determinant with rank- $k$  updates: as  $k$  grows, so does  $\det V_\tau$ . The condition number has the same growth in each case. Theorem 4.3 outperforms the condition number-based bound for smaller values of  $\det V_\tau$  and loses ground as the determinant grows relative to  $d \log \kappa(V_\tau)$ . We use  $d = 20, c = 1$  and  $U_0 = I_d$ . Full simulation details can be found in Appendix C.

Figure 2 compares Theorem 4.3 to the sub-gamma bound presented by Whitehouse et al. [59]. We fix the growth rate of the condition number,  $\kappa(V_t)$ , and vary that of  $\det(V_t)$ . As expected, our bound outperforms the condition number-based bound of Whitehouse et al. [60] when the determinant grows slowly but is uniformly weaker when the determinant grows quickly.

Intriguingly, the bound of Whitehouse et al. [60] seems to always dominate our bound at small sample sizes (roughly  $< 200$ ). This is due to the factor of  $H_{c,t}$  in our bounds, which is small at low sample sizes (or, more specifically, until  $\gamma_{\min}(V_t)$  grows sufficiently large). This highlights a drawback of our bounds: if  $\gamma_{\min}(V_t)$  never grows, or grows extremely slowly, then  $H_{c,t}$  can remain extremely small, blowing up the bound and making it vacuous.

It's helpful to write Theorem 4.3 in terms of  $\psi_{G,c}^*$ , in which case (4.5) becomes

$$\|S_\tau\|_{V_\tau^{-1}} \lesssim \frac{(\psi_{G,c}^*)^{-1}(D_\tau)}{H_{c,\tau}}. \quad (4.8)$$

The appearance of  $(\psi_{G,c}^*)^{-1}$  is comforting: reliance on the inverse convex conjugate of  $\psi$  is consistent with past work on sub- $\psi$  concentration, such as Whitehouse et al. [60] (see (1.1) and (1.2) in Section 1.1) and Manole and Ramdas [41]. The reliance here also suggests that we may be able to obtain a bound for general  $\psi$  that relies on  $(\psi^*)^{-1}$ . This is the goal of Theorem 4.7 below.

Before we get there, however, let us compare Theorem 4.3 with the line-crossing inequality in Theorem 4.1. Figure 3 examines the sub- $\psi_{G,c}$  boundary both for fixed values of  $\lambda$  and when we stitch over  $\lambda$ . For reasonable sample sizes, some values of  $\lambda$  dominate the stitched boundary of Theorem 4.3 (dotted black line). This is not out of step with previous work: stitching is often deployed as a theoretical tool to obtain optimal rates, not necessarily as a practical bound. We find that smaller values of  $\lambda$  tend to do better, but not monotonically. In practice we recommend sample splitting (if possible) to choose  $\lambda$ .

Let us now move on to Theorem 4.7, which explicitly relates the growth of  $\|S_t\|_{V_t^{-1}}$  to the inverse of  $\psi^*$ . As stated, the result is more of mathematical interest than of practical value.

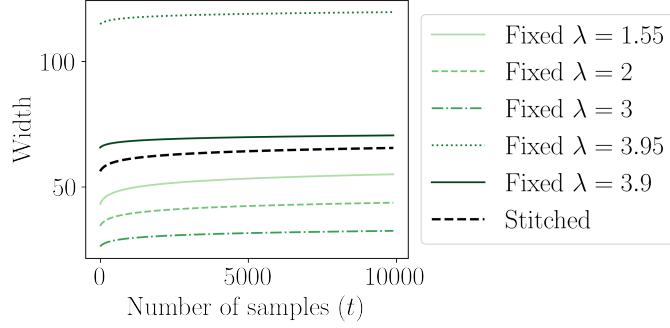


Figure 3: Comparison of our bound instantiated with a fixed  $\lambda$  (Theorem 4.1) versus stitching over lambda (Theorem 4.3). Here  $\psi = \psi_{G,c}$  with  $c = 1/4$ . Darker green lines correspond to higher values of  $\lambda$ .

We consider general CGF-like functions  $\psi$  which have a nonnegative third derivative and  $\lambda_{\max} < \infty$ , and show how to relate the growth rate to the convex conjugate of  $\psi$ . Since we stitch over epochs of the form  $\{t : \eta^k \leq \|S_\tau\|_{V_\tau^{-1}} < \eta^{k+1}\}$ , the iterated logarithm term of the bound is a function of  $\|S_\tau\|_{V_\tau^{-1}}$  itself instead of  $\det V_\tau$  as in Theorem 4.3. Such self-reference makes the bound difficult to apply (though this is remedied somewhat by Corollary 4.8 below). Still, as iterated logarithm terms grow extremely slowly,<sup>11</sup> we feel the result gives a sense of the growth rate of  $\|S_t\|_{V_t^{-1}}$ . The proof is provided in Appendix B.4. We refer again to Remark 4.2 which points out that the assumption that  $\psi(\lambda)/\lambda^2$  is nondecreasing is a minor one.

**Theorem 4.7.** *Let  $(S_t, V_t)$  be a sub- $\psi$  process where  $\psi$  is CGF-like,  $\psi(\lambda)/\lambda^2$  is nondecreasing, and  $\psi''' = \frac{d^3\psi}{d\lambda^3} \geq 0$ . In addition, assume that  $\lambda_{\max} < \infty$  and suppose that for all  $t \geq 1$ ,  $V_t \succeq U_0$  where  $U_0$  is positive-definite and satisfies  $\gamma_{\min}^{-1}(U_0) < \psi(\lambda_{\max})$ . Then, with probability  $1 - \delta$ , for all stopping times  $\tau$ ,*

$$\|S_\tau\|_{V_\tau^{-1}} \lesssim (\psi^*)^{-1} \left( \frac{\log(\det V_\tau) + \log(\log(\|S_\tau\|_{V_\tau^{-1}})) + \log(1/\delta)}{\sqrt{1 - 1/\gamma_{\min}(V_\tau)}} \right). \quad (4.9)$$

If we are prepared to sacrifice some tightness, we can move the iterated logarithm term outside of the convex conjugate in (4.9) and bound it by a constant times  $\|S_\tau\|_{V_\tau^{-1}}$ . This results in the following bound. The multiplicative factor in front of the convex conjugate is the cost of removing the self-reference from Theorem 4.7. The proof is in Appendix B.5.

**Corollary 4.8.** *Consider the same setup as Theorem 4.7. Let  $\varphi(y) = (\psi^*)^{-1}(y)$  and  $\rho = \gamma_{\min}(U_0)$ . Fix  $\delta \in (0, 1)$  and suppose that  $\varphi'(\log(1/\delta)) < \sqrt{1 - 1/\rho}$ . Then, with probability  $1 - \delta$ , for all stopping times  $\tau$ ,*

$$\|S_\tau\|_{V_\tau^{-1}} \lesssim \left( \frac{\sqrt{1 - 1/\rho}}{\sqrt{1 - 1/\rho} - \varphi'(\log(1/\delta))} \right) (\psi^*)^{-1} \left( \frac{\log(\det V_\tau) + \log(1/\delta)}{\sqrt{1 - 1/\gamma_{\min}(V_\tau)}} \right). \quad (4.10)$$

<sup>11</sup>For instance,  $\log_2(\log_2(10^{200})) < 10$ .  $10^{200}$  is roughly the number of books in Jorge Luis Borges' *Library of Babel* [8], while the approximate number of the atoms in the observable universe is a meager  $10^{80}$ .



## 5 Bernstein and Bennett Inequalities

In the scalar settings, two standard inequalities for light-tailed random variables are Bennett’s inequality and Bernstein’s inequality (see Boucheron et al. [9, Theorem 2.9 and Theorem 2.10]). The former holds for bounded random variables and the latter holds under a moment condition which has come to be known as “Bernstein’s condition.” Both results rely on upper bounding the log-MGF with a particular  $\psi$ -function;  $\psi_{P,c}$  in the case of Bennett’s inequality and  $\psi_{G,c}$  in the case of Bernstein’s. Here we provide dimension-free, self-normalized versions of these inequalities.

Consider a stream  $(X_t)$  of random vectors taking values in  $\mathbb{R}^d$  with conditional mean 0. That is,  $\mathbb{E}_{t-1}[X_t] = 0$ , where we use the shorthand  $\mathbb{E}_{t-1}[\cdot] = \mathbb{E}[\cdot | \mathcal{F}_{t-1}]$ . We assume a conditional mean of 0 for convenience only. If this condition is not met, we may consider the centered vectors  $X'_t = X_t - \mathbb{E}_{t-1}X_t$ . Note that assuming a constant *conditional* mean is more general than an iid assumption. In particular, it accommodates martingale dependence, making the resulting bounds useful in bandit applications (cf. [59, 1, 39]).

Our Bernstein and Bennett inequalities will both rely on the process  $(N_t^B(\theta; \lambda))$  for  $\theta \in \mathbb{S}^{d-1}$  and some  $\lambda > 0$  defined by

$$N_t^B(\theta; \lambda) = \prod_{k \leq t} \exp \{ \lambda \langle \theta, X_k \rangle - \log \mathbb{E}_{k-1} \exp(\lambda \langle \theta, X \rangle) \}, \quad (5.1)$$

which, as long as the log-MGF term  $\log \mathbb{E}_{k-1} \exp(\lambda \langle \theta, X \rangle)$  is finite, is easily seen to satisfy  $\mathbb{E}_{t-1} M_t(\theta) = M_{t-1}(\theta)$  and is hence a nonnegative martingale. Depending on the assumptions we make on  $(X_t)$ , we can bound the log-MGF by some function of  $\mathbb{E}_{k-1} X_k X_k^\top$ , which will give rise to a sub- $\psi$  process. The following lemma showcases this when the random vectors are bounded in each direction. The resulting process will provide our self-normalized Bennett’s inequality.

**Lemma 5.1.** *Let  $(X_t)$  be a stream of random vectors in  $\mathbb{R}^d$  with conditional mean 0 and  $\sup_{\theta \in \mathbb{S}^{d-1}} \langle \theta, X_t \rangle \leq b$  almost surely for some constant  $b > 0$ . Then  $S_t = \sum_{j \leq t} X_j$  and  $V_t = U_0 + \sum_{j \leq t} \mathbb{E}_{j-1} X_j X_j^\top$  constitute a sub- $\psi_{P,b}$  process for any PSD matrix  $U_0$ .*

The proof of Lemma 5.1 may be found in Appendix B.6. Next we show how to obtain a sub- $\psi$  process when the random vectors obey a Bernstein condition—the inequality in (5.2)—in each direction. We note that Bernstein’s condition is more general than boundedness, in the sense that bounded random vectors satisfy (5.2) but the converse is not true in general. The proof of the following result is in Appendix B.7.

**Lemma 5.2.** *Let  $(X_t)$  be a stream of random vectors with conditional mean 0. Suppose that for some  $c > 0$  and all  $\theta \in \mathbb{S}^{d-1}$ ,  $t \geq 1$ ,*

$$\mathbb{E}_{t-1}[\langle \theta, X_t \rangle^q] \leq \frac{q! c^{q-2}}{2} \mathbb{E}_{t-1} \langle \theta, X_t \rangle^2 \quad \text{for all integers } q \geq 3. \quad (5.2)$$

*Then  $S_t = \sum_{j \leq t} X_j$  and  $V_t = U_0 + \sum_{j \leq t} \mathbb{E}_{j-1} X_j X_j^\top$  constitute a sub- $\psi_{G,c}$  process for any PSD matrix  $U_0$ .*

Using these two results in conjunction with Theorem 4.1 we obtain the following line crossing inequalities. Figure 4 plots  $\psi_{P,1}^{-1}$  versus  $\psi_{G,1}^{-1}$  to give a sense of the constraints on  $\lambda$  in the following result. It also plots  $\psi_{G,1}$  versus  $\psi_{P,1}$  to give some sense of how (5.3) scales compared to (5.4).

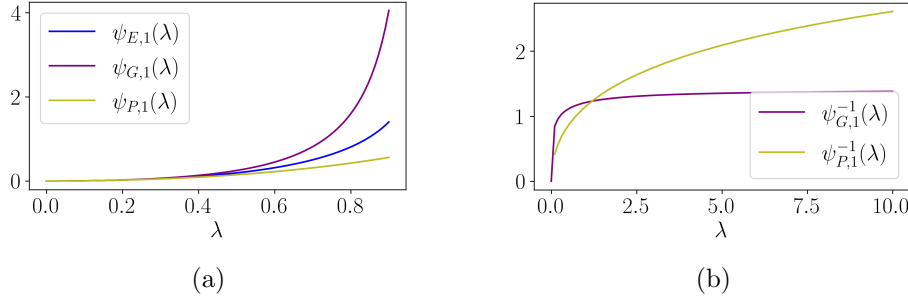


Figure 4: *Left:* Comparison of  $\psi_{E,1}$ ,  $\psi_{G,1}$ , and  $\psi_{P,1}$ . As shown in Lemma 6.2,  $\psi_{G,1}$  dominates  $\psi_{E,1}$  for all  $\lambda \in [0, 1]$ . *Right:* Comparison of  $\psi_{P,1}^{-1}$  and  $\psi_{G,1}^{-1}$ , which define the bounds for  $\lambda$  in Corollary 5.3.

**Corollary 5.3.** *Let  $(X_t)$  be a sequence of random vectors with conditional mean 0. Suppose either that (i)  $\sup_{\theta \in \mathbb{S}^{d-1}} \langle \theta, X_t \rangle \leq b$  for all  $t \geq 1$  and some  $b > 0$  or (ii) that (5.2) holds. Let  $S_t = \sum_{k \leq t} X_k$  and  $V_t = U_0 + \sum_{k \leq t} \mathbb{E}_{k-1}[X_k X_k^T]$  for some positive definite matrix  $U_0$ . Set  $\rho = \|U_0\|$ . Fix any  $\delta \in (0, 1)$ . If (i) holds, then for any  $\lambda \in [\psi_{P,b}^{-1}(1/\gamma_{\min}(U_0)), \infty)$ , with probability  $1 - \delta$ , for any stopping time  $\tau$ ,*

$$\|S_\tau\|_{V_\tau^{-1}} \leq \frac{\sqrt{\rho}}{\lambda g_{\psi_{P,b},\tau}(\lambda)} \left( \frac{1}{2} \log \left( \frac{\det V_\tau}{\det U_0} \right) + 1 + \log(1/\delta) \right) + \frac{g_{\psi_{P,b},\tau}(\lambda) \psi_{P,b}(\lambda)}{\lambda \sqrt{\rho}}, \quad (5.3)$$

*where  $g_{\psi_{P,b},\tau}$  is defined as in (4.3). If (ii) holds, then for any  $\lambda \in [\psi_{G,c}^{-1}(1/\gamma_{\min}(U_0)), 1/c)$ , with probability  $1 - \delta$ , for any stopping time  $\tau$ ,*

$$\|S_\tau\|_{V_\tau^{-1}} \leq \frac{\sqrt{\rho}}{\lambda g_{\psi_{G,c},\tau}(\lambda)} \left( \frac{1}{2} \log \left( \frac{\det V_\tau}{\det U_0} \right) + 1 + \log(1/\delta) \right) + \frac{g_{\psi_{G,c},\tau}(\lambda) \psi_{G,c}(\lambda)}{\lambda \sqrt{\rho}}. \quad (5.4)$$

Since the process is sub-gamma when  $(X_t)$  satisfy the Bernstein condition in (5.2), we can in that case apply Theorem 4.3 to obtain a stitched boundary. We omit the statement here for brevity, but we will provide a similar result in the next section when studying an empirical Bernstein bound.

## 6 An Empirical Bernstein Inequality

In this section we provide a self-normalized “empirical Bernstein” result. Empirical Bernstein bounds are useful because they do not require a priori knowledge of the variance in order to be instantiated. Instead, they adapt to the unknown variance. Such bounds have been studied in scalar random variables [45, 58, 34], vector-valued random variables [37, 42, 43, 14], u-statistics [50], and matrices [57]. As far as we are aware, Whitehouse et al. [60] gave the first empirical Bernstein result for self-normalized, vector-valued processes. Ours is thus the first dimension-free, self-normalized, empirical Bernstein inequality.

Let  $(X_t)$  be a stream of random vectors with conditional mean 0. Suppose that  $\|X_t\| \leq 1/2$  (one can replace  $1/2$  with any constant and rescale the bound accordingly). Recall that

$\psi_{E,1}(\lambda) = -\log(1 - \lambda) - \lambda$ . Denoting  $\hat{\mu}_t = \frac{1}{t} \sum_{j \leq t} X_j$ , the process defined by

$$N_t(\theta) = \prod_{k \leq t} \exp \{ \lambda \langle \theta, X_k \rangle - \psi_{E,1}(\lambda) \langle \theta, X_k - \hat{\mu}_{k-1} \rangle^2 \}, \quad N_0(\theta) = 1 \quad (6.1)$$

is a nonnegative supermartingale (see Chugg et al. [14, Lemma E.1] for an explicit proof). It is a multivariate extension of a supermartingale used by Howard et al. [34], which they use to prove an empirical Bernstein bound for scalar random variables. This supermartingale is, in turn, an extension of an inequality of Fan et al. [21], where the mean is replaced by the key term  $\hat{\mu}_{k-1}$  in the empirical variance. This reduces the variance when the mean of  $X_k$  is nonzero and delivers a certain asymptotic sharpness property of the resulting confidence set widths in scalar, vector, and matrix settings [58, 42, 57].

That the process in (6.1) is a supermartingale implies that  $(S_t, V_t)$  is a sub- $\psi_{E,1}$  process where

$$S_t = \sum_{k \leq t} X_k, \quad V_t = \rho I_d + \sum_{k \leq t} (X_k - \hat{\mu}_{k-1})(X_k - \hat{\mu}_{k-1})^\top, \quad (6.2)$$

for any fixed  $\rho \geq 0$ . (The term  $\rho I_d$  is not in (6.1), but adding to  $V_t$  only lowers the value of the process.) Since  $\psi_{E,1}$  is super-Gaussian, we may apply Theorem 4.1 to obtain the following result.

**Corollary 6.1.** *Let  $(X_t)$  be a stream of random vectors such that  $\|X_t\|_2 \leq 1/2$  for all  $t \geq 1$ . Let  $S_t = \sum_{k=1}^t X_k$  and  $V_t = \rho I_d + \sum_{k=1}^t (X_k - \hat{\mu}_{k-1})(X_k - \hat{\mu}_{k-1})^\top$  for any  $\rho > 1$ . Let  $g_t(z) = g_{\psi_{E,1},t}(z)$  for  $g_{\psi,t}$  as in (4.3). Then, for any  $\delta \in (0, 1)$  and any  $\lambda \in [\psi_{E,1}^{-1}(1/\rho), 1)$ , with probability  $1 - \delta$ , for any stopping time  $\tau$ ,*

$$\|S_\tau\|_{V_\tau^{-1}} \leq \frac{\sqrt{\rho}}{\lambda g_\tau(\lambda)} \left( \frac{1}{2} \log \left( \frac{\det V_\tau}{\det \rho I_d} \right) + 1 + \log(1/\delta) \right) + \frac{g_\tau(\lambda) \psi_{E,1}(\lambda)}{\lambda \sqrt{\rho}}. \quad (6.3)$$

In Corollary 6.1, we need to constrain  $\rho$  to be larger than 1 to ensure that a valid domain exists for  $\lambda$ . To elaborate, in Theorem 4.1 we require that  $\lambda \in [\psi^{-1}(1/\gamma_{\min}(U_0)), \lambda_{\max}]$ . Here,  $\psi^{-1}(1/\gamma_{\min}(U_0)) = \psi_{E,1}^{-1}(1/\rho)$  and  $\lambda_{\max} = 1$ . To have  $\psi_{E,1}^{-1}(1/\rho) < 1$  requires that  $\rho > 1$ .

A skeptical reader might eye the term  $\det(\rho I_d)$  in (6.3) with suspicion. This term depends on  $d$ , after all! However,  $\rho I_d$  can be replaced with any  $U_0$  such that  $\gamma_{\min}(U_0) > 1$ , which would make the bound truly dimension-free. We set  $U_0 = \rho I_d$  only for easier comparison to previous work.

In order to remove the dependence on  $\lambda$  in Corollary 6.1, we again turn to stitching. Since stitching with  $\psi = \psi_{E,1}$  is difficult to do directly, we instead upper bound  $\psi_{E,c}$  with  $\psi_{G,c}$  using the following lemma, and then apply Theorem 4.3.

**Lemma 6.2.** *For all  $\lambda \in [0, 1)$ ,  $\psi_{E,1}(\lambda) \leq \psi_{G,1}(\lambda)$ .*

*Proof.* Set  $h(\lambda) = \psi_{G,c}(\lambda) - \psi_{E,1}(\lambda)$ ,  $0 \leq \lambda < 1$ . Observe that  $h'(\lambda) = \lambda^2/[2(1 - \lambda)^2] \geq 0$ , implying that  $h(\lambda)$  is increasing in  $\lambda$ . Since  $h(0) = 0$  this implies that  $h(\lambda) \geq 0$  for  $\lambda$  in the desired range.  $\blacksquare$

As illustrated by Figure 4a, for small  $\lambda$ ,  $\psi_{E,1}$  and  $\psi_{G,1}$  are nearly identical. They separate more as  $\lambda$  grows (note that  $\lambda_{\max} = 1$ ). Using Lemma 6.2 and applying Theorem 4.3 with  $(S_t, V_t)$  as in (6.2) gives the following result. As in Corollary 6.1, we require that  $\rho > 1$  in order to ensure that  $[\psi_{G,1}^{-1}(1/\rho), 1)$  is non-empty.

**Corollary 6.3.** *Suppose  $\|X_t\|_2 \leq 1/2$ . Let  $V_t = \rho I_d + \sum_{k=1}^t (X_k - \hat{\mu}_{k-1})(X_k - \hat{\mu}_{k-1})^\top$  for any  $\rho > 1$ . Fix  $\delta \in (0, 1/\sqrt{2}]$ . Then, with probability  $1 - \delta$ , for all stopping times  $\tau$ ,*

$$\|S_\tau\|_{V_\tau^{-1}} \leq \frac{D_\tau + 1.60\sqrt{D_\tau} + \max\left\{\frac{1+\sqrt{1+2\rho}}{2\rho}, \sqrt{\frac{D_\tau}{2}}\right\}}{H_\tau}, \quad (6.4)$$

where

$$D_\tau = \frac{1}{2} \log\left(\frac{\det V_\tau}{\det(\rho I_d)}\right) + 1.5 + 2 \log(\log_2(\det V_\tau) + 1) + \log(1/\delta), \quad (6.5)$$

and

$$H_\tau = 0 \vee \left( \sqrt{1 - \frac{\rho}{\gamma_{\min}(V_t)}} - \sqrt{\frac{2\rho}{\gamma_{\min}(V_t)(1 + \sqrt{3})}} \right). \quad (6.6)$$

## 7 Summary

This work has asked and answered the question: Can self-normalized, dimension-free, bounds be given for processes beyond the sub-Gaussian case? We provide such bounds for sub- $\psi$  processes, a general class of stochastic process which includes sub-exponential, sub-Poisson, sub-gamma, and sub-Gaussian distributions. Our methods are based on the variational approach to concentration, a relatively new and (we believe) exciting approach to multivariate concentration.

Our results bridge a gap in the literature between determinant-based bounds, previously confined to sub-Gaussian or bounded observations [1, 3, 65, 46], and the dimension-dependent, condition number-based bounds of Whitehouse et al. [60]. We show that one can enjoy both dimension-free bounds *and* the distributional generality provided by Whitehouse et al.

The story does not end here. For one, the optimality (or suboptimality) of our bounds has not been established. Can we remove the dependence on  $H_{c,\tau}$  in Theorem 4.3, for instance? Is the function  $g_\psi$  in Theorem 4.1 necessary, or is it simply an artifact of the proof technique? While  $g_\psi$  is easy to compute, its presence makes it tougher to reason about these bounds analytically.

Second, the fact that our bounds are dimension-free suggests that they can be applied to infinite-dimensional spaces. It is typically possible to extend dimension-free results from finite dimensional spaces to infinite dimensional spaces (e.g., [47]). While we expect this is also possible for our bounds, some of the details pose challenges. For instance, Theorem 4.1 relies on using uniform distributions, which are not well-defined in general Hilbert spaces as there is no Lebesgue measure in infinite dimensions.

More general spaces beyond Hilbert spaces are also of interest. Several previous works have studied mean concentration in smooth Banach spaces, for instance [51, 42, 61]. Can self-normalized results be given in such a general setting? Can the variational technique be applied in such a setting?

## Acknowledgments

BC was supported in part by NSERC-PGSD, grant no. 567944. Both authors acknowledge support from NSF grants IIS-2229881 and DMS-2310718.

## References

- [1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011. 1, 2, 3, 4, 5, 7, 8, 15, 18, 23
- [2] 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. 2
- [3] Arya Akhavan, Amitis Shidani, Alex Ayoub, and David Janz. Bernstein-type dimension-free concentration for self-normalised martingales. *arXiv preprint arXiv:2507.20982*, 2025. 3, 4, 18
- [4] Pierre Alquier et al. User-friendly introduction to PAC-Bayes bounds. *Foundations and Trends® in Machine Learning*, 17(2):174–303, 2024. 7
- [5] Bernard Bercu and Abderrahmen Touati. Exponential inequalities for self-normalized martingales with applications. *The Annals of Applied Probability*, 18:1848–1869, 2008. 3
- [6] Bernard Bercu and Taieb Touati. New insights on concentration inequalities for self-normalized martingales. *Electronic Communications in Probability*, 24, 2019. 2, 3
- [7] David Blackwell and David Freedman. On the amount of variance needed to escape from a strip. *The Annals of Probability*, pages 772–787, 1973. 5
- [8] Jorge Luis Borges. *Labyrinths: Selected stories & other writings*. Number 186. New Directions Publishing, 1964. 14
- [9] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. 6, 7, 15
- [10] Olivier Catoni and Ilaria Giulini. Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv preprint arXiv:1712.02747*, 2017. 2, 4
- [11] Olivier Catoni and Ilaria Giulini. Dimension-free PAC-Bayesian bounds for the estimation of the mean of a random vector. *(Almost) 50 Shades of Bayesian Learning: PAC-Bayesian Trends and Insights. NeurIPS Workshop.*, 2018. 2, 4
- [12] Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pages 844–853. PMLR, 2017. 2, 3, 4
- [13] Ben Chugg, Hongjian Wang, and Aaditya Ramdas. A unified recipe for deriving (time-uniform) PAC-Bayes bounds. *Journal of Machine Learning Research*, 24(372):1–61, 2023. 7, 11, 12
- [14] Ben Chugg, Hongjian Wang, and Aaditya Ramdas. Time-uniform confidence spheres for means of random vectors. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. 2, 4, 11, 16, 17, 25
- [15] Victor H de la Peña, Michael J Klass Lai, and Tze Leung. Pseudo-maximization and self-normalized processes. *Probability Surveys*, 4:172–192, 2007. 1, 6, 7, 12
- [16] 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. 1, 2, 3
- [17] Victor H De La Peña and Guodong Pang. Exponential inequalities for self-normalized processes with applications. *Electronic Communications in Probability*, 14:372–381, 2009. 1

- [18] Victor H 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), 2004. 1, 2, 3, 4, 6
- [19] Victor H de la Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media, 2008. 1, 3, 4, 7, 23
- [20] Bradley Efron. Student’s t-test under symmetry conditions. *Journal of the American Statistical Association*, 64(328):1278–1302, 1969. 3
- [21] Xiequan Fan, Ion Grama, and Quansheng Liu. Exponential inequalities for martingales with applications. *Electron. J. Probab*, 20(1):1–22, 2015. 17
- [22] Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pages 3052–3060. PMLR, 2020. 3, 4
- [23] Hamish Flynn, David Reeb, Melih Kandemir, and Jan R Peters. Improved algorithms for stochastic linear bandits using tail bounds for martingale mixtures. *Advances in Neural Information Processing Systems*, 36:45102–45136, 2023. 7
- [24] David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975. 6
- [25] Evarist Giné and David M Mason. On the LIL for self-normalized sums of IID random variables. *Journal of Theoretical Probability*, 11(2):351–370, 1998. 3
- [26] Evarist Giné, Friedrich Götze, and David M Mason. When is the Student  $t$ -statistic asymptotically standard normal? *The Annals of Probability*, 25(3):1514–1531, 1997. 3
- [27] Ilaria Giulini. Robust dimension-free Gram operator estimates. *Bernoulli*, 24(4B), 2018. 4
- [28] Philip Griffin and James Kuelbs. Some extensions of the LIL via self-normalizations. *The Annals of Probability*, pages 380–395, 1991. 3
- [29] Philip S Griffin and James D Kuelbs. Self-normalized laws of the iterated logarithm. *The Annals of Probability*, pages 1571–1601, 1989. 3
- [30] Benjamin Guedj. A primer on PAC-Bayesian learning. *arXiv preprint arXiv:1901.05353*, 2019. 7
- [31] Nicholas J Higham. *Accuracy and stability of numerical algorithms*. SIAM, 2002. 2
- [32] Harold Hotelling. The generalization of student’s ratio. *The Annals of Mathematical Statistics*, 2(3):360–378, 1931. 3
- [33] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. 2, 3, 6, 9, 11
- [34] 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. 5, 6, 9, 11, 12, 16, 17
- [35] Bing-Yi Jing and Qiying Wang. An exponential nonuniform Berry-Esseen bound for self-normalized sums. *The Annals of Probability*, 27(4):2068–2088, 1999. 3



- [36] Johannes Kirschner, Andreas Krause, Michele Meziu, and Mojmir Mutny. Confidence estimation via sequential likelihood mixing. *arXiv preprint arXiv:2502.14689*, 2025. 4, 5
- [37] Jonas Moritz Kohler and Aurelien Lucchi. Sub-sampled cubic regularization for non-convex optimization. In *International Conference on Machine Learning*, pages 1895–1904. PMLR, 2017. 16
- [38] 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*, pages 154–166, 1982. 2
- [39] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020. 2, 15
- [40] Benjamin F Logan, CL Mallows, SO Rice, and Larry A Shepp. Limit distributions of self-normalized sums. *The Annals of Probability*, 1(5):788–809, 1973. 3
- [41] Tudor Manole and Aaditya Ramdas. Martingale methods for sequential estimation of convex functionals and divergences. *IEEE Transactions on Information Theory*, 69(7):4641–4658, 2023. 13
- [42] Diego Martinez-Taboada and Aaditya Ramdas. Empirical Bernstein in smooth Banach spaces. *arXiv preprint arXiv:2409.06060*, 2024. 16, 17, 18
- [43] Diego Martinez-Taboada and Aaditya Ramdas. Sharp empirical Bernstein bounds for the variance of bounded random variables. *arXiv preprint arXiv:2505.01987*, 2025. 16
- [44] Nikolai Matni and Stephen Tu. A tutorial on concentration bounds for system identification. In *2019 IEEE 58th conference on decision and control (CDC)*, pages 3741–3749. IEEE, 2019. 2
- [45] Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. *Conference on Learning Theory*, 2009. 16
- [46] Alberto Maria Metelli, Simone Drago, and Marco Mussi. Generalized kernelized bandits: Self-normalized Bernstein-like dimension-free inequality and regret bounds. *arXiv preprint arXiv:2508.01681*, 2025. 3, 4, 18
- [47] Mattes Mollenhauer and Claudia Schillings. On the concentration of subGaussian vectors and positive quadratic forms in Hilbert spaces. *arXiv preprint arXiv:2306.11404*, 2023. 18
- [48] Shogo Nakakita, Pierre Alquier, and Masaaki Imaizumi. Dimension-free bounds for sums of dependent matrices and operators with heavy-tailed distributions. *Electronic Journal of Statistics*, 18(1):1130–1159, 2024. 2, 4
- [49] Roberto Imbuzeiro Oliveira. The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields*, 166:1175–1194, 2016. 2, 4
- [50] Thomas Peel, Sandrine Anthoine, and Liva Ralaivola. Empirical Bernstein inequalities for u-statistics. *Advances in Neural Information Processing Systems*, 23, 2010. 16
- [51] Iosif Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *The Annals of Probability*, pages 1679–1706, 1994. 18
- [52] Qi-Man Shao. Self-normalized large deviations. *The Annals of Probability*, 25(1):285–328, 1997. 3

- [53] Xiaofeng Shao. A self-normalized approach to confidence interval construction in time series. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(3):343–366, 2010. [2](#)
- [54] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908. [2](#)
- [55] Michel Talagrand. *Upper and lower bounds for stochastic processes*, volume 60. Springer, 2014. [8](#)
- [56] Daniel Vial, Advait Parulekar, Sanjay Shakkottai, and R Srikant. Improved algorithms for mis-specified linear Markov decision processes. In *International Conference on Artificial Intelligence and Statistics*, pages 4723–4746. PMLR, 2022. [2](#)
- [57] Hongjian Wang and Aaditya Ramdas. Sharp matrix empirical Bernstein inequalities. *arXiv preprint arXiv:2411.09516*, 2024. [16](#), [17](#)
- [58] Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. [16](#), [17](#)
- [59] 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. [2](#), [13](#), [15](#)
- [60] Justin Whitehouse, Zhiwei Steven Wu, and Aaditya Ramdas. Time-uniform self-normalized concentration for vector-valued processes. *arXiv preprint arXiv:2310.09100*, 2023. [2](#), [3](#), [4](#), [6](#), [11](#), [12](#), [13](#), [16](#), [18](#), [34](#), [36](#), [37](#)
- [61] Justin Whitehouse, Ben Chugg, Diego Martinez-Taboada, and Aaditya Ramdas. Mean estimation in Banach spaces under infinite variance and martingale dependence. *arXiv preprint arXiv:2411.11271*, 2024. [18](#)
- [62] Weitong Zhang, Zhiyuan Fan, Jiafan He, and Quanquan Gu. Achieving constant regret in linear Markov decision processes. *arXiv preprint arXiv:2404.10745*, 2024. [2](#)
- [63] Nikita Zhivotovskiy. Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle. *Electronic Journal of Probability*, 29:1–28, 2024. [2](#), [4](#)
- [64] Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture Markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021. [3](#), [4](#)
- [65] Ingvar Ziemann. A vector Bernstein inequality for self-normalized martingales. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. [2](#), [4](#), [5](#), [8](#), [18](#)

## A Method of Mixtures for sub-Gaussian Processes

In this section we show that the method of mixtures—the proof method used by Abbasi-Yadkori et al. [1] and de la Peña et al. [19]—can also be used to prove Theorem 3.1. To repeat what was said in that section, existing proofs assume that  $V_t$  is predictable, not adapted. Here we show that the same proof technique can accommodate an adapted variance process.

Let  $(S_t, V_t)$  be a sub- $\psi_N$  process, implying that for all  $\theta \in \mathbb{R}^d$ ,

$$M_t(\theta) = \exp\{\langle \theta, S_t \rangle - \psi_N(1)\langle \theta, V_t \theta \rangle\} \leq N_t(\theta),$$

where  $(N_t(\theta))$  is a nonnegative supermartingale. (To see why we can consider all  $\theta \in \mathbb{R}^d$  instead of  $\theta \in \mathbb{S}^{d-1}$ , see Section B.1 below.) Let  $\nu$  be a Gaussian with mean 0 and covariance  $U_0^{-1}$  and consider the process with increments

$$M_t = \int_{\mathbb{R}^d} \exp\{\langle \theta, S_t \rangle - \psi_N(1)\langle \theta, V_t \theta \rangle\} \nu(d\theta). \quad (\text{A.1})$$

To compute  $M_t$ , notice that we can write

$$\langle \theta, S_t \rangle - \psi_N(1)\langle \theta, V_t \theta \rangle = \frac{1}{2}\|S_t\|_{V_t^{-1}}^2 - \frac{1}{2}\|\theta - V_t^{-1}S_t\|_{V_t}^2,$$

and

$$\|\theta - V_t^{-1}S_t\|_{V_t}^2 + \langle \theta, U_0 \theta \rangle = \|\theta - (U_0 + V_t)^{-1}S_t\|_{U_0+V_t}^2 - 2\|S_t\|_{V_t^{-1}}^2 + 2\|S_t\|_{(U_0+V_t)^{-1}}^2.$$

Hence, writing out the density of  $\nu$ ,

$$\begin{aligned} M_t &= \frac{\exp(\frac{1}{2}\|S_t\|_{V_t^{-1}}^2)}{\sqrt{2\pi \det(U_0^{-1})}} \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\|\theta - V_t^{-1}S_t\|_{V_t}^2 - \frac{1}{2}\langle \theta, U_0 \theta \rangle\right) d\theta \\ &= \frac{\exp(\frac{1}{2}\|S_t\|_{(U_0+V_t)^{-1}}^2)}{\sqrt{2\pi \det(U_0^{-1})}} \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\|\theta - (U_0 + V_t)^{-1}S_t\|_{U_0+V_t}^2\right) d\theta \\ &= \frac{\exp(\frac{1}{2}\|S_t\|_{(U_0+V_t)^{-1}}^2)}{\sqrt{2\pi \det(U_0^{-1})}} \sqrt{2\pi \det((U_0 + V_t)^{-1})} \\ &= \sqrt{\frac{\det(U_0)}{\det(U_0 + V_t)}} \exp\left(\frac{1}{2}\|S_t\|_{(U_0+V_t)^{-1}}^2\right). \end{aligned}$$

Since  $\int N_t(\theta) d\nu(\theta)$  is a supermartingale by Fubini's theorem,  $M_t$  remains upper bounded by a nonnegative supermartingale. We may thus apply Ville's inequality to obtain  $P(M_\tau \geq 1/\delta) \leq \mathbb{E}[M_1]\delta \leq \delta$ . In other words, with probability  $1 - \delta$ ,  $\log M_\tau \leq 1/\delta$ , which translates to

$$\frac{1}{2}\|S_\tau\|_{(U_0+V_\tau)^{-1}}^2 \leq \frac{1}{2}\log\left(\frac{\det(U_0 + V_\tau)}{\det U_0}\right) + \log(1/\delta),$$

which is precisely Theorem 3.1.

## B Omitted Proofs

### B.1 Proof of Theorem 3.1

Let  $(S_t)$  be a sub-Gaussian process with variance proxy  $V_t$ . For each  $\lambda \geq 0$  and  $\theta \in \mathbb{R}^d$ , let

$$Z_t^\lambda(\theta) = \exp \{ \lambda \langle \theta, S_t \rangle - \psi_N(\lambda) \langle \theta, V_t \theta \rangle \}. \quad (\text{B.1})$$

For  $\theta \in \mathbb{S}^{d-1}$ ,  $(Z_t(\theta))$  is upper-bounded by a nonnegative supermartingale by Definition 2.1. We claim that this property extends to all  $\theta \in \mathbb{R}^d$ . Indeed, given any such  $\theta$ , let  $\phi = \theta / \|\theta\|_2$ . Note that  $\phi \in \mathbb{S}^{d-1}$ . Observe that

$$Z_t^\lambda(\theta) = \exp \{ \lambda \|\theta\| \langle \phi, S_t \rangle - \psi_N(\lambda \|\theta\|) \langle \phi, V_t \phi \rangle \} = Z_t^{\lambda \|\theta\|}(\phi),$$

which is upper bounded by  $L_t^{\lambda \|\theta\|}(\phi)$  (note that  $\lambda_{\max} = \infty$  in this case). Thus, the process defined by (B.1) is upper bounded by a nonnegative supermartingale for all  $\theta \in \mathbb{R}^d$ , enabling us to apply Proposition 2.2 with  $\Theta = \mathbb{R}^d$ .

Let  $\tau$  be a stopping time and consider a Gaussian distribution  $\rho = \rho_\tau$  with mean  $\mu_\rho$  and  $\Sigma_\rho$ . We have

$$\begin{aligned} \int \log Z_\tau(\theta) d\rho &= \lambda \langle \mu_\rho, S_\tau \rangle - \psi_N(\lambda) \int \langle \theta, V_\tau \theta \rangle d\rho \\ &= \lambda \langle \mu_\rho, S_\tau \rangle - \psi_N(\lambda) (\langle \mu_\rho, V_\tau \mu_\rho \rangle + \text{Tr}(V_\tau \Sigma_\rho)). \end{aligned} \quad (\text{B.2})$$

Define  $\widehat{V}_t = V_t + U_0$  and consider setting  $\mu_\rho = \widehat{V}_\tau^{-1} S_\tau$ . Then

$$\begin{aligned} \int \log Z_\tau(\theta) \rho(d\theta) &= \lambda \|S_\tau\|_{\widehat{V}_\tau^{-1}}^2 - \psi_N(\lambda) \|S_\tau\|_{\widehat{V}_\tau^{-1} V_\tau \widehat{V}_\tau^{-1}}^2 - \psi_N(\lambda) \text{Tr}(V_\tau \Sigma_\rho) \\ &= (\lambda - \psi_N(\lambda)) \|S_\tau\|_{\widehat{V}_\tau^{-1}}^2 + \psi_N(\lambda) \|S_\tau\|_{\widehat{V}_\tau^{-1} U_0 \widehat{V}_\tau^{-1}}^2 - \psi_N(\lambda) \text{Tr}(V_\tau \Sigma_\rho). \end{aligned}$$

Let  $\nu$  be a Gaussian with mean 0 and covariance  $\Sigma_\nu$ . Then

$$\begin{aligned} D_{\text{KL}}(\rho \| \nu) &= \frac{1}{2} \left\{ \text{Tr}(\Sigma_\nu^{-1} \Sigma_\rho) + \langle \mu_\rho, \Sigma_\nu^{-1} \mu_\rho \rangle - d + \log \left( \frac{\det \Sigma_\nu}{\det \Sigma_\rho} \right) \right\} \\ &= \frac{1}{2} \left\{ \text{Tr}(\Sigma_\nu^{-1} \Sigma_\rho) + \|S_\tau\|_{\widehat{V}_\tau^{-1} \Sigma_\nu^{-1} \widehat{V}_\tau^{-1}}^2 - d + \log \left( \frac{\det \Sigma_\nu}{\det \Sigma_\rho} \right) \right\} \end{aligned}$$

In order to match the term  $\psi_N(\lambda) \|S_\tau\|_{\widehat{V}_\tau^{-1} U_0 \widehat{V}_\tau^{-1}}^2$  above, we choose  $\Sigma_\nu = (2\psi_N(\lambda) U_0)^{-1}$ . Proposition 2.2 implies that with probability  $1 - \delta$ ,

$$\int \log Z_\tau(\theta) d\rho \leq D_{\text{KL}}(\rho \| \nu) + \log(1/\delta),$$

which in our case rearranges to

$$(\lambda - \psi_N(\lambda)) \|S_\tau\|_{\widehat{V}_\tau^{-1}}^2 \leq \psi_N(\lambda) \text{Tr}(\widehat{V}_\tau \Sigma_\rho) - \frac{d}{2} + \frac{1}{2} \log \left( \frac{\det \Sigma_\nu}{\det \Sigma_\rho} \right) + \log(1/\delta). \quad (\text{B.3})$$

Taking  $\Sigma_\rho = (2\psi_N(\lambda) \widehat{V}_\tau)^{-1}$ , gives  $\psi_N(\lambda) \text{Tr}(\widehat{V}_\tau \Sigma_\rho) = d/2$ . Also,  $\det \Sigma_\nu / \det \Sigma_\rho = \det \widehat{V}_\tau / \det U_0$ . Assuming  $\lambda - \psi_N(\lambda) > 0$ , (B.3) thus becomes

$$\|S_\tau\|_{(V_\tau + U_0)^{-1}}^2 \leq \frac{1}{\lambda - \psi_N(\lambda)} \left( \frac{1}{2} \log \left( \frac{\det(V_\tau + U_0)}{\det U_0} \right) + \log(1/\delta) \right). \quad (\text{B.4})$$

Finally, we optimize over  $\lambda$  to achieve the maximum value of  $\lambda^* - \psi_N(\lambda^*) = 1/2$  at  $\lambda^* = 1$ . This completes the proof.

## B.2 Proof of Theorem 4.1

Let  $(S_t, V_t)$  be a sub- $\psi$  process and set

$$Z_t(\theta) = \exp \{ \lambda \langle \theta, S_t \rangle - \psi(\lambda) \langle \theta, V_t \theta \rangle \}. \quad (\text{B.5})$$

For any  $\theta \in \mathbb{S}^{d-1}$ , the process  $(Z_t(\theta))$  is upper bounded by some nonnegative supermartingale by assumption. We claim that the same holds for all  $\theta \in \mathbb{B}^d$ . This is shown in Chugg et al. [14, Lemma 2.10], but let us prove it for completeness. For any  $\theta \in \mathbb{B}^d$ , let  $\phi = \theta / \|\theta\|$  and notice that

$$\begin{aligned} & \exp \{ \lambda \langle \theta, S_t \rangle - \psi(\lambda) \langle \theta, V_t \theta \rangle \} \\ &= \exp \{ \lambda \|\theta\| \langle \phi, S_t \rangle - \psi(\lambda) \|\theta\|^2 \langle \phi, V_t \phi \rangle \} \\ &\leq \exp \{ \lambda \|\theta\| \langle \phi, S_t \rangle - \phi(\|\theta\| \lambda) \langle \phi, V_t \phi \rangle \}, \end{aligned}$$

where the inequality follows from the fact that  $\phi$  is super-Gaussian. We may henceforth assume that  $(Z_t(\theta))$  is upper-bounded by a nonnegative supermartingale for all  $\theta \in \mathbb{B}^d$ , enabling us to take  $\Theta = \mathbb{B}^d$  in Proposition 2.2.

Let  $\tau$  be a stopping time and let  $\rho = \rho_\tau$  be a distribution over  $\mathbb{B}^d$  with mean  $\mu_\rho$  and covariance  $\Sigma_\rho$ . Then,

$$\begin{aligned} \int \log Z_\tau(\theta) d\rho &= \lambda \langle \mu_\rho, S_\tau \rangle - \psi(\lambda) \int \langle \theta, V_\tau \theta \rangle d\rho \\ &= \lambda \langle \mu_\rho, S_\tau \rangle - \psi(\lambda) (\langle \mu_\rho, V_\tau \mu_\rho \rangle + \text{Tr}(V_\tau \Sigma_\rho)). \end{aligned} \quad (\text{B.6})$$

For some  $\beta > 0$  to be determined later, we take  $\rho$  to be the uniform distribution over the ellipsoid with mean  $\mu_\rho = \beta V_\tau^{-1} S_\tau$  and shape  $A_\rho = \psi^{-1}(\lambda) V_\tau^{-1}$ . That is,  $\rho$  is uniform over  $\mathcal{E}_\rho := \{ \theta \in \mathbb{R}^d : (\theta - \mu_\rho)^t A_\rho^{-1} (\theta - \mu_\rho) \leq 1 \}$  which we assume for now is a subset of  $\mathbb{B}^d$ . Note that similarly to the proof of Theorem 3.1, both  $\mu_\rho$  and  $A_\rho$  are functions of  $\tau$  but we suppress this dependence for convenience. The distribution  $\rho$  has covariance  $\Sigma_\rho = (d+2)^{-1} A_\rho$  and

$$\psi(\lambda) \text{Tr}(V_\tau \Sigma_\rho) = \frac{d}{d+2} \leq 1.$$

With these choices, (B.6) becomes

$$\int \log Z_\tau(\theta) d\rho = (\lambda\beta - \beta^2\psi(\lambda)) \langle V_\tau^{-1} S_\tau, S_\tau \rangle - \frac{d}{d+2}. \quad (\text{B.7})$$

Proposition 2.2 then gives that with probability  $1 - \delta$ ,

$$(\lambda\beta - \beta^2\psi(\lambda)) \|S_\tau\|_{V_\tau^{-1}}^2 \leq D_{\text{KL}}(\rho\|\nu) + 1 + \log(1/\delta). \quad (\text{B.8})$$

for a data-independent prior  $\nu$ . If  $\nu$  is uniform over the ellipsoid  $\mathcal{E}_\nu = \{ \theta : \theta^t A_\nu^{-1} \theta \leq 1 \}$  (which, again, we require to be a subset of  $\mathbb{B}^d$ ) and  $\beta$  is small enough such that  $\mathcal{E}_\rho \subseteq \mathcal{E}_\nu$ , then

$$D_{\text{KL}}(\rho\|\nu) = \int \log \left( \frac{d\rho}{d\nu}(\theta) \right) d\rho = \int \log \left( \frac{\text{vol}(\mathcal{E}_\nu)}{\text{vol}(\mathcal{E}_\rho)} \right) d\rho = \log \left( \frac{\sqrt{\det A_\nu}}{\sqrt{\det A_\rho}} \right).$$

If we take  $A_\nu = \psi^{-1}(\lambda)U_0^{-1}$  then

$$D_{\text{KL}}(\rho\|\nu) = \frac{1}{2} \log \left( \frac{\det V_\tau}{\det U_0} \right). \quad (\text{B.9})$$

To ensure that  $\mathcal{E}_\nu \subseteq \mathbb{B}^d$  it suffices that the minimum eigenvalue of  $A_\nu^{-1}$  is at least 1, which holds if  $\psi(\lambda)\gamma_{\min}(U_0) \geq 1$ . That is,  $\lambda \geq \psi^{-1}(1/\gamma_{\min}(U_0))$ . This gives the lower bound on  $\lambda$  in the statement of the theorem.

It remains to pick  $\beta$ , which must be small enough such that  $\mathcal{E}_\rho \subseteq \mathcal{E}_\nu$ . That is, if  $(\theta - \beta V_\tau^{-1}S_t)^t A_\rho^{-1}(\theta - \beta V_\tau^{-1}S_t) \leq 1$  we want to ensure that  $\theta^t A_\nu^{-1} \theta \leq 1$ . Equivalently, we may shift  $\theta$  by  $\beta V_\tau^{-1}S_\tau$  and show that if  $\theta^t V_\tau \theta \leq \psi^{-1}(\lambda)$  then  $(\theta + \beta V_\tau^{-1}S_\tau)^t U_0(\theta + \beta V_\tau^{-1}S_\tau) \leq \psi^{-1}(\lambda)$ , where we've recalled the definition of  $A_\rho$  and  $A_\nu$ . Put  $y = y_\tau = V_\tau^{-1}S_\tau$ . Suppose that  $\theta^t V_\tau \theta \leq \psi^{-1}(\lambda)$ . For any  $\epsilon > 0$ ,

$$\begin{aligned} (\theta + \beta y)^t U_0(\theta + \beta y) &= \theta^t U_0 \theta + 2\beta \theta^t U_0 y + \beta^2 y^t U_0 y \\ &\leq (1 + \epsilon) \theta^t U_0 \theta + \beta^2 (1 + \epsilon^{-1}) y^t U_0 y && \text{Young's inequality} \\ &\leq (1 + \epsilon) \alpha_\tau \theta^t V_\tau \theta + \beta^2 (1 + \epsilon^{-1}) y^t U_0 y && \text{Rayleigh coef.} \\ &\leq (1 + \epsilon) \alpha_\tau \psi^{-1}(\lambda) + \beta^2 (1 + \epsilon^{-1}) y^t U_0 y && \theta^t V_\tau \theta \leq \psi^{-1}(\lambda) \\ &=: h(\epsilon, \beta). \end{aligned}$$

Minimizing  $h(\epsilon, \beta)$  over  $\epsilon > 0$  gives

$$\epsilon^* = \beta \sqrt{\frac{y^t U_0 y}{\alpha_\tau \psi^{-1}(\lambda)}},$$

so that  $h(\epsilon^*, \beta) = \beta^2 \|y\|_{U_0}^2 + 2\beta \|y\|_{U_0} \sqrt{\alpha_\tau \psi^{-1}(\lambda)} + \alpha_\tau \psi^{-1}(\lambda)$ . Setting  $h(\epsilon^*, \beta) = \psi^{-1}(\lambda)$  and solving for  $\beta$  gives

$$\beta^* = \frac{\sqrt{\alpha_\tau \psi^{-1}(\lambda) + 1 - \alpha_\tau} - \sqrt{\alpha_\tau \psi^{-1}(\lambda)}}{\|y\|_{U_0}} = \frac{g_\tau(\lambda)}{\|y\|_{U_0}}. \quad (\text{B.10})$$

Notice that  $\|y\|_{U_0}^2 = S_\tau^t V_\tau^{-1} U_0 V_\tau^{-1} S_\tau \leq \|U_0\| \|S_\tau\|_{V_\tau^{-1}}^2$ . Therefore, setting

$$\beta^{**} = \frac{g_\tau(\lambda)}{\sqrt{\|U_0\|} \|S_\tau\|_{V_\tau^{-1}}} \leq \beta^*,$$

we obtain  $(\theta + \beta^{**} y)^t U_0(\theta + \beta^{**} y) \leq h(\epsilon^*, \beta^{**}) \leq h(\epsilon^*, \beta^*) = \psi^{-1}(\lambda)$  (where we've used that  $h$  is monotonically increasing in  $\beta$ ), so (B.8) holds with  $\beta = \beta^{**}$ . Noting that

$$(\lambda \beta^{**} - (\beta^{**})^2 \psi(\lambda)) \|S_\tau\|_{V_\tau^{-1}}^2 = \frac{\lambda g_\tau(\lambda)}{\sqrt{\|U_0\|}} \|S_\tau\|_{V_\tau^{-1}} - \frac{g_\tau^2(\lambda) \psi(\lambda)}{\|U_0\|},$$

and rearranging (B.8) completes the proof.



### B.3 Proof of Theorem 4.3

Theorem 4.3 follows from the following more general result after taking  $\eta = 2$  and  $\ell(k) = (k+1)^2\pi^2/6$ . In this case  $\sup_{k \geq 0} \ell(k+2)/\ell(k+1) \leq 4$  so  $\alpha_{\delta,\eta} \leq 5.07$ , and

$$\log(\ell(\log_\eta(\det V_\tau) + 1)) = \log\left(\frac{\pi^2}{6}\right) + 2\log(\log_2(\det V_\tau) + 1),$$

which gives rise to the constants in Theorem 4.3.

**Theorem B.1.** *Let  $(S_t, V_t)$  be a sub- $\psi_{G,c}$  process for any  $c > 0$  and suppose that  $V_1 \succeq I_d$  and  $V_t \succeq U_0$  for some positive-definite  $U_0$ . Let  $\ell : \mathbb{N}_0 \rightarrow \mathbb{R}_{>0}$  satisfy  $\sum_{k \geq 0} \ell^{-1}(k) = 1$ . Fix  $\delta \in (0, 1)$  and take any  $1 < \eta < (e/\delta)^2$ . Set*

$$\alpha_{\delta,\eta} := 1 + \frac{\log(\eta)}{1 + \log(1/\delta) - \log(\eta)/2} + \sup_{k \geq 0} \frac{\ell(k+2)}{\ell(k+1)}. \quad (\text{B.11})$$

Then, with probability  $1 - \delta$ , for all stopping times  $\tau$ ,

$$\|S_\tau\|_{V_\tau^{-1}} \leq \frac{cD_{\tau,\eta}(U_0, V_\tau) + \sqrt{\frac{\alpha_{\delta,\eta}}{2}D_{\tau,\eta}(U_0, V_\tau)} + \max\left\{\frac{c + \sqrt{c^2 + 2\rho}}{2\rho}, \sqrt{\frac{D_{\tau,\eta}(U_0, V_\tau)}{2}}\right\}}{H_c(U_0, V_\tau)}, \quad (\text{B.12})$$

where

$$D_{\tau,\eta}(U_0, V_\tau) = \frac{1}{2} \log\left(\frac{\det V_\tau}{\det U_0}\right) + 1 + \log(\ell(\log_\eta(\det V_\tau))) + \log(1/\delta), \quad (\text{B.13})$$

and

$$H_c(U_0, V_t) = 0 \vee \left( \sqrt{1 - \frac{\gamma_{\max}(U_0)}{\gamma_{\min}(V_t)}} - \sqrt{\frac{\gamma_{\max}(U_0)}{\gamma_{\min}(V_t)} \frac{2}{(c + \sqrt{c^2 + 2c})}} \right). \quad (\text{B.14})$$

*Proof.* Consider  $\psi_{G,c}(\lambda)$  for  $c > 0$ . We consider breaking intrinsic time into geometric epochs,  $E_k = \{t : \eta^k \leq \det V_t < \eta^{k+1}\}$  for all  $k \geq 0$  and some  $\eta > 1$ . We assume that  $V_1 \succeq I$  so that  $\cup_{k \geq 0} E_k$  is a bonafide partition of the sample space. Let  $k(t)$  denote the unique  $k \in \mathbb{N}$  such that  $\eta^k \leq \det V_t \leq \eta^{k+1}$  and set

$$D_t = \frac{1}{2} \log\left(\frac{\det V_t}{\det U_0}\right) + 1 + \log(1/\delta_{k(t)}). \quad (\text{B.15})$$

By definition,  $k(t) \leq \log_\eta \det V_t$ , so

$$\log(1/\delta_{k(t)}) = \log(\ell(k(t))/\delta) \leq \log(\ell(\log_\eta(\det V_t))) + \log(1/\delta),$$

which is where the iterated logarithm term will come from in the final bound. In epoch  $E_k$  we apply Theorem 4.1 with  $\delta = \delta_k$  and  $\lambda = \lambda_k$ , for some  $\delta_k, \lambda_k$  to be chosen later. This gives that with probability  $1 - \delta_k$ , for all  $t \in E_k$ ,

$$\begin{aligned} \|S_t\|_{V_t^{-1}} &\leq \frac{D_t}{g(\lambda_k)\lambda_k} + \frac{g(\lambda_k)\psi_{G,c}(\lambda_k)}{\lambda_k} \\ &= \frac{1}{g(\lambda_k)} \left( \frac{D_t}{\lambda_k} + \frac{g^2(\lambda_k)\psi_{G,c}(\lambda_k)}{\lambda_k} \right) \\ &\leq \frac{1}{g(\lambda_k)} \left( \frac{D_t}{\lambda_k} + \frac{\psi_{G,c}(\lambda_k)}{\lambda_k} \right) =: W_c(t, \lambda_k), \end{aligned}$$

where we've used that  $g(\lambda_k) \leq 1$ . Choose  $\delta_k = \delta/\ell(k)$ . Then,

$$\begin{aligned} \mathbb{P}(\exists t \geq 1 : \|S_t\|_{V_t^{-1}} \geq W_c(t, \lambda_{k(t)})) &= \mathbb{P}\left(\bigcup_{k \geq 0} \{\exists t \in E_k : \|S_t\|_{V_t^{-1}} \geq W_c(t, \lambda_k)\}\right) \\ &\leq \sum_{k \geq 0} \mathbb{P}(\exists t \in E_k : \|S_t\|_{V_t^{-1}} \geq W_c(t, \lambda_k)) \\ &\leq \sum_{k \geq 0} \frac{\delta}{\ell(k)} = \delta. \end{aligned}$$

It remains to choose  $\lambda_k$  and analyze the width of  $W_c(t, \lambda_{k(t)})$ . Set

$$B_k = \frac{1}{2} \log \left( \frac{\eta^k}{\det U_0} \right) + 1 + \log(1/\delta_k),$$

and consider setting

$$\lambda_k = \max \left\{ \psi_{G,c}^{-1}(1/\rho), \frac{\sqrt{2B_k}}{1 + c\sqrt{2B_k}} \right\}, \quad \text{where } \rho = \gamma_{\min}(U_0). \quad (\text{B.16})$$

Observe that  $x/(1+cx) < 1/c$  for all  $x \geq 0$ , so  $\lambda_k < \lambda_{\max}$ . Therefore,  $\lambda_k \in [\psi_{G,c}^{-1}(1/\rho), \lambda_{\max}]$  and is a legal choice in Theorem 4.1.

Next we want to bound  $B_{k(t)}$  and relate it to  $D_t$ . Note that  $B_k$  is increasing in  $k$ , and  $\lambda_k$  is increasing in  $B_k$  hence also in  $k$ . For  $t \in E_k$  we have  $B_k \leq D_t \leq B_{k+1}$  by construction. To relate  $B_{k+1}$  and  $B_k$  we use the following lemma.

**Lemma B.2.** *Define*

$$\alpha_{\delta,\eta} := 1 + \frac{\log(\eta)}{1 + \log(1/\delta) - \log(\eta)/2} + \sup_k \log(\ell(k+1)/\ell(k)). \quad (\text{B.17})$$

If  $\eta < (e/\delta)^2$ , then  $B_{k+1} \leq \alpha_{\delta,\eta} B_k$ .

*Proof.* Write

$$\begin{aligned} B_{k+1} &= \frac{1}{2} \left( \log \left( \frac{\eta^{k+1}}{\det U_0} \right) \right) + 1 + \log(\ell(k+1)/\delta) \\ &= \frac{1}{2} \left( \log(\eta) + \log \left( \frac{\eta^k}{\det U_0} \right) \right) + 1 + \log \left( \frac{\ell(k+1)\ell(k)}{\ell(k)\delta} \right) \\ &= B_k + \frac{1}{2} \log \eta + \log(\ell(k+1)/\ell(k)) \\ &\leq B_k + \frac{1}{2} \log(\eta) + s, \end{aligned} \quad (\text{B.18})$$

where  $s = \sup_k \log(\ell(k+1)/\ell(k))$ . We claim that  $\log(\eta) \leq u B_k$  for some constant  $u$  and all  $k$ . Note, however, that we need only consider those  $k$  such that  $E_k$  actually occurs, i.e., the minimum  $k$  we need to consider satisfies  $\eta^k \leq \det V_1 < \eta^{k+1}$ . That is, we may assume that

$$\frac{\eta^k}{\det U_0} > \frac{\det V_1}{\eta \det U_0}.$$

Therefore,

$$\begin{aligned}
uB_k &= \frac{u}{2} \log \left( \frac{\eta^k}{\det U_0} \right) + u + u \log(1/\delta_k) \\
&\geq \frac{u}{2} \log \left( \frac{\det V_1}{\eta \det U_0} \right) + u + u \log(1/\delta) \\
&\geq \frac{u}{2} \log \left( \frac{1}{\eta} \right) + u + u \log(1/\delta),
\end{aligned}$$

where we've used that  $V_1 \succeq U_0$  by assumption. Therefore, to have  $\log(\eta) \leq uB_k$  it suffices that  $\log(\eta) \leq u \log(1/\eta)/2 + u + u \log(1/\delta)$ , i.e.,

$$u \geq \frac{\log(\eta)}{1 + \log(1/\delta) - \log(\eta)/2}, \quad (\text{B.19})$$

where we are assured that the denominator on the right hand side is greater than 0 since  $\eta < (e/\delta)^2$ . Set  $u^\circ$  to be the right hand side of (B.19). Then, from (B.18) we have

$$B_{k+1} \leq B_k + \frac{u^\circ}{2} B_k + s \leq (1 + u^\circ + s) B_k = \alpha_{\delta, \eta} B_k.$$

In summary, for  $\eta < (e/\delta)^2$ , we have that  $B_k \leq D_t \leq \alpha_{\delta, \eta} B_k$  for all  $t \in E_k$ , which completes the proof.  $\blacksquare$

Since  $\lambda_k$  is increasing in  $B_k$ , Lemma B.2 implies that

$$\frac{\sqrt{2D_t/\alpha_{\delta, \eta}}}{1 + c\sqrt{2D_t/\alpha_{\delta, \eta}}} \leq \frac{\sqrt{2B_{k(t)}}}{1 + c\sqrt{2B_{k(t)}}} \leq \frac{\sqrt{2D_t}}{1 + c\sqrt{2D_t}}. \quad (\text{B.20})$$

Assume for the moment that  $\lambda_{k(t)} = \frac{\sqrt{2B_{k(t)}}}{1 + c\sqrt{2B_{k(t)}}}$ . Note that  $\lambda \mapsto \psi_{G,c}(\lambda)/\lambda = \lambda/[2(1 - c\lambda)]$  is also an increasing function of  $\lambda$ . Using the lower bound in (B.20) in the first term and the upper bound in the second, we have

$$\begin{aligned}
\frac{D_t}{\lambda_{k(t)}} + \frac{\psi_{G,c}(\lambda_{k(t)})}{\lambda_{k(t)}} &= \frac{D_t}{\lambda_{k(t)}} + \frac{\lambda_{k(t)}}{2(1 - c\lambda_{k(t)})} \\
&\leq \frac{(1 + c\sqrt{\frac{2D_t}{\alpha_{\delta, \eta}}})\sqrt{D_t}}{\sqrt{2/\alpha_{\delta, \eta}}} + \frac{\sqrt{2D_t}}{2(1 + c\sqrt{2D_t})(1 - c\frac{\sqrt{2D_t}}{1 + c\sqrt{2D_t}})} \\
&= \sqrt{\frac{\alpha_{\delta, \eta} D_t}{2}} + cD_t + \sqrt{\frac{D_t}{2}} \\
&= cD_t + \left( \sqrt{\frac{\alpha_{\delta, \eta}}{2}} + \frac{1}{\sqrt{2}} \right) \sqrt{D_t}.
\end{aligned}$$

Meanwhile, if  $\lambda_{k(t)} = \psi_{G,c}^{-1}(1/\rho)$ , then we may still use the lower bound in (B.20) and we

obtain

$$\begin{aligned}
\frac{D_t}{\lambda_{k(t)}} + \frac{\psi_{G,c}(\lambda_{k(t)})}{\lambda_{k(t)}} &= \frac{D_t}{\lambda_{k(t)}} + \frac{1/\rho}{\psi_{G,c}^{-1}(1/\rho)} \\
&\leq \frac{(1 + c\sqrt{\frac{2D_t}{\alpha_{\delta,\eta}}})\sqrt{D_t}}{\sqrt{2/\alpha_{\delta,\eta}}} + \frac{c + \sqrt{c^2 + 2\rho}}{2\rho} \\
&= \sqrt{\frac{\alpha_{\delta,\eta}D_t}{2}} + cD_t + \frac{c + \sqrt{c^2 + 2\rho}}{2\rho},
\end{aligned}$$

where we've used that

$$\psi_{G,c}^{-1}(u) = \frac{2}{c + \sqrt{c^2 + 2/u}}.$$

We can summarize both cases with the bound

$$\frac{D_t}{\lambda_{k(t)}} + \frac{\psi_{G,c}(\lambda_{k(t)})}{\lambda_{k(t)}} \leq \sqrt{\frac{\alpha_{\delta,\eta}D_t}{2}} + cD_t + \max\left\{\frac{c + \sqrt{c^2 + 2\rho}}{2\rho}, \sqrt{D_t/2}\right\}. \quad (\text{B.21})$$

To bound the width of  $W_c(t, \lambda_{k(t)})$  it remains to analyze  $g(\lambda_{k(t)})$ . We claim that

$$\psi_{G,c}^{-1}(\lambda_k) \nearrow \frac{2}{c + \sqrt{c^2 + 2c}}. \quad (\text{B.22})$$

To see this, note that as  $D_k \rightarrow \infty$ ,  $\lambda_k \nearrow 1/c$ . Moreover,

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

is strictly increasing in  $u$ . (B.22) follows. Further, notice that

$$\alpha_t = \sup_{\theta \in \mathbb{R}^d} \frac{\langle \theta, U_0 \theta \rangle}{\langle \theta, V_t \theta \rangle} \leq \frac{\sup_{\theta \in \mathbb{S}^{d-1}} \langle \theta, U_0 \theta \rangle}{\inf_{\vartheta \in \mathbb{S}^{d-1}} \langle \vartheta, V_t \vartheta \rangle} = \frac{\gamma_{\max}(U_0)}{\gamma_{\min}(V_t)}.$$

Therefore, for  $t \in E_k$ ,

$$\begin{aligned}
g(\lambda_k) &= \sqrt{\alpha_t \psi_{G,c}^{-1}(\lambda_k) + 1 - \alpha_t} - \sqrt{\alpha_t \psi_{G,c}^{-1}(\lambda_k)} \\
&\geq 0 \vee \left( \sqrt{1 - \alpha_t} - \sqrt{\alpha_t \psi_{G,c}^{-1}(\lambda_k)} \right) \\
&\geq 0 \vee \left( \sqrt{1 - \frac{\gamma_{\max}(U_0)}{\gamma_{\min}(V_t)}} - \sqrt{\frac{\gamma_{\max}(U_0)}{\gamma_{\min}(V_t)} \frac{2}{(c + \sqrt{c^2 + 2c})}} \right) \\
&= H_c(U_0, V_t).
\end{aligned}$$

We have thus shown that

$$W_c(t, \lambda_{k(t)}) = \frac{1}{\lambda_{k(t)}} \left( \frac{D_t}{\lambda_{k(t)}} + \frac{\psi_{G,c}(\lambda_{k(t)})}{\lambda_{k(t)}} \right) \leq \frac{cD_t + \sqrt{\frac{\alpha_{\delta,\eta}}{2}}D_t + \max\left\{\frac{c + \sqrt{c^2 + 2\rho}}{2\rho}, \sqrt{D_t/2}\right\}}{H_c(U_0, V_t)},$$

which completes the proof. ■

## B.4 Proof of Proposition 4.7

We begin with some properties of the convex conjugate.

**Lemma B.3.** *Let  $\psi : [0, \lambda_{\max}) \rightarrow \mathbb{R}_{\geq 0}$  be CGF-like and  $\psi^* : [0, u_{\max}) \rightarrow \mathbb{R}_{\geq 0}$  denote the Legendre-Fenchel transform of  $\psi$ . For any  $a \in \mathbb{R}$ , let*

$$\lambda^*(a) = \operatorname{argmax}_{\lambda \in [0, \lambda_{\max})} \lambda a - \psi(\lambda).$$

Then,

- (1). For all  $a \in \operatorname{Im}(\psi')$ ,  $a = \psi'(\lambda^*(a))$  and  $\lim_{a \rightarrow \infty} \lambda^*(a) = \lambda_{\max}$ ,
- (2). For all  $a \in \operatorname{Im}(\psi')$ ,  $\lambda^*(a) = (\psi')^{-1}(a)$ , and
- (3). If  $\psi'''(\lambda) \geq 0$ , then  $\lambda^*(ra) \leq r\lambda^*(a)$  for all  $r \geq 1$ .

Additionally, if  $\psi(\lambda)/\lambda^2$  is increasing then,

- (4). For all  $\lambda \in [0, \lambda_{\max})$ ,  $\psi'(\lambda) \geq 2\psi(\lambda)/\lambda$  and
- (5). For all  $u \in [0, u_{\max})$ ,  $\psi^*(u) \geq u\lambda^*(u)/2$ .

*Proof.* We begin with (1). For fixed  $a \in \operatorname{Im}(\psi')$ , let  $f(x) = ax - \psi(x)$ . Since  $\psi$  is convex,  $f$  is concave, so  $f$  is maximized at  $x^*$  satisfying  $a = \psi'(x^*)$  ( $x^*$  exists and is unique because  $a \in \operatorname{Im}(\psi')$  and  $\psi'$  is strictly increasing). But  $x^* = \lambda^*(a)$  by definition, so  $a = \psi'(\lambda^*(a))$ . Now, as  $a \rightarrow \infty$ ,  $\lambda = \lambda^*(a)$  is chosen to maximize  $a - \psi'(\lambda)$ , which grows as  $\lambda \rightarrow \lambda_{\max}$  since  $\psi'$  is strictly increasing (since  $\psi$  is strictly convex). (2) now also follows: since  $\psi'$  is strictly increasing it is invertible, so may invert (1) to get (2). For (3), note that  $\psi''' \geq 0$  iff  $\psi''$  is increasing iff  $\psi'$  is convex, hence its inverse  $\lambda^*$  is concave (and nonnegative). Therefore, for all  $x, y$ , we have  $\lambda^*(\alpha x + (1 - \alpha)y) \geq \alpha\lambda^*(x) + (1 - \alpha)\lambda^*(y)$  for all  $\alpha \in [0, 1]$ . Consider  $y = 0$ , so  $\lambda^*(\alpha x) \geq \alpha\lambda^*(x) + (1 - \alpha)\lambda^*(0) \geq \alpha\lambda^*(x)$ . Taking  $x = ra$  and  $\alpha = 1/r \in [0, 1]$ , the result follows. For (4), if  $h(\lambda) = \psi(\lambda)/\lambda^2$  is increasing, then  $h'(\lambda) \geq 0$ , i.e.,

$$0 \leq \frac{\lambda^2 \psi'(\lambda) - 2\lambda \psi(\lambda)}{\lambda^4},$$

thus implying that  $\lambda \psi'(\lambda) - 2\psi(\lambda) \geq 0$ . Finally, (5) involves combining the definition of  $\lambda^*$  with (1) and (4) to notice that

$$\begin{aligned} \psi^*(u) &= u\lambda^*(u) - \psi(\lambda^*(u)) \geq u\lambda^*(u) - \frac{\lambda^*(u)\psi'(\lambda^*(u))}{2} \\ &= u\lambda^*(u) - \frac{\lambda^*(u)u}{2} = \frac{\lambda^*(u)u}{2}, \end{aligned}$$

which completes the proof. ■

Now let us return to the bound of Theorem 4.1. Let  $\rho = \gamma_{\min}(U_0)$ . For all  $\lambda \in [\psi^{-1}(1/\rho), \lambda_{\max})$ , with probability  $1 - \delta$ ,

$$\lambda \|S_\tau\|_{V_\tau^{-1}} - \psi(\lambda) \leq \frac{\|U_0\|_{\text{op}}^{1/2} D_\tau(\delta)}{g_{\psi, \tau}(\lambda)} + \frac{g_{\psi, \tau}(\lambda) \psi(\lambda)}{\|U_0\|_{\text{op}}^{1/2}} - \psi(\lambda).$$

Since  $g_{\psi,\tau}(\lambda) \leq 1$ , if we assume that  $U_0 \succeq I_d$  (in which case  $\|U_0\|_{\text{op}} \geq 1$ ), then the above display implies that

$$\lambda \|S_\tau\|_{V_\tau^{-1}} - \psi(\lambda) \leq \frac{\|U_0\|_{\text{op}}^{1/2} D_\tau(\delta)}{g_{\psi,\tau}(\lambda)}. \quad (\text{B.23})$$

For  $j \in \mathbb{N}$  and some  $\eta > 1$  to be determined, consider the event  $E_j = \{\eta^j \leq \|S_\tau\|_{V_\tau^{-1}} < \eta^{j+1}\}$ . Let  $E_0 = \{0 \leq \|S_\tau\|_{V_\tau^{-1}} < 2\}$ . Note that  $E_0, E_1, \dots$  partitions the sample space. Our strategy is to apply Theorem 4.1 once on each event  $E_j$  with a different  $\delta = \delta_j$  and  $\lambda = \lambda_j$ , where the latter can be optimized as (roughly) a function of  $\|S_\tau\|_{V_\tau^{-1}}$  since we know how this quantity behaves on  $E_j$ . Rewriting (B.23) with  $\delta = \delta_j$  and  $\lambda = \lambda_j$  gives

$$\lambda_j \|S_\tau\|_{V_\tau^{-1}} - \psi(\lambda_j) \leq h_\tau(\lambda_j) D_\tau(\delta_j), \quad (\text{B.24})$$

where  $h_\tau(\lambda) = \|U_0\|_{\text{op}}^{1/2} / g_{\psi,\tau}(\lambda)$ . Now, conditioning on  $E_j$ , we have

$$\begin{aligned} \lambda_j \eta^{j+1} - \psi(\lambda_j) &= \lambda_j \|S_\tau\|_{V_\tau^{-1}} - \psi(\lambda_j) + \lambda_j (\eta^{j+1} - \|S_\tau\|_{V_\tau^{-1}}) \\ &\leq h_\tau(\lambda_j) D_\tau(\delta_j) + \lambda_j (\eta^{j+1} - \eta^j). \end{aligned} \quad (\text{B.25})$$

We want to choose an appropriate  $\lambda_j$  on the event  $E_j$ . As in Lemma B.3, let

$$\lambda^*(a) = \operatorname{argmax}_{\lambda \in (0, \lambda_{\max})} \lambda a - \psi(\lambda),$$

and take

$$\lambda_j^* = \max \{ \lambda^*(\eta^{j+1}), \psi^{-1}(1/\rho) \}. \quad (\text{B.26})$$

Let us suppose for now that  $\lambda_j^* = \lambda^*(\eta^{j+1})$  so that  $\lambda_j^* \eta^{j+1} - \psi(\lambda_j^*) = \psi^*(\eta_{j+1})$  and we can rearrange (B.25) to read

$$\psi^*(\|S_\tau\|_{V_\tau^{-1}}) \leq \psi^*(\eta^{j+1}) \leq h_\tau(\lambda_j^*) D_\tau(\delta_j) + \eta^j \lambda_j^* (\eta - 1), \quad (\text{B.27})$$

where we've used that  $\|S_\tau\|_{V_\tau^{-1}} \leq \eta^{j+1}$  on  $E_j$  and the fact that  $\psi^*$  is increasing. Now, we claim that

$$\eta^j \lambda_j^* (\eta - 1) \leq \frac{\psi^*(\eta^j)}{2}.$$

Let  $s = \eta^j$  and define  $R(s) = s \lambda_j^* (\eta - 1) / \psi^*(s)$ . We want to show that  $R(s) \leq 1/2$ . Using Lemma B.3, we have  $\psi^*(s) \geq s \lambda^*(s) / 2$ . Moreover, if  $\psi''' \geq 0$ , then  $\lambda^*(\eta s) \leq \eta \lambda^*(s)$ , so

$$R(s) \leq \frac{2s \lambda^*(\eta s) (\eta - 1)}{s \lambda^*(s)} \leq 2\eta (\eta - 1),$$

for any  $\eta \geq 1$ . We must thus choose  $\eta$  such that  $\eta > 1$  (required in the definition of  $E_j$ ) and  $2\eta(\eta - 1) \leq 1/2$ , which holds for any  $\eta$  satisfying

$$1 < \eta \leq \frac{1 + \sqrt{2}}{2} \approx 1.207.$$

For such  $\eta$  (B.27) implies

$$\psi^*(\|S_\tau\|_{V_\tau^{-1}}) \leq h_\tau(\lambda_j^*) D_\tau(\delta_j) + \frac{\psi^*(\eta^j)}{2} \leq h_\tau(\lambda_j^*) D_\tau(\delta_j) + \frac{\psi^*(\|S_\tau\|_{V_\tau^{-1}})}{2},$$



whence  $\psi^*(\|S_\tau\|_{V_\tau^{-1}}) \leq 2h_\tau(\lambda_j^*)D_\tau(\delta_j)$ , i.e.,

$$\|S_\tau\|_{V_\tau^{-1}} \leq (\psi^*)^{-1} \left( 2h_\tau(\lambda_j^*)D_\tau(\delta_j) \right). \quad (\text{B.28})$$

Now we return to the scenario when  $\lambda_j^* = \psi^{-1}(1/\rho)$  in (B.26). In this case, we have

$$\|S_\tau\|_{V_\tau^{-1}} \leq \frac{h_\tau(\psi^{-1}(1/\rho))D_\tau(\delta_j) + 1/\rho}{\psi^{-1}(1/\rho)}.$$

We have thus shown that

$$\mathbb{P} \left( \|S_\tau\|_{V_\tau^{-1}} \geq B_\tau | E_j \right) \leq \delta,$$

where  $B_\tau$  is the piecewise boundary

$$B_\tau = \begin{cases} (\psi^*)^{-1} \left( 2h_\tau(\lambda^*(\eta^{j(\tau)+1}))D_\tau(\delta_{j(\tau)}) \right), & \text{if } \lambda^*(\eta^{j(\tau)+1}) \geq \psi^{-1}(1/\rho), \\ \frac{h_\tau(\psi^{-1}(1/\rho))D_\tau(\delta_j) + 1/\rho}{\psi^{-1}(1/\rho)}, & \text{otherwise,} \end{cases} \quad (\text{B.29})$$

and  $j(\tau)$  is the unique  $j \in \mathbb{N}_0$  such that  $\|S_\tau\|_{V_\tau^{-1}} \in [\eta^j, \eta^{j+1})$ . Therefore,

$$\begin{aligned} \mathbb{P} \left( \|S_\tau\|_{V_\tau^{-1}} \geq B_\tau \right) &= \sum_{j \geq 0} \mathbb{P} \left( \|S_\tau\|_{V_\tau^{-1}} \geq B_\tau | E_j \right) \mathbb{P}(E_j) \\ &\leq \sum_{j \geq 0} \mathbb{P} \left( \|S_\tau\|_{V_\tau^{-1}} \geq B_\tau | E_j \right) \\ &\leq \sum_{j \geq 0} \delta_j = \delta \sum_{j \geq 0} \ell^{-1}(j) = \delta. \end{aligned}$$

Now, by Lemma B.3,  $\lambda^*(a)$  is increasing in  $a$ . In particular,

$$\lambda^*(a) \xrightarrow{a \rightarrow \infty} \lambda_{\max}.$$

Therefore, for large enough  $t$ , as long as  $\psi^{-1}(1/\rho) < \lambda_{\max}$ , we have

$$B_t = (\psi^*)^{-1} \left( 2h_\tau(\lambda^*(\eta^{j(t)+1}))D_\tau(\delta_{j(t)}) \right).$$

We may thus write that with probability  $1 - \delta$ ,

$$\|S_\tau\|_{V_\tau^{-1}} \leq B_\tau \lesssim (\psi^*)^{-1} \left( 2h_\tau(\lambda^*(\eta^{j(\tau)+1}))D_\tau(\delta_{j(\tau)}) \right). \quad (\text{B.30})$$

It remains to bound  $h_\tau(\lambda^*(\eta^{j(\tau)+1}))$ . When  $\lambda_{\max} < \infty$ , we have that

$$\lim_{j \rightarrow \infty} \psi^{-1}(\lambda^*(\eta^{j+1})) \nearrow \psi^{-1}(\lambda_{\max}).$$

To see this, again start with the fact that  $\lambda^*$  is increasing. Then, since  $\psi$  is strictly increasing,  $\psi^{-1}$  is continuous, so

$$\lim_{j \rightarrow \infty} \psi^{-1}(\lambda^*(\eta^{j+1})) = \psi^{-1} \left( \lim_{j \rightarrow \infty} \lambda^*(\eta^{j+1}) \right) = \psi^{-1}(\lambda_{\max}).$$

That the sequence *increases* to its limit follows since  $\psi^{-1}$  is increasing. As was done in the proof of Proposition 4.3, notice that  $\alpha_t \leq \gamma_{\max}(U_0)/\gamma_{\min}(V_t)$ . Then,

$$\begin{aligned} g_{\psi_\tau}(\lambda_j^*) &= \sqrt{\alpha_t \psi^{-1}(\lambda^*(\eta^{j(\tau)+1})) + 1 - \alpha_t} - \sqrt{\alpha_t \psi^{-1}(\lambda^*(\eta^{j(\tau)+1}))} \\ &\geq \sqrt{1 - \alpha_\tau} - \sqrt{\alpha_t \psi^{-1}(\lambda_{\max})} \\ &\geq \sqrt{1 - \frac{\gamma_{\max}(U_0)}{\gamma_{\min}(V_\tau)}} - \sqrt{\frac{\gamma_{\max}(U_0) \psi^{-1}(\lambda_{\max})}{\gamma_{\min}(V_\tau)}} \\ &\gtrsim \sqrt{1 - \frac{1}{\gamma_{\min}(V_\tau)}}, \end{aligned}$$

hence

$$h_\tau(\lambda^*(\eta^{j(\tau)+1})) = \frac{\|U_0\|_{\text{op}}^{1/2}}{g_{\psi,\tau}(\lambda^*(\eta^{j(\tau)+1}))} \lesssim \frac{1}{\sqrt{1 - 1/\gamma_{\min}(V_\tau)}}.$$

Returning to (B.30), we have that with probability  $1 - \delta$ ,

$$\begin{aligned} \|S_\tau\|_{V_\tau^{-1}} &\lesssim (\psi^*)^{-1} \left( \frac{1}{\sqrt{1 - 1/\gamma_{\min}(V_\tau)}} \left\{ \log(\det V_\tau) + \log\left(\frac{\ell(j(\tau))}{\delta}\right) \right\} \right) \\ &\lesssim (\psi^*)^{-1} \left( \frac{1}{\sqrt{1 - 1/\gamma_{\min}(V_\tau)}} (\log(\det V_\tau) + \log(\log_\eta(\|S_\tau\|_{V_\tau^{-1}}/\delta)) \right), \end{aligned}$$

where we've used that  $\ell(j(\tau)) \asymp j(\tau)^2$  and  $\log(j(\tau)^2) \asymp \log(\log_\eta \|S_\tau\|_{V_\tau^{-1}})$ , which completes the proof.

## B.5 Proof of Corollary 4.8

Since  $\psi$  is CGF-like, so too is  $\psi^*$  [60, Proposition A.1]. In particular,  $\psi^*$  is strictly convex and increasing, implying that  $\varphi$  is strictly increasing and concave. It follows that for all  $x, y$ ,

$$\varphi(x + y) - \varphi(x) \leq \varphi'(x)y. \quad (\text{B.31})$$

(A concave function is bounded by its first order Taylor approximation.) Set

$$A_\tau = \frac{\log(\det V_\tau) + \log(1/\delta)}{\sqrt{1 - 1/\gamma_{\min}(V_\tau)}}, \quad B_\tau = \frac{\log \log(\|S_\tau\|_{V_\tau^{-1}})}{\sqrt{1 - 1/\gamma_{\min}(V_\tau)}}.$$

Since  $1 - 1/\gamma_{\min}(V_\tau) \leq 1$  we can bound  $A_\tau$  as  $A_\tau \geq A_{\min} := \log(1/\delta)$ . Further, since  $V_\tau \succeq U_0$ ,  $\gamma_{\min}(V_\tau) \geq \gamma_{\min}(U_0) := \rho$ , hence

$$B_\tau \leq \frac{\log \log(\|S_\tau\|_{V_\tau^{-1}})}{\sqrt{1 - 1/\rho}} \leq \frac{\|S_\tau\|_{V_\tau^{-1}}}{\sqrt{1 - 1/\rho}}.$$

Take  $x = A_\tau$  and  $y = B_\tau$  in (B.31), we obtain that the right hand side of (4.9) is bounded as

$$\varphi(A_\tau + B_\tau) \leq \varphi(A_\tau) + \varphi'(A_\tau)B_\tau = \varphi(A_\tau) + \frac{B_\tau}{(\psi^*)'(\varphi(A_\tau))}.$$

Since  $\psi^*$  is strictly increasing and convex,  $(\psi^*)'$  is also increasing. Therefore the term  $(\psi^*)'(\varphi(x))$  is increasing in  $x$ , and

$$\varphi'(A_\tau) \leq \frac{1}{(\psi^*)'(\varphi(A_\tau))} \leq \frac{1}{(\psi^*)'(\varphi(\log(1/\delta)))} = \varphi'(\log(1/\delta)),$$

whence

$$\varphi(A_\tau + B_\tau) \leq \varphi(A_\tau) + \frac{\varphi'(\log(1/\delta))\|S_\tau\|_{V_\tau^{-1}}}{\sqrt{1 - 1/\rho}}.$$

Rearranging gives the desired result.

## B.6 Proof of Lemma 5.1

Our goal is to upper bound the term  $\log \mathbb{E}_{t-1} \exp(\lambda \langle \theta, X \rangle)$  in (5.1). Since  $\log x \leq x - 1$  for all  $x > 0$ , we have

$$\log \mathbb{E}_{t-1} \exp(\lambda \langle \theta, X \rangle) \leq \mathbb{E}_{t-1} \exp(\lambda \langle \theta, X \rangle) - 1 = \mathbb{E}_{t-1} \psi_{P,1}(\lambda \langle \theta, X \rangle),$$

where the final inequality follows since  $\mathbb{E}_{t-1} \langle \theta, X \rangle = 0$  (recall that  $\psi_{P,1}(x) = e^x - x - 1$ ). Now, notice that  $f(c) = \psi_{P,c}(x)$  is an increasing function of  $c$  (not of  $x$ !). Since  $\langle \theta, X_t \rangle \leq b$  almost surely, we have

$$\frac{e^{\lambda \langle \theta, X_t \rangle} - \lambda \langle \theta, X_t \rangle - 1}{\langle \theta, X_t \rangle^2} \leq \psi_{P,b}(\lambda).$$

Rearranging and taking expectations gives

$$\mathbb{E}_{t-1} \psi_{P,1}(\lambda \langle \theta, X_t \rangle) \leq \mathbb{E}_{t-1} [\langle \theta, X_t \rangle^2] \psi_{P,b}(\lambda).$$

This proves that

$$N_t^{\text{Ben}}(\theta) = \prod_{k \leq t} \exp \{ \lambda \langle \theta, X_k \rangle - \psi_{P,b}(\lambda) \mathbb{E}_{k-1} [\langle \theta, X_k \rangle^2] \} \leq N_t^B(\theta),$$

hence  $(N_t^{\text{Ben}}(\theta))$  is upper bounded by a nonnegative martingale. Noticing that  $N_t^{\text{Ben}}(\theta) = \exp \{ \lambda \langle \theta, S_t \rangle - \psi_{P,b}(\lambda) \langle \theta, V_t \theta \rangle \}$  proves that  $(S_t, V_t)$  is a sub- $\psi$  process for  $S_t = \sum_{k \leq t} X_k$ ,  $V_t = \sum_{k \leq t} \mathbb{E}_{k-1} X_k X_k^\top$ . Adding any PSD matrix  $U_0$  to  $V_t$  can only make the process smaller, which completes the proof.

## B.7 Proof of Lemma 5.2

In the proof of Lemma 5.1, we showed that  $\log \mathbb{E}_{t-1} \exp(\lambda \langle \theta, X \rangle) \leq \mathbb{E}_{t-1} \psi_{P,1}(\lambda \langle \theta, X \rangle)$ . Using the Taylor expansion  $e^x = \sum_{q \geq 0} x^q / q!$  it follows that  $\psi_{P,1}(x) = \sum_{q \geq 2} \frac{x^q}{q!}$ , hence

$$\mathbb{E}_{t-1} \psi_{P,1}(\lambda \langle \theta, X_t \rangle) = \sum_{q \geq 2} \frac{\lambda^q \mathbb{E}_{t-1} [\langle \theta, X_t \rangle^q]}{q!}.$$

Using assumption (5.2) on the moments of  $\langle \theta, X \rangle$ , we have

$$\sum_{q \geq 2} \frac{\lambda^q \mathbb{E}_{t-1} [\langle \theta, X_t \rangle^q]}{q!} \leq \sum_{q \geq 2} \frac{\lambda^q c^{q-2} \mathbb{E}_{t-1} [\langle \theta, X_t \rangle^2]}{2} \leq \psi_{G,c}(\lambda) \mathbb{E}_{t-1} [\langle \theta, X_t \rangle^2],$$

if  $\lambda < 1/c$ . This implies that the process

$$N_t^{\text{Bern}}(\theta) = \prod_{k \leq t} \exp \{ \lambda \langle \theta, X_k \rangle - \psi_{G,c}(\lambda) \langle \theta, \mathbb{E}_{k-1} X_k X_k^\top \theta \rangle \},$$

is upper bounded by  $N_t^B(\theta)$  in (5.1) and hence is itself a nonnegative supermartingale. As in the proof of Lemma 5.1, noticing that we can rewrite  $N_t^{\text{Bern}}(\theta)$  as  $N_t^{\text{Bern}}(\theta) = \exp \{ \lambda \langle \theta, S_t \rangle - \psi_{G,c}(\lambda) \langle \theta, V_t \theta \rangle \}$  for  $V_t = \sum_{k \leq t} \mathbb{E}_{k-1} X_k X_k^\top$  shows that  $(S_t, V_t)$  is sub- $\psi_{G,c}$ , and adding any PSD matrix to  $V_t$  maintains this property.

## C Simulation Details

Code may be found at <https://github.com/bchugg/sn-concentration>.

**Figure 1** In both figures we take  $d = 10$ ,  $U_0 = I_d$ , and  $V_t = U_0 + \sum_{k \leq t} X_k X_k^\top$  where  $X_k$  is chosen based on UCB in a linear bandit setting. In particular,  $X_k$  is the eigenvector corresponding to the maximum eigenvalue of  $V_{k-1}^{-1}$ .

**Figure 2.** We take  $U_0 = I_d$  and  $d = 20$ . For the first 100 time steps, we make a full rank update in order to grow the minimum eigenvalue sufficiently. These updates are dampened based on the size of the eigenvalue, however, so there is still anisotropic growth of the matrix. The dampening factor is 1 for the top eigenvector (i.e., corresponding to the largest eigenvalue) and 0.1 for the bottom, with a linear interpolation between 1 and 0.1 in the remaining directions.

After 100 timesteps, we make rank  $k$  updates to the matrix (in the directions corresponding to the top  $k$  eigenvalues). That is,

$$V_t = V_{t-1} + \sum_{j \leq k} u_j u_j^\top,$$

where  $u_1$  is the eigenvector corresponding to the largest eigenvalue,  $u_2$  that corresponding to the second largest eigenvalue, and so on. For small growth of the determinant (Figure 2a) we use  $k = 1$ , for moderate growth (Figure 2b) we use  $k = 4$  and for large growth (Figure 2c) we use  $k = 8$ .

We compare Theorem 4.3 to Corollary 2 in Whitehouse et al. [60], which is tailored explicitly to sub- $\psi_{G,c}$  processes. It states that if  $(S_t, V_t)$  is a sub- $\psi_{G,c}$  process with  $V_t \succeq \rho I_d$  for all  $t$ , then, with probability  $1 - \delta$ , for all  $t$ ,

$$\|S_t\|_{V_t^{-1}} \leq \frac{1}{1 - \epsilon} \sqrt{4(M_1(t) + M_2(t) + M_3(t))} + \frac{c\eta}{\sqrt{\gamma_{\min}(V_t)}} (M_1(t) + M_2(t) + M_3(t)),$$

where

$$\begin{aligned} M_1(t) &= B \log \left( A \log_{\eta} \left( \frac{\gamma_{\max}(V_t)}{\rho} \right) \right), \\ M_2(t) &= \log \left( \frac{1}{\delta(1 - 1/\beta)} \right), \\ M_3(t) &= (d + 1) \log \left( \frac{\beta \sqrt{\kappa(V_{\tau})}}{\epsilon} \right). \end{aligned}$$

We take  $\beta = 2$ ,  $\eta = 2$ , and  $\epsilon = 1/2$ , as suggested in their paper. In all plots, we omit the first 200 time steps because our bound blows up when  $\gamma_{\min}(V_t)$  is too small. As we say in the main paper, this is drawback of our bound from which the bound of Whitehouse et al. [60] does not suffer.

**Figure 3.** For Figure 3 we generate  $V_t$  in the same way as in Figure 2. We use  $k = 5$ ,  $d = 20$ , and  $U_0 = I_d$ . We use  $c = 1/4$  as the sub-gamma parameter. Note that  $\psi_{G,1/4}^{-1}(1) \approx 1.54$  so our choices of  $\lambda$  are all legal.