A variational approach to dimension-free self-normalized concentration

Anonymous authors
Paper under double-blind review

Abstract

We study the self-normalized concentration of vector-valued stochastic processes. We focus on bounds for "sub-ψ" processes, a well-known and quite general class of process that encompasses a wide variety of well-known tail conditions (including sub-exponential, sub-Gaussian, sub-gamma, sub-Poisson, and several heavy-tailed settings without a moment generating function such as symmetric or bounded 2nd or 3rd moments). Our results recover and generalize the influential bound of de la Peña et al. (2004) (proved again in Abbasi-Yadkori et al. 2011) in the sub-Gaussian case. Further, we fill a gap in the literature between determinant-based bounds and more recent bounds based on condition numbers. As applications we prove a Bernstein inequality for random vectors satisfying a moment condition (a more general condition 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.

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 (de la Peña et al., 2004; de la Peña et al., 2007; de la Peña et al., 2008; de la Peña & Pang, 2009; de la Peña et al., 2009). Much of this work was focused on scalar-valued processes until de la Peña et al. (2009) 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. (2004) provide bounds—much earlier than the oft-cited bound of Abbasi-Yadkori et al. (2011)—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. (2009) 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. (2025) gave self-normalized concentration bounds for vector-valued sub-ψ processes, a powerful and general abstraction first introduced by Howard et al. (2020) which captures many distributions of interest, including sub-Poisson, sub-exponential, sub-gamma, in addition to several heavy tailed settings where the moment generating function may not exist (see Section 2). While comprehensive, the bounds of Whitehouse et al. (2025) bounds 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-ψ processes?

This paper answers in the affirmative, providing dimension-free, determinant-based self-normalized concentration inequalities for vector-valued sub-ψ processes. Our techniques are based on the variational approach to concentration (sometimes called the PAC-Bayesian approach (Catoni & Giulini, 2017; 2018)), which has

	Condition	*Dimension-free*	*Dependence*
Faury et al. (2020)	Bounded		$\det V_t$
Zhou et al. (2021)	Bounded		UB on $\|S_t\|$
Ziemann (2025)	Bounded		$\det V_t$
Akhavan et al. (2025)**	Bounded	✓	$\det V_t$
Metelli et al. (2025)**	Bounded	✓	$\det V_t$
Kirschner et al. (2025)	Bounded		$\det V_t$
Kirschner et al. (2025)	Gaussian	✓	$\det V_t$
de la Peña et al. (2004)	sub-Gaussian	✓	$\det V_t$
Abbasi-Yadkori et al. (2011)	sub-Gaussian	✓	$\det V_t$
de la Peña et al. (2008)*	Gen. canonical	✓	$\det V_t$
Martinez-Taboada et al. (2026)**	Bernstein	✓	$\det V_t$
Whitehouse et al. (2025)	sub-ψ		$\kappa(V_t)$
This work	sub-ψ	✓	$\det V_t$

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). "Gen. canonical" refers to the generalized canonical assumption of de la Peña et al. (2008, Section 14.1.2). (*) indicates that the results are not in closed-form. (**) indicates that these works were concurrent to this one.

been successfully leveraged in recent years to study the concentration of random vectors and matrices (Zhivotovskiy, 2024; Oliveira, 2016; Nakakita et al., 2024; Ziemann, 2025; Chugg et al., 2025). 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.

Related work. We give a full treatment of related work in Appendix A. See Table 1 for a concise summary of previous bounds and their domain of applicability. Our main point of comparison throughout this work will be that of Whitehouse et al. (2025) who is the only other work to-date which studies self-normalized concentration of vector-valued sub-ψ processes in full generality. Whitehouse et al. (2025) show that if (S_t, V_t) is a sub-ψ process (to be defined shortly), then

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

where $\kappa(V_t) = \gamma_{\max}(V_\tau)/\gamma_{\min}(V_\tau)$ is the ratio of the largest to smallest eigenvalue of V_τ (the condition number) and ψ^* is the convex conjugate of ψ. Our bounds will be similar, but will have the factor $\log\det(V_t)$ instead of $d\log\kappa(V_t)$.

Contributions and outline. After defining sub-ψ processes and giving an overview of the variational approach in Section 2, Section 3 provides concentration results for general sub-ψ processes evolving in $\mathbb{R}^d$. In particular, Theorem 3.1 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 λ optimize the threshold at different (intrinsic) times.

Theorem 3.4 then employs a method known as *stitching* (Howard et al., 2021) which iteratively applies Theorem 3.1 with different parameters in different geometrically-spaced epochs of time, in order to give a bound which remains tight at all times. This procedure costs only an iterated logarithm term (this is similar to a doubling trick, chaining, peeling, etc.).

As an application, Section 4 provides the first dimension-free, self-normalized empirical Bernstein inequality for bounded random vectors. It is 'empirical' because it does not rely on a priori knowledge of the variance, but instead uses the empirical variance to adapt to distribution over time. In the spirit of more traditional applications, Appendix D applies our results to obtain self-normalized (non-empirical) Bennett and Bernstein inequalities.

We also study sub-Gaussian processes specifically in Appendix B, which both serves as an introduction to our method in a familiar setting, and also demonstrates that our technique precisely recovers the bounds of de la Peña et al. (2004) and Abbasi-Yadkori et al. (2011). This gives a sense of its scope and generality.

2 Preliminaries

Fix a filtered probability space $(\Omega, \mathcal{F} \equiv (\mathcal{F}_t)_{t\geq 0}, P)$. We assume that we are working in discrete time, so $\mathcal{F}$ is indexed by $\mathbb{N} \cup \{0\}$. We consider two stochastic processes $(S_t)_{t\geq 1}$ and $(V_t)_{t\geq 0}$ on $(\Omega, \mathcal{F}, P)$ 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; S is a mnemonic for sum, and V for variance. 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 $\theta \in \mathbb{R}^d$. Typically, we will also have $V_t \preceq V_{t+1}$, where $\preceq$ is the positive semidefinite partial (Loewner) order.

We are interested obtaining bounds on $\|S_\tau\|_{V_\tau^{-1}}$ for any $\mathcal{F}$-adapted stopping time τ 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 & Freedman (1973), a measure of the intrinsic time-scale of (S_t). This is formalized with the notion of a sub-ψ 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 $\mathcal{F}$ if it is $\mathcal{F}$-adapted and $\mathbb{E}[A_t|\mathcal{F}_{t-1}] \leq A_{t-1}$ for all $t \geq 1$.

Definition 2.1 (Sub-ψ process in $\mathbb{R}^d$)**.** Let $\psi : [0, \lambda_{\max}) \to \mathbb{R}_{\geq 0}$. 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 say that (S_t) is a sub-ψ process with variance proxy (V_t) if, for all $\theta \in \mathbb{S}^{d-1}$, all $\lambda \in [0, \lambda_{\max})$, and all $t \geq 0$,

$$\exp\left\{\lambda\langle\theta, S_t\rangle - \psi(\lambda)\langle\theta, V_t\theta\rangle\right\} \leq L_t^\lambda(\theta), \tag{2}$$

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

For scalar S_t and V_t, sub-ψ processes were originally proposed and studied by Howard et al. (2020) for scalar and matrix settings, generalizing and unifying a wide array of earlier work, such as de la Peña et al. (2004). The vector case has been recently studied by Whitehouse et al. (2025). Like both Howard et al. (2020) and Whitehouse et al. (2025), we assume that ψ is *CGF-like*, meaning that it is strictly convex and twice continuously differentiable, with $\lim_{\lambda\to 0^+} \mathrm{d}\psi(\lambda)/\mathrm{d}\lambda = 0 = \psi(0)$.

While the definition of a sub-ψ process may appear abstract at first glance, let us reassure the reader that the definition captures many familiar (and unfamiliar) examples, including certain heavy-tailed settings.

1. $\psi_N = \psi_N(\lambda) := \lambda^2/2$ for $\lambda \in [0, \lambda_{\max})$ results in a sub-exponential process (sub-Gaussian if $\lambda_{\max} = \infty$). If X_t is Σ_t-sub-Gaussian[1] then $S_t = \sum_{k\leq t} X_k$ is a sub-ψ_N process with $\lambda_{\max} = \infty$ and $V_t = \sum_{k\leq t} \Sigma_k$. Here (Σ_t) can be $\mathcal{F}$-adapted (not just predictable). As a heavy-tailed example, if $\mathbb{E}[\langle\theta, X_t\rangle|\mathcal{F}_{t-1}] < \infty$ for all $\theta \in \mathbb{S}^{d-1}$, then $S_t = \sum_{k\leq t} X_k$ is sub-ψ_N with $\lambda_{\max} = \infty$ and $V_t = \frac{1}{3}\sum_{k\leq t} X_k X_k^\intercal + \frac{2}{3}\sum_{k\leq t} \mathbb{E}[X_k X_k^\intercal|\mathcal{F}_{k-1}]$ (see Howard et al. (2020, Lemma 3f)). 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. (2007) implies that $S_t = \sum_{k\leq t} X_k$ and $V_t = \sum_{k\leq t} X_k X_k^\intercal$ define a sub-ψ_N process with $\lambda_{\max} = \infty$; note here that we did not require even a single moment to exist, let alone a moment generating function, in order to yield a sub-ψ_N process. As in the scalar case, bounded random vectors can also be shown to be sub-ψ_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; the terminology stems from the fact that $\psi_{G,c}$ is an *upper bound* on the CGF of a gamma random variable (Boucheron et al., 2013, Section 2.4). 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 D.2. Whitehouse et al. (2025) 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 $V_t = \sum_{j\leq t}(X_j X_j^\intercal + \mathbb{E}[\|X_j\|^3|\mathcal{F}_{j-1}]I_d)$. Moreover, Howard et al. (2020, Proposition 1) show that any twice differentiable, CGF-like ψ can be bounded by $a\psi_{G,c}$ for some $a, c > 0$. Thus, any sub-ψ process is sub-$\psi_{G,c}$ after scaling (though transforming to sub-gamma may cause some looseness in the resulting bounds).

[1] That is, $\mathbb{E}[\exp\{\langle\theta, X_t\rangle) - \psi_N(\lambda)\langle\theta, \Sigma_t\theta\rangle\}|\mathcal{F}_{t-1}] \leq 1$ for all $\theta \in \mathbb{S}^{d-1}$ with $\lambda_{\max} = \infty$. In the scalar case, X_t is σ_t-sub-Gaussian (conditional on the past) if $\mathbb{E}[\exp(\lambda X_t)|\mathcal{F}_{t-1}] \leq \exp(\psi_N(\lambda)\sigma_t^2)$ for all real λ.

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-exponential process, which is equivalent[2] up to constants to the sub-exponential condition mentioned under ψ_N; the terminology stems from the fact that $\psi_{E,c}$ is the CGF of a centered unit-rate negative-exponential random variable. Importantly, sums of bounded random vectors can be shown to be sub-exponential with an *adapted* (i.e. not $\mathcal{F}$-predictable) variance V_t; see Section 4. Indeed, this observation undergirds our empirical Bernstein bound. As in the sub-gamma case, Howard et al. (2020, Proposition 1) also shows that any CGF-like ψ can be upper bounded by $a\psi_{E,c}$ for some $a, c \geq 0$ (but as mentioned above for sub-gamma, exploiting this transformation may yield looser bounds than directly dealing with ψ).

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; this terminology stems from the fact that $\psi_{P,c}$ is the CGF of a centered unit-rate Poisson random variable. Most proofs of Bennett's inequality (e.g., Boucheron et al. (2013, Theorem 2.9)) involve showing that a bounded random variable is sub-Poisson. In the multivariate setting, Lemma D.1 shows that vectors (X_t) with $\|X_t\| \leq b$ are sub-$\psi_{P,b}$ with $V_t = \sum_{j \leq t} \mathbb{E}[X_j X_j^\intercal | \mathcal{F}_{j-1}]$.

Let us now turn to the techniques we use to prove our results. The bounds of de la Peña et al. (2008) and Abbasi-Yadkori et al. (2011) are proved using the so-called "method of mixtures" (sometimes also called the pseudo-maximization technique (de la Peña et al., 2007)). At its core, the method of mixtures involves (i) finding a family of supermartingales $Z(\theta)$ indexed by some parameter space Θ, (ii) constructing a new supermartingale by integrating this family with respect to some measure ρ over Θ, say $Z_t = \int_\Theta Z_t(\theta) \mathrm{d}\rho$, 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) and ρ 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*. Here we also begin 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 ρ and some prior distribution. The benefit of this technique is that ρ can be data-dependent. This approach lies at the heart of (and originated in) PAC-Bayesian learning theory (Guedj, 2019; Alquier et al., 2024). A generic template for such a bound, which applies to general stochastic processes and handles stopping times, is given by Chugg et al. (2023, Theorem 4).

Proposition 2.2 (Variational Template)**.** *Let Θ 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 $L(\theta) \equiv (L_t(\theta))_{t \geq 0}$. Assume that all processes are adapted to the same filtration $\mathcal{F}$ and that $L_0(\theta) \leq 1$ for each $\theta \in \Theta$. Let ν be a data-free distribution over Θ. Then, for all $\delta \in (0, 1)$, with probability $1 - \delta$, for any $\mathcal{F}$-adapted stopping time τ and any $\mathcal{F}_\tau$-measurable distribution ρ_τ over Θ,*

$$\int \log Z_\tau(\theta) \rho_\tau(\mathrm{d}\theta) \leq D_{KL}(\rho_\tau \| \nu) + \log(1/\delta). \tag{3}$$

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 ρ_τ and ν. 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 $\gamma_{\min}(M) = \gamma_1(M) \leq \cdots \leq \gamma_d(M) = \gamma_{\max}(M)$. We let $\|M\|_{\mathrm{op}} = \gamma_{\max}(M)$ be the operator norm of M. 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 $A^\intercal$ to refer to the transpose of the matrix A.

[2] In particular, Appendix E in Howard et al. (2020) shows that $\log \mathbb{E} e^{\lambda X} \leq \psi_{E,c}(\lambda)\sigma^2$ for all $0 \leq \lambda < 1/c$ and some $\sigma, c > 0$ if and only if $\log \mathbb{E} e^{\lambda X} \leq \psi_N(\lambda) u$ for all $0 \leq \lambda < 1/b$ for some $u, b > 0$.

3 General Sub-ψ Processes

3.1 A line-crossing inequality for sub-ψ processes

We must define two quantities before stating our first result. First, for nonzero positive definite matrices M_1 and M_2, define their *Rayleigh-Ritz quotient*[3]

$$\alpha(M_1, M_2) \equiv \sup_{\theta \in \mathbb{R}^d} \frac{\langle \theta, M_1 \theta \rangle}{\langle \theta, M_2 \theta \rangle}. \tag{4}$$

When using the Rayleigh-Ritz quotient 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 PSD matrix U_0, let

$$\alpha_t = \alpha(U_0, U_0 + V_t), \tag{5}$$

and for $t \geq 1$ define the function $g_{\psi,t} : \mathrm{Im}(\psi) \to [0, 1]$ by

$$g_{\psi,t}(z) = \sqrt{\psi^{-1}(z)} - \sqrt{\alpha_t \psi^{-1}(z)}. \tag{6}$$

Clearly, $g_{\psi,t}(z) \to \sqrt{\psi^{-1}(z)}$ as $\alpha_t \to 0$, which occurs for example if $\gamma_{\min}(V_t) \to \infty$. For most common sub-$\psi$ processes and matrices V_t which have at least some spread in all directions, $g_{\psi,t}(z)$ is rapidly approaching $\sqrt{\psi^{-1}(z)}$ after 100 samples for all $\lambda \in \mathrm{Im}(\psi)$; see Figure 1. We now present the line-crossing inequality for sub-ψ processes. The proof is in Appendix G.2.

Theorem 3.1. *Let (S_t, V_t) be a sub-ψ process such that $\lambda \mapsto \psi(\lambda)/\lambda^2$ is nondecreasing. Let U_0 be positive definite and define the Rayleigh-Ritz quotient $\alpha_t = \alpha(U_0, U_0 + V_t)$. Then, for any $\delta \in (0, 1)$ and any $\lambda \in [\psi^{-1}(1/\gamma_{\min}(U_0)), \lambda_{\max}) \cap \mathrm{Im}(\psi)$, with probability $1 - \delta$, for any stopping time τ,*

$$\|S_\tau\|_{(V_\tau + U_0)^{-1}} \leq \frac{1}{\lambda g_{\psi,\tau}(\lambda)} \left(\frac{1}{2} \log \left(\frac{\det(V_\tau + U_0)}{\det U_0} \right) + 1 + \log(1/\delta) \right) + \frac{g_{\psi,\tau}(\lambda)\psi(\lambda)}{\lambda}. \tag{7}$$

Remark 3.2. Theorem 3.1 assumes that ψ is super-Gaussian, i.e., that $\psi(\lambda)/\lambda^2$ is nondecreasing. This assumption is without loss of generality in the following sense: As we detailed in Section 2, any ψ function can be upper bounded by a constant times $\psi_{G,c}$, which is super-Gaussian. (In fact, all sub-ψ processes mentioned in Section 2 are super-Gaussian). Thus, Theorem 3.1 may be applied to any sub-ψ process, at the cost of some looseness in constants if ψ is strictly sub-Gaussian (e.g., sub-Bernoulli).

Let us provide a sketch of the proof of Theorem 3.1. Due to the super-Gaussian assumption we can assume that the sub-ψ condition holds for all $\theta \in \mathbb{B}^d$—the unit *ball* as opposed to the unit sphere. We thus apply Proposition 2.2 to the process $Z_t(\theta) = \exp\{\lambda\langle\theta, S_t\rangle - \psi(\lambda)\langle\theta, (V_t + U_0)\theta\rangle\}$ for $\theta \in \mathbb{B}^d$, where we take distributions ν and ρ to be uniform over ellipsoids $\mathcal{E}_\nu$ and $\mathcal{E}_\rho$ sitting inside $\mathbb{B}^d$. The shape (hence volume) of $\mathcal{E}_\nu$ is a function of U_0^{-1}; that of $\mathcal{E}_\rho$ a function of $(V_\tau + U_0)^{-1}$ (recall that ρ can be $\mathcal{F}_\tau$-measurable). With these choices, integrating out $\int \log Z_\tau(\theta) \mathrm{d}\rho$ and applying Proposition 2.2 gives that with probability $1 - \delta$,

$$(\lambda\beta - \beta^2\psi(\lambda))\|S_\tau\|_{(V_\tau + U_0)^{-1}} \leq D_{\mathrm{KL}}(\rho\|\nu) + 1 + \log(1/\delta).$$

The KL-divergence between ρ and ν is the ratio of their volumes, which is $\frac{1}{2}\log(\det(V_\tau + U_0)/\det U_0)$, giving rise to the determinant-based bound. The parameter β is a multipler on the mean of $\mathcal{E}_\rho$, ensuring that it sits inside $\mathcal{E}_\nu$. Optimizing β gives the final result.

Several remarks on Theorem 3.1 are in order. First, the lower bound on λ, i.e., $\lambda \geq \psi^{-1}(1/\gamma_{\min}(U_0))$ comes from the proof technique described above. The precise form of $\mathcal{E}_\nu$ is $\{\theta : \psi(\lambda)\theta^t U_0 \theta \leq 1\}$, which we require to be a subset of $\mathbb{B}^d$. This holds if $\psi(\lambda)U_0 \succeq I_d$, i.e., $\lambda \geq \psi^{-1}(\gamma_{\min}^{-1}(U_0))$. Further, the requirement that $\lambda \leq \lambda_{\max}$ comes from Definition 2.1, and the requirement that $\lambda \in \mathrm{Im}(\psi)$ comes from the definition of $g_{\psi,t}$. Note, however, that for common ψ functions including ψ_E, ψ_P, ψ_G and ψ_N, $\mathrm{Im}(\psi) = \mathbb{R}_{\geq 0}$, so requiring that $\lambda \in \mathrm{Im}(\psi)$ is often a vacuous constraint.

[3]This is also the maximum over the generalized Rayleigh coefficient between M_1 and M_2, and the maximum solution to the generalized eigenvalue problem $M_1 x = \lambda M_2 x$.

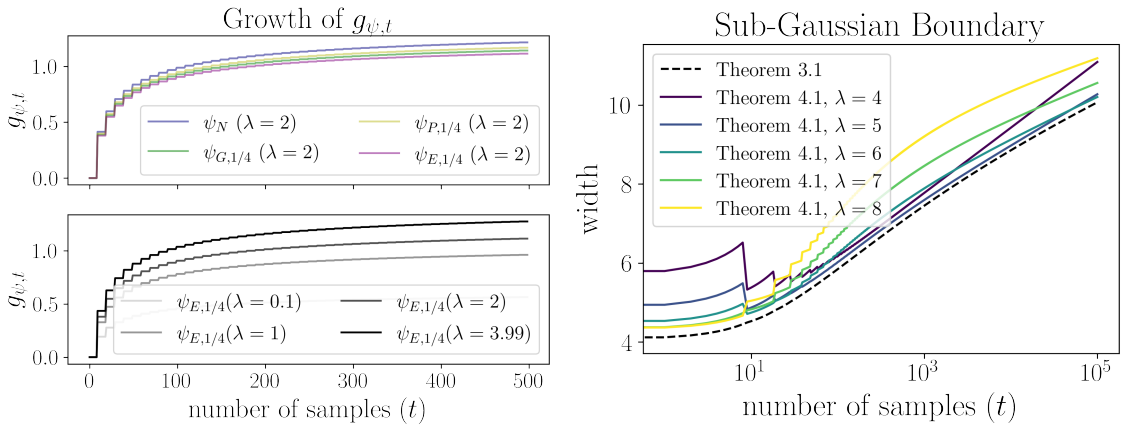


Figure 1: *Left*: The growth of g_t for various ψ functions and across various values of λ. For sub-gamma, sub-Poisson, and sub-exponential we fix $c = 1/4$. *Right:* A comparison of the sub-Gaussian bound (dotted black line) of de la Peña et al. (2004); Abbasi-Yadkori et al. (2011) (Theorem B.1) and Theorem 3.1 (various colored lines) instantiated for $\psi = \psi_N$. In both figures we use $U_0 = I_d$ and $V_t = \sum_{k \le t} X_s X_s^\intercal$ where the vectors X_t are chosen based on a bandit algorithm. Details may be found in Appendix H.

Remark 3.3. That λ is bounded away from 0 in Theorem 3.1 may seem like a cause for concern. In many non-self-normalized concentration inequalities (e.g., traditional Hoeffding, Bernstein, Bennett bounds) one typically chooses λ as a decreasing function of n. However, that is not the case with the self-normalized inequalities studied here, whose widths are growing with time. For this reason, the optimal value of λ at a particular time $t = n$ is an increasing function of n (this is illustrated concretely by the preceding discussion, and by the proof of Theorem G.1).

It's worth asking how Theorem 3.1 compares with the famous bound of Abbasi-Yadkori et al. (2011), which holds when (S_t, V_t) is a sub-Gaussian process. This bound is stated as Theorem B.1 in Appendix B. The right hand panel of Figure 1 plots the comparison for various values of λ. As one would expect, Theorem 3.1 is looser since it handles a wider class of distributions and is not explicitly optimized for sub-Gaussian processes. That said, Theorem 3.1 loses surprisingly little over Theorem B.1, even for fixed values of λ. As we can see, one faces a tradeoff when choosing λ, 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). The early staggered jumps in the sub-Gaussian boundary in Figure 1 are a result of small values of $g_{\psi,t}$ early on, before α_t shrinks sufficiently.

Figure 2 compares Theorem 3.1 to the bound presented by Whitehouse et al. (2023). We fix the growth rate of the condition number, $\kappa(V_t)$, and vary that of $\det(V_t)$. As expected, for most values of λ, our bound outperforms the condition number-based bound of Whitehouse et al. (2025) when the determinant grows slowly but is uniformly weaker when the determinant grows quickly.

Intriguingly, the bound of Whitehouse et al. (2025) seems to always dominate our bound at small sample sizes (roughly < 200). The is due to the factor of $g_{\psi,t}(\lambda)$ in our bounds, which is small at low sample sizes (or, more specifically, until $\gamma_{\min}(V_t + U_0)$ grows sufficiently large). This highlights a drawback of our bounds: if $\gamma_{\min}(V_t + U_0)$ never grows, or grows extremely slowly, then $g_{\psi,t}(\lambda)$ can remain extremely small, blowing up the bound and making it vacuous.

3.2 Optimizing λ via Stitching

In order to overcome the hurdle of optimizing λ in Theorem 3.1, will apply the bound of Theorem 3.1 iteratively over geometrically spaced epochs, choosing a different λ 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. (2021). In this section we provide the stitching result for sub-gamma

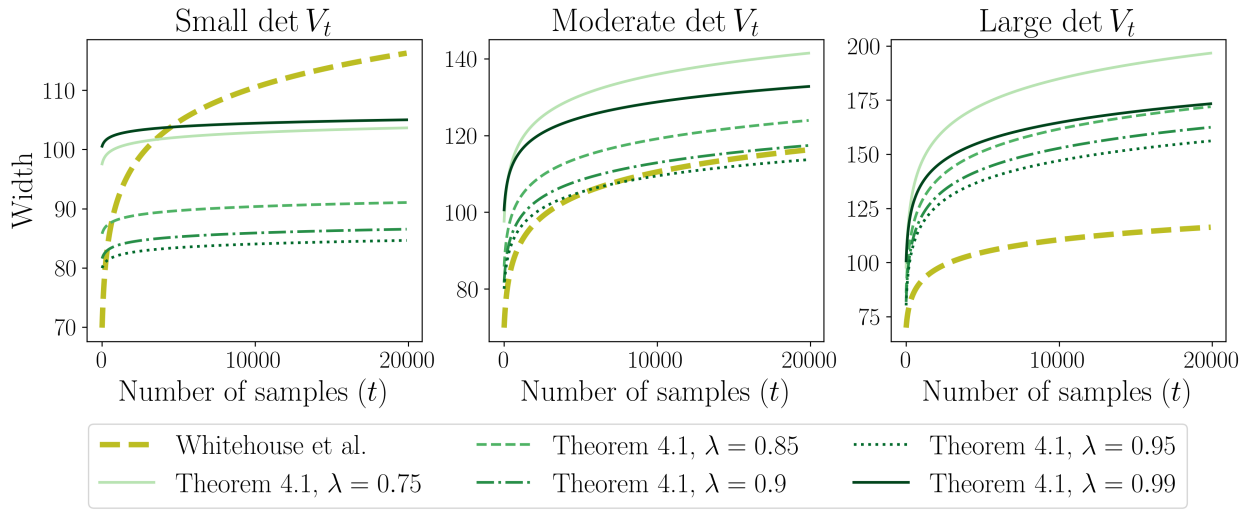


Figure 2: A comparison of Theorem 3.1 with the bound of Whitehouse et al. (2025), which we recall is based on the condition number of V_τ. We control the growth of the determinant with rank-k updates: as k grows, so does $\det V_\tau$. The condition number has the same growth in each case. As the sample size grows, Theorem 3.1 outperforms the condition number-based bound for smaller values of $\det V_\tau$, though the performance depends substantially on the choice of λ. It 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 H.

processes. A result for more general processes is given in the appendix; see Theorem C.2. As we mentioned in Section 2, all sub-ψ functions are sub-gamma, in the sense that for all twice-differentiable $\psi : [0, \lambda_{\max}) \to \mathbb{R}$ with $\psi(0) = \lim_{x \to 0+} \psi'(x) = 0$, there exist constants $a, c > 0$ such that $\psi(\lambda) \leq a\psi_{G,c}(\lambda)$ (Howard et al., 2020, Proposition 1). Sub-gamma bounds can thus be applied to all sub-ψ processes after appropriate scaling.

As far as stitching is concerned, in the proof of Theorem 3.4 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 can be found in Appendix G.3.

Theorem 3.4. *Let (S_t, V_t) be a sub-$\psi_{G,c}$ process for any $c > 0$. Further suppose that $V_t + U_0 \succeq I_d$ for all $t \geq 1$ and some positive-definite U_0 with $\upsilon = \gamma_{\min}(U_0)$. Fix $\delta \in (0, 1/e]$. Let $\xi(x) = 2/(c + \sqrt{c^2 + 2x})$. Then, with probability $1 - \delta$, for all stopping times τ,*

$$\|S_\tau\|_{(V_\tau + U_0)^{-1}} \leq \frac{cD_\tau + 1.51\sqrt{D_\tau} + \xi(c) \max\left\{\frac{1}{\xi(\upsilon)\upsilon}, \sqrt{\frac{D_\tau}{2}}\right\}}{H_{c,\tau}}, \tag{8}$$

where

$$D_\tau = \frac{1}{2} \log\left(\frac{\det(V_\tau + U_0)}{\det U_0}\right) + 1.5 + 2\log(\log_2(\det(V_\tau + U_0)) + 1) + \log(1/\delta), \tag{9}$$

and

$$H_{c,\tau} = 0 \vee \left(\sqrt{\frac{\sqrt{c^2 + 2c} - c}{3 + c}} - \sqrt{\frac{\gamma_{\max}(U_0)}{\gamma_{\min}(V_t + U_0)}\xi(c)}\right). \tag{10}$$

The denominator in Theorem G.1 should be viewed as a penalty we pay for the appearance of $g_{\psi,t}$ in Theorem 3.1. If $\gamma_{\min}(V_t + U_0) \to \infty$ then $H_{c,t}$ converges to a constant depending on c, just as $g_{\psi,t}$. Note that $H_{c,t}$ is larger than 0 as soon as $\gamma_{\min}(V_t + U_0) > \frac{(3+c)\gamma_{\max}(U_0)}{c}$. Thus, as $c \to 0$, it takes either more

observations or a faster growth rate of $\gamma_{\min}(V_t + U_0)$ for the bound to be informative. We suspect that the appearance of $H_{c,\tau}$ in (8) is sub-optimal, but it's unclear how to remove this term with our current proof techniques.

Theorem 3.4 assumes that $V_1 + U_0 \succeq I_d$ to ensure that the union of the epochs $E_k = \{t : \eta^k \leq \det(V_t + U_0) < \eta^{k+1}\}$, $k \geq 0$ 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-ψ process then Whitehouse et al. (2025, Proposition 2.4) shows that $(S_t/\sqrt{\beta}, V_t/\beta)$ is a sub-ψ_β process for $\psi_\beta(\lambda) = \beta\psi(\lambda/\sqrt{\beta})$. We can thus apply Theorem 3.4 to this rescaled process.

Remark 3.5. The appearance of both the logarithmic term $\log\det V_\tau$ and the iterated logarithmic term $\log\log\det V_\tau$ in Theorem 3.4 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. (2025) and is comparable to the dependence on variance-like terms in non self-normalized bounds (see, e.g., Chugg et al. 2023, Theorem 2.3).

Remark 3.6. 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. (2025) 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. Howard et al., 2021). 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 $\det V_\tau$ only is impossible, as shown by the following example.

Example 3.7. We borrow the conclusion of an elegant example from de la Peña et al. (2007), also studied by Whitehouse et al. (2025) (see their Example 4.6). In particular, de la Peña et al. (2007) 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 $\gamma_{\min}(V_t) \sim \log(t)/(1+a)$ for a constant a. Hence $\det V_t \sim t\log t$ implying that $\|S_t\|_{V_t^{-1}} \sim \log(\det V_t)$.

3.3 A note on suspected suboptimality

Consider Theorem 3.4 with V_t such that $\gamma_{\min}(V_t) \to \infty$ and $c = 1$. With probability $1 - \delta$, we obtain that $\|S_t\|_{(V_t+U_0)^{-1}} \lesssim \log\det(V_t) + \log(1/\delta) + \sqrt{\log\det(V_t) + \log(1/\delta)}$. To put this in more familiar terms, suppose that $V_t = tI_d$. Then we obtain the bound

$$\left\|\frac{S_t}{t}\right\| \lesssim \frac{d\log(t) + \log(1/\delta)}{\sqrt{t}} + \sqrt{\frac{d\log(t) + \log(1/\delta)}{t}}.$$

This resembles the terms in a Bernstein bound, except for a missing factor of $1/\sqrt{t}$ on the first term. Indeed, we expect the linear term in a Bernstein bound to shrink at rate $1/t$, not $1/\sqrt{t}$, thus making it asymptotically smaller than the square root term. This makes us suspect that Theorem 3.4 is suboptimal in terms of dependence on D_t. We imagine that the correct dependence would resemble $cD_t/\gamma_{\min}(V_t) + \sqrt{D_t}$, akin to the bound of Whitehouse et al. (2025) (though we emphasize that theirs is not a log-determinant bound, and thus not dimension-free).

It is instructive to compare our bound with several concurrent works which study log-determinant Bernstein bounds (Metelli et al., 2025; Akhavan et al., 2025; Martinez-Taboada et al., 2026). Like us, Akhavan et al. (2025) obtain a bound which scales linearly in $\log\det(V_t)$. Meanwhile, both Metelli et al. (2025) and Martinez-Taboada et al. (2026) avoid this linear dependence, but at the cost of a multiplicative dependence between $\log\det(V_t)$ and $\log(1/\delta)$. That is, their bounds scale as

$$\|S_t\|_{(V_t+U_0)^{-1}} \lesssim \log(1/\delta) + \sqrt{\log(\det(V_t))\log(1/\delta)}.$$

Our bound will thus be tighter for $\log(\det(V_t)) \lesssim \log(1/\delta)$, and theirs when $\log(1/\delta) \lesssim \log(\det(V_t))$. There are other differences as well. For instance, the result of Metelli et al. (2025) only applies to bounded observations; those of Martinez-Taboada & Ramdas (2026) are not time-uniform or either require a priori upper bounds on $\log\det(V_t)$. Whether one can obtain a time-uniform log-determinant bound for sub-gamma processes with (i) additive dependence between $\log\det(V_t)$ and $\log(1/\delta)$ and (ii) a linear term vanishing at a faster rate, is thus an open question.

4 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 (Maurer & Pontil, 2009; Waudby-Smith & Ramdas, 2024; Howard et al., 2021; Chugg & Ramdas, 2025), vector-valued random variables (Kohler & Lucchi, 2017; Martinez-Taboada & Ramdas, 2026; 2025; Chugg et al., 2025), u-statistics (Peel et al., 2010), and matrices (Wang & Ramdas, 2024; Chugg & Ramdas, 2025). As far as we are aware, Whitehouse et al. (2025) 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 μ. Suppose that $\|X_t\| \leq 1/2$ (one can replace $1/2$ with any constant and rescale the bound accordingly). Recall that $\psi_{E,1}(\lambda) = -\log(1-\lambda) - \lambda$. For any adapted sequence $(\widehat{\mu}_t)_{t\geq 0}$,[4] the process defined by

$$N_t(\theta) = \prod_{k\leq t} \exp\left\{\lambda\langle\theta, X_k - \mu\rangle - \psi_{E,1}(\lambda)\langle\theta, X_k - \widehat{\mu}_{k-1}\rangle^2\right\}, \quad N_0(\theta) = 1. \tag{11}$$

is a nonnegative supermartingale (see Chugg et al. (2025, Lemma E.1) for an explicit proof). It is a multivariate extension of a supermartingale used by Howard et al. (2021), 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. (2015), where the mean is replaced by the key term $\widehat{\mu}_{k-1}$ 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 (Waudby-Smith & Ramdas, 2024; Martinez-Taboada & Ramdas, 2026; Wang & Ramdas, 2024).

That the process in (11) is a supermartingale implies that (S_t, V_t) is a sub-$\psi_{E,1}$ process where

$$S_t = \sum_{k\leq t}(X_k - \mu), \quad V_t = \Gamma + \sum_{k\leq t}(X_k - \widehat{\mu}_{k-1})(X_k - \widehat{\mu}_{k-1})^\intercal, \tag{12}$$

and Γ is any fixed PSD matrix (could be the zero matrix). Since $\psi_{E,1}$ is super-Gaussian, we may apply Theorem 3.1 to obtain the following result.

Corollary 4.1. *Let (X_t) be a stream of random vectors such that $\|X_t\|_2 \leq 1/2$ for all $t \geq 1$. Let U_0 be positive definite and V_t be as above. Let $g_t(z) = g_{\psi_{E,1},t}(z)$ for $g_{\psi,t}$ as in (6). Then, for any $\delta \in (0,1)$ and any $\lambda \in [\psi_{E,1}^{-1}(1/\gamma_{\min}(U_0)), 1)$, with probability $1-\delta$, for any stopping time τ,*

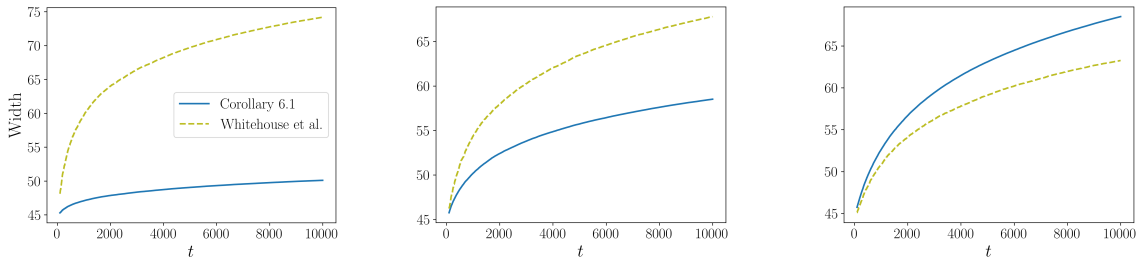
$$\left\|\sum_{k\leq\tau}(X_k - \mu)\right\|_{(V_\tau+U_0)^{-1}} \leq \frac{1}{\lambda g_\tau(\lambda)}\left(\frac{1}{2}\log\left(\frac{\det(V_\tau+U_0)}{\det U_0}\right) + 1 + \log(1/\delta)\right) + \frac{g_\tau(\lambda)\psi_{E,1}(\lambda)}{\lambda}.$$

The upper bound of 1 on λ comes from the fact that $\lambda_{\max} = 1$ for $\psi_{E,1}$. Note that $\psi^{-1}(1/\kappa) < 1$ for all $\kappa > 0$ so there is no constraint on the minimum eigenvalue of U_0 other than positivity.

We added the PSD matrix Γ to the process (V_t) in Corollary 4.1 because we can often obtain better bounds in practice by splitting $U_0 = \Gamma + U_0'$. To elaborate, suppose we want to bound $\|S_t\|_{(V_t+U_0)^{-1}}$ for some fixed U_0 using Corollary 4.1. If the X_k are mostly pointing in one direction, then V_t will be low-rank, in which case $\alpha_t = \alpha(U_0, U_0 + V_t) \approx 1$ and $g_t(\lambda) \approx 0$, blowing up the bound. Thus, applying the bound as stated can lead to vacuous results, at least until $\gamma_{\min}(V_t)$ grows sufficiently large. However, we can also apply the bound with $U_0' = (1-\epsilon)U_0$ and $\Gamma = \epsilon U_0$ for any $0 < \epsilon < 1$, which ensures that α_t is bounded away from 1. Of course, this is bounding the same process since $V_t + \Gamma + U_0' = V_t + U_0$ by construction. This leads to much better results in practice.

Figure 3 employs this stabilization trick and compares Corollary 4.1 to the self-normalized empirical Bernstein bound of Whitehouse et al. (2025, Theorem 4.1). Translated to our setting, this latter bounds reads: For all

[4] That is, $\widehat{\mu}_t$ is $\mathcal{F}_t$-measurable.



(a) $k = 2,\ d = 20$ (b) $k = 5,\ d = 20$ (c) $k = 10,\ d = 20$

Figure 3: Comparison of our empirical Bernstein bound, Corollary 4.1 (blue line) , with the empirical Bernstein bound of Whitehouse et al. (2025) (dotted line). Here each X_t is generated to have random uniform noise in k directions, so $V_{t+1} = V_t + (X_{t+1} - \widehat{\mu}_t)(X_{t+1} - \widehat{\mu}_t)^\top$ acts roughly as a rank-k update. As k grows, the determinant grows relative to the condition number, thus our determinant-based bound deteriorates with respect to the condition number bound. Simulation details can be found in Appendix H.

$\delta \in (0, 1)$, with probability $1 - \delta$, for all stopping times τ,

$$\|S_t\|_{(V_\tau + U_0)^{-1}} \leq 2\sqrt{\gamma_{\min}(V_\tau + U_0)} \cdot (\psi^*_{E,1})^{-1}\left(\frac{2L(V_\tau)}{\gamma_{\min}(V_\tau + U_0)}\right), \tag{13}$$

where $L(V_t) = \log\left(\frac{2}{\delta}\right) + \log\left(h\left(\log_2\left(\frac{\gamma_{\max}(V_t+U_0)}{\gamma_{\min}(U_0)}\right)\right)\right) + \log\left(2\sqrt{\kappa_t} \cdot N_{d-1}\left(\frac{1/4}{\sqrt{\kappa_t}}\right)\right)$, and $\kappa_t = \kappa(V_t + U_0)$ is the condition number of $V_t + U_0$. We have simplified the bound somewhat by fixing certain parameters that are left as variables in the original. Here $N_{d-1}(\epsilon)$ is the ϵ-covering number of the unit sphere $\mathbb{S}^{d-1}$, i.e., the minimum size of an ϵ-cover of $\mathbb{S}^{d-1}$. This is where the dimension-dependence comes from, as $N_{d-1}(\epsilon) = \Omega(\epsilon^{-(d-1)})$.

When comparing our bound with (13) we take $(\widehat{\mu}_t)$ to be the sequence of sample means, i.e., $\widehat{\mu}_t = t^{-1}\sum_{k \leq t} X_k$. The results echo those in Section 3: our bound beats the condition number-based bound of Whitehouse et al. insofar as $\log\det(V_t) \lesssim d\log(\kappa(V_t))$, where $\kappa(V_t)$ is the condition number of V_t. As the determinant grows relative to the condition number, our bound becomes looser than (13). In practice, if we do not know how the determinant will grow, we might use a union bound to take advantage of both bounds. Just as in Section 3.2, we can use stitching to obtain a bound that does not depend on λ. We provide the details in Appendix E.

5 Conclusion

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-ψ 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 (Abbasi-Yadkori et al., 2011; Akhavan et al., 2025; Ziemann, 2025; Metelli et al., 2025), and the dimension-dependent, condition number-based bounds of Whitehouse et al. (2025). 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, as discussed in Section 3.3, we expect that our bounds are suboptimal, in the sense that they do not recover expected Bernstein-like behavior for $V_t = tI_d$. Can this issue be fixed? Second, can we remove the dependence on $H_{c,\tau}$ in Theorem 3.4? Is the function g_ψ in Theorem 3.1 necessary, or is it simply an artifact of the proof technique? While g_ψ is easy to compute, its presence makes it tougher to reason about these bounds analytically.

References

Yasin Abbasi-Yadkori. *Online learning for linearly parametrized control problems*. PhD thesis, University of Alberta, Edmonton, Canada, 2013. 14

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011. 1, 2, 4, 6, 10, 14, 15, 16, 18, 20

Arya Akhavan, Amitis Shidani, Alex Ayoub, and David Janz. Bernstein-type dimension-free concentration for self-normalised martingales. *arXiv preprint arXiv:2507.20982*, 2025. 2, 8, 10, 14

Pierre Alquier et al. User-friendly introduction to PAC-Bayes bounds. *Foundations and Trends® in Machine Learning*, 17(2):174–303, 2024. 4

Bernard Bercu and Abderrahmen Touati. Exponential inequalities for self-normalized martingales with applications. *The Annals of Applied Probability*, 18:1848–1869, 2008. 14

Bernard Bercu and Taieb Touati. New insights on concentration inequalities for self-normalized martingales. *Electronic Communications in Probability*, 24, 2019. 14

David Blackwell and David Freedman. On the amount of variance needed to escape from a strip. *The Annals of Probability*, pp. 772–787, 1973. 3

Jorge Luis Borges. *Labyrinths: Selected stories & other writings*. Number 186. New Directions Publishing, 1964. 16

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. 3, 4, 18

Olivier Catoni and Ilaria Giulini. Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv preprint arXiv:1712.02747*, 2017. 1, 15

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. 1, 15

Ben Chugg and Aaditya Ramdas. Closed-form empirical Bernstein confidence sequences for scalars and matrices. *arXiv preprint arXiv:2512.21300*, 2025. 9

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. 4, 8

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, 9, 15, 22

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, 3, 4, 8

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, 14

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

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, 6, 14

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, 2, 4, 13, 14, 20

Bradley Efron. Student's t-test under symmetry conditions. *Journal of the American Statistical Association*, 64(328):1278–1302, 1969. 13

Xiequan Fan, Ion Grama, and Quansheng Liu. Exponential inequalities for martingales with applications. *Electron. J. Probab*, 20(1):1–22, 2015. 9

Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pp. 3052–3060. PMLR, 2020. 2, 14

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. 13

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. 13

Ilaria Giulini. Robust dimension-free Gram operator estimates. *Bernoulli*, 24(4B), 2018. 15

Philip Griffin and James Kuelbs. Some extensions of the LIL via self-normalizations. *The Annals of Probability*, pp. 380–395, 1991. 13

Philip S Griffin and James D Kuelbs. Self-normalized laws of the iterated logarithm. *The Annals of Probability*, pp. 1571–1601, 1989. 13

Benjamin Guedj. A primer on PAC-Bayesian learning. *arXiv preprint arXiv:1901.05353*, 2019. 4

Harold Hotelling. The generalization of student's ratio. *The Annals of Mathematical Statistics*, 2(3):360–378, 1931. 14

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. 1, 3, 4, 7, 14, 17

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. 2, 6, 8, 9

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. 13

Johannes Kirschner, Andreas Krause, Michele Meziu, and Mojmir Mutny. Confidence estimation via sequential likelihood mixing. *arXiv preprint arXiv:2502.14689*, 2025. 2, 15

Jonas Moritz Kohler and Aurelien Lucchi. Sub-sampled cubic regularization for non-convex optimization. In *International Conference on Machine Learning*, pp. 1895–1904. PMLR, 2017. 9

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020. 18

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. 13

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. 16

Diego Martinez-Taboada and Aaditya Ramdas. Sharp empirical Bernstein bounds for the variance of bounded random variables. *arXiv preprint arXiv:2505.01987*, 2025. 9

Diego Martinez-Taboada and Aaditya Ramdas. Empirical Bernstein in smooth Banach spaces. *Annals of Applied Probability*, 2026. 8, 9

Diego Martinez-Taboada, Tomas Gonzalez, and Aaditya Ramdas. Vector-valued self-normalized concentration inequalities beyond sub-Gaussianity. *Algorithmic Learning Theory (ALT)*, 2026. 2, 8, 14

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. *Conference on Learning Theory*, 2009. 9

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. 2, 8, 10, 14

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, 15

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, 15

Artsiom Patarusau, Nikita Puchkin, Maxim Rakhuba, and Fedor Noskov. Structured covariance estimation via tensor-train decomposition. *arXiv preprint arXiv:2510.08174*, 2025. 15

Thomas Peel, Sandrine Anthoine, and Liva Ralaivola. Empirical Bernstein inequalities for u-statistics. *Advances in Neural Information Processing Systems*, 23, 2010. 9

Nikita Puchkin and Maxim Rakhuba. Dimension-free structured covariance estimation. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 4276–4306. PMLR, 2024. 15

Qi-Man Shao. Self-normalized large deviations. *The Annals of Probability*, 25(1):285–328, 1997. 13

Student. The probable error of a mean. *Biometrika*, pp. 1–25, 1908. 13

Michel Talagrand. *Upper and lower bounds for stochastic processes*, volume 60. Springer, 2014. 15

Hongjian Wang and Aaditya Ramdas. Sharp matrix empirical Bernstein inequalities. *arXiv preprint arXiv:2411.09516*, 2024. 9

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. 9

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. 6, 18

Justin Whitehouse, Zhiwei Steven Wu, and Aaditya Ramdas. Time-uniform self-normalized concentration for vector-valued processes. *Annals of Applied Probability*, 2025. 1, 2, 3, 6, 7, 8, 9, 10, 14, 16, 31, 33

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, 15

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture Markov decision processes. In *Conference on Learning Theory*, pp. 4532–4576. PMLR, 2021. 2, 14

Ingvar Ziemann. A vector Bernstein inequality for self-normalized martingales. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. 2, 10, 15

A Related Work

The canonical example of a self-normalized process is the t-statistic, introduced by William Gosset (aka Student) in 1908 (Student, 1908). 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 $X_1, \ldots, X_n$ 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 (Efron, 1969; Logan et al., 1973) and gaining significant momentum in the 90s (Griffin & Kuelbs, 1989; 1991; Jing & Wang, 1999; Giné et al., 1997; Giné & Mason, 1998; Shao, 1997).

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 (de la Peña

et al., 2008); the book is mostly focused on the scalar setting (with some sub-Gaussian vector bounds as we describe in the following paragraph). In the scalar setting, subsequent work by Bercu & Touati (2008; 2019) 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. (2020) further generalized this framework by the studying time-uniform concentration of sub-ψ 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 scalar (and matrix) bounds.

Beyond the scalar setting, de la Peña et al. (2009) (see also de la Peña et al. (2008, 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 (Hotelling, 1931)—the generalization of the Student t-statistic to the multivariate setting—is an example of such a process. De la Peña (de la Peña et al., 2009) 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 (de la Peña et al., 2008, Equation (14.5))). This assumption is similar in spirit to the sub-ψ assumption of Howard et al. (2020) (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. (2011) apply results of de la Peña et al. (2004) and give what is now a famous self-normalized inequality. They show that, if $S_t = \sum_{k \le t} X_k \epsilon_k$ and $V_t = U_0 + \sum_{k \le t} X_k X_k^\intercal$ where ϵ_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 τ. This bound was extended to the infinite-dimensional case in Abbasi-Yadkori (2013).

Recently, Whitehouse et al. (2025) extended the general sub-ψ condition of Howard et al. (2020) to $\mathbb{R}^d$ and showed that, if (S_t) and (V_t) obey such a condition, then

$$\|S_\tau\|_{V_\tau^{-1}} \lesssim \gamma_{\min}^{1/2}(V_\tau)(\psi^*)^{-1}\left(\gamma_{\min}^{-1}(V_\tau)(\log\log(\gamma_{\max}(V_\tau)) + d\log(\kappa(V_\tau))\right), \tag{14}$$

where $\kappa(V_t) = \gamma_{\max}(V_\tau)/\gamma_{\min}(V_\tau)$ is the ratio of the largest to smallest eigenvalue of V_τ (the condition number) and ψ^* is the convex conjugate of ψ. In the sub-Gaussian case we have $(\psi^*)^{-1}(x) \asymp \sqrt{x}$, so the bound of Whitehouse et al. (2025) scales as

$$\|S_\tau\|_{V_\tau^{-1}} \lesssim \sqrt{\log\log(\gamma_{\max}(V_\tau)) + d\log\kappa(V_t)}. \tag{15}$$

This qualitatively new bound was also shown to be sharp with a corresponding lower bound. Neither $\det V_\tau$ (the product of all eigenvalues) or $\kappa(V_\tau)$ dominates the other, making the bound of Whitehouse et al. (2025) and Abbasi-Yadkori et al. (2011) 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. (2025) Metelli et al. (2025) gave dimension-free, self-normalized Bernstein inequalities, building off the earlier work of Faury et al. (2020) and Zhou et al. (2021), both of which provide dimension-dependent Bernstein-type bounds. Akhavan et al. (2025) study the case in which the normalization process (V_t) is the quadratic variation of (S_t) and Metelli et al. (2025) study the bandit setting (as do Faury et al. (2020) and Zhou et al. (2021)). All of these results apply to bounded processes and do not extend to general sub-ψ processes. They are closest to the Bennett and Bernstein results we give in Section D, though we note that our Bernstein bound applies to random vectors satisfying the Bernstein condition—a more general condition than boundedness.

In the contextual bandit setting, Martinez-Taboada et al. (2026) also give a dimension-free determinant-based bound for random vectors satisfying the Bernstein condition. It differs from ours in several ways: It is time-uniform for the finite horizon $1 \le t \le n$ for some fixed n, it has a multiplicative dependence between $\log(1/\delta)$ and $\log\det(V_t)$, and it requires a deterministic upper bound on $\log\det(V_t)$ to instantiate. Nevertheless, it has several appealing properties, and will dominate our bound in some regimes. We provide

a more involved discussion in Section 3.3. 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 (2016) in 2016, who was interested in bounding the tails of quadratic forms. Since then, a number of authors of have employed it to solve various problems.

In 2018, Giulini (2018) used the variational approach to estimate the Gram operator in Hilbert spaces. Around the same time, Catoni & Giulini (2017; 2018) used it to estimate the mean of random vectors and the operator norm of random matrices in finite-dimensional Euclidean spaces. More recently, all of Nakakita et al. (2024); Zhivotovskiy (2024); Puchkin & Rakhuba (2024); Patarusau et al. (2025) return to the question of matrix concentration/covariance estimation, each giving dimension-free bounds on sums of random matrices under various conditions. Chugg et al. (2025) study the concentration of random vectors using the variational method, giving bounds under a variety of distributional assumptions (including sub-ψ condition which we study here). All of the above works using the variational approach are in non-self-normalized settings.

Kirschner et al. (2025) and Ziemann (2025) recently noted that variational ideas can be used to obtain bounds similar to (or identical to) the ones presented by Abbasi-Yadkori et al. (2011). Ziemann (2025) also gives a dimension-dependent Bernstein inequality for bounded observations in a contextual bandit setting. One can view our work here as extending the ideas of Kirschner et al. (2025) and Ziemann (2025) to much more general stochastic processes.

B Warmup: Sub-Gaussian Processes

Let us begin by considering the special case of sub-Gaussian processes[5] . 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. (2011), 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 F.

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 (2025) recently made a similar observation, though he focuses on the bandit setting in particular. The proof can be found in Appendix G.1.

Theorem B.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 τ,*

$$\|S_\tau\|^2_{(V_\tau+U_0)^{-1}} \le \log\left(\frac{\det(V_\tau+U_0)}{\det U_0}\right) + 2\log(1/\delta). \tag{16}$$

To give a sense of how Proposition 2.2 is used to prove Theorem B.1, let us give a brief outline of the proof here.

If (S_t, V_t) is a sub-Gaussian process, then it follows that the process $(N_t(\theta))$ with $N_t(\theta) = \exp\{\lambda\langle\theta, S_t\rangle - \psi_N(\lambda)\langle\theta, V_t\theta\rangle\}$ is upper bounded by a supermartingale for all $\theta \in \mathbb{R}^d$ (not only for every $\theta \in \mathbb{S}^{d-1}$; this is

[5] 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 (2014), which is a collection of zero-mean random variable $\{Y_\theta : \theta \in \Theta\}$ such that $\mathbb{E}[\exp(\lambda(X_\theta - X_\phi))] \le \exp(\psi_N(\lambda)d^2(\theta,\phi))$ for some metric d and all $\theta, \phi \in \Theta$.

because $\lambda_{\max} = \infty$). We may therefore apply Proposition 2.2 with $\Theta = \mathbb{R}^d$. In particular, we take ρ_τ to be Gaussian with mean $(V_\tau + U_0)^{-1} S_\tau$ and covariance $(2\psi_N(\lambda))^{-1}(V_\tau + U_0)^{-1}$, and ν to be Gaussian with mean 0 and covariance $(2\psi_N(\lambda))^{-1} U_0^{-1}$. (Recall that ρ can be data-dependent while ν must be data-free.) Then $\int \log N_t(\theta) \mathrm{d}\rho_\tau = (\lambda - \psi_N(\lambda)) \|S_\tau\|^2_{(V_\tau + U_0)^{-1}} + (\text{extra terms})$ and $D_{\mathrm{KL}}(\rho_\tau \| \nu) = \frac{1}{2} \log(\det(V_\tau + U_0) / \det U_0) +$ (the same extra terms). Rearranging and taking $\lambda = 1$ gives the result.

Arithmetic aside, the crux of the argument is noticing that we may work with $\Theta = \mathbb{R}^d$ instead of $\Theta = \mathbb{S}^{d-1}$, enabling us to choose ρ and ν as Gaussians. This strategy unfortunately won't be available to us for more general sub-ψ distributions, which makes those settings more challenging. We will end up using uniform distributions instead.

Theorem B.1 recovers the bound of Abbasi-Yadkori et al. (2011) 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 η_k is scalar σ-sub-Gaussian noise. Then (S_t) is sub-ψ_N with variance proxy (V_t) defined as $V_t = \sigma^2 \sum_{k=1}^t X_k X_k^\intercal$. Taking $U_0 = \sigma^2 V_0$ for some V_0 and rearranging (16) gives Theorem 1 of Abbasi-Yadkori et al. (2011). More generally, Theorem B.1 applies to any sub-Gaussian process described in Section 2, such as $S_t = \sum_{k \leq t} X_k$ and $V_t = \sum_{k \leq t} X_k X_k^\intercal$ when X_t is conditionally symmetric.

While Theorem 3.1 is stated in terms of stopping times, it can equivalently be stated as a time-uniform bound of the form: $P(\forall t \geq 1 : (16) \text{ holds}) \geq 1 - \delta$. In this paper we state our bounds in terms of stopping times to conform to previous work in this area.

C A general bound via the convex conjugate

While Theorem 3.4 removes the need to choose λ, we find that in practice fixing λ and applying Theorem 3.1 results in the tightest bounds. Indeed, Theorem 3.1 typically outperforms Theorem 3.4 by a factor of three to five. 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. If one is interested in using the line-crossing inequality, we find that a good rule of thumb is to choose $\lambda \approx 0.85 \lambda_{\max}$. If possible, we recommend sample splitting to choose λ.

It's helpful to write Theorem 3.4 in terms of $\psi^*_{G,c}$, in which case (8) becomes

$$\|S_\tau\|_{(V_\tau + U_0)^{-1}} \lesssim \frac{(\psi^*_{G,c})^{-1}(D_\tau)}{H_{c,\tau}}. \tag{17}$$

The appearance of $(\psi^*_{G,c})^{-1}$ is comforting: reliance on the inverse convex conjugate of ψ is consistent with past work on sub-ψ concentration, such as Whitehouse et al. (2025) (see (14) and (15) in Section A) and Manole & Ramdas (2023). The reliance here also suggests that we may be able to obtain a bound for general ψ that relies on $(\psi^*)^{-1}$. This is the goal of Theorem C.2 below, which explicitly relates the growth of $\|S_t\|_{(V_t + U_0)^{-1}}$ to the inverse of ψ^*.

As stated, Theorem C.2 is more of mathematical interest than of practical value. We consider general CGF-like functions ψ which have a nonnegative third derivative and show how to relate the growth rate to the convex conjugate of ψ. Since we stitch over epochs of the form $\{t : \eta^k \leq \|S_\tau\|_{(V_\tau + U_0)^{-1}} < \eta^{k+1}\}$, the width of the bound is itself function of $\|S_\tau\|_{(V_\tau + U_0)^{-1}}$ itself instead of $\det V_\tau$ as in Theorem 3.4. In particular, the width depends on an iterated-logarithm term involving $\|S_\tau\|_{(V_\tau + U_0)^1}$ and a second term $f_\psi(2\|S_\tau\|_{(V_\tau + U_0)^{-1}})$, for a function f defined below. Such self-reference makes the bound difficult to apply. Still, as iterated logarithm terms grow extremely slowly,[6] and f is often sublinear, we feel the result gives a sense of the growth rate of $\|S_t\|_{(V_t + U_0)^{-1}}$.

To state the result, for a given CGF-like ψ, let

$$\lambda^*(a) := \operatorname*{argmax}_{\lambda \in (0, \lambda_{\max})} \lambda a - \psi(\lambda), \tag{18}$$

[6] For instance, $\log_2(\log_2(10^{200})) < 10$. 10^{200} is roughly the number of books in Jorge Luis Borges' *Library of Babel* (Borges, 1964), while the approximate number of the atoms in the observable universe is a meager 10^{80}.

which we note is well-defined; see Lemma G.3 in the appendix. Then define the function $f_\psi : \mathbb{R} \to \mathbb{R}$ as

$$f_\psi(x) := \psi(\lambda^*(x)) \left(\sqrt{\psi^{-1}(\lambda^*(x))} - 1 \right). \tag{19}$$

The function f_ψ is somewhat opaque, so let us give a few examples of how it grows as a function of a. The proof is provided in Appendix G.4.

Lemma C.1. *Let (a_n) be an increasing sequence. If $\psi = \psi_{G,c}$ for some $c > 0$, then*

$$f_\psi(a_n) = O\left(\frac{\sqrt{a_n}}{c^{3/2}} \right) \left(\sqrt{\frac{2}{c + \sqrt{c^2 + 2c}}} - 1 \right). \tag{20}$$

If $\psi = \psi_{E,c}$ for some $c > 0$, then

$$f_\psi(a_n) = O(\log(ca_n)) \left(\frac{1}{\sqrt{c}} - 1 \right). \tag{21}$$

Here the asymptotics are in terms of the sequence (a_n). We've dropped other constants but kept the dependence on c. Note that in both cases, if c is large enough, then $f(a_n)$ is negative.

We now state Theorem C.2. The proof is provided in Appendix G.5. We refer again to Remark 3.2 which points out that the assumption that $\psi(\lambda)/\lambda^2$ is nondecreasing is a minor one.

Theorem C.2. *Let (S_t, V_t) be a sub-ψ process where ψ is CGF-like, $\psi(\lambda)/\lambda^2$ is nondecreasing, and $\psi''' = \frac{\mathrm{d}^3\psi}{\mathrm{d}\lambda^3} \geq 0$. Let U_0 be positive-definite and satisfy $\gamma_{\min}^{-1}(U_0) < \psi(\lambda_{\max})$. Set $\widehat{V}_t = V_t + U_0$ for all t. Then, with probability $1 - \delta$, for all stopping times τ,*

$$\|S_\tau\|_{\widehat{V}_\tau^{-1}} \lesssim (\psi^*)^{-1} \left(\frac{\log(\det \widehat{V}_\tau) + \log(\log(\|S_\tau\|_{\widehat{V}_\tau^{-1}})/\delta)}{0 \vee \sqrt{1 - \gamma_{\max}(U_0)/\gamma_{\min}(\widehat{V}_\tau)}} + f_\psi\left(2\|S_\tau\|_{\widehat{V}_\tau^{-1}}\right) \right). \tag{22}$$

Given that $f_\psi(u)$ scales as $\sqrt{u}$ for $\psi = \psi_{G,c}$ and as $\log(u)$ for $\psi = \psi_{E,c}$, a natural question is whether one should convert a given sub-gamma process into a sub-exponential process before applying Theorem C.2. Indeed, any sub-$\psi_{G,c}$ is also a sub-$\psi_{E,3c/2}$ process (see Howard et al. 2020, Table 5). Such a transformation will not make much of a difference, however. Both $(\psi_{G,c}^*)^{-1}$ and $(\psi_{E,c}^*)^{-1}$ are dominated by linear terms (indeed $(\psi_{G,c}^*)^{-1}(u) = cu + \sqrt{2u}$ and $(\psi_{E,c}^*)^{-1}(u) = cu + O(c^{-1}\log(1 + c^2 u))$). Since f_ψ is sublinear in both cases, if we write out the bounds and bring all the terms depending on $\|S_t\|$ to one side, the difference between the sub-gamma and sub-exponential cases is a bound on $\|S_t\| - O(\sqrt{\|S_t\|})$ vs $\|S_t\| - O(\log(\|S_t\|)$, which have the same growth rate.

If we are prepared to sacrifice some tightness and to assume that f is sublinear (as is the case in Lemma C.1) we can move the terms depending on $\|S_\tau\|_{\widehat{V}_\tau^{-1}}$ outside of the convex conjugate in (22) and bound it by a constant times $\|S_\tau\|_{\widehat{V}_\tau^{-1}}$. This results in Corollary C.3. The proof is in Appendix G.6.

Corollary C.3. *Consider the same setup as Theorem C.2. Let $\varphi(y) = (\psi^*)^{-1}(y)$ and $\upsilon = \gamma_{\min}(U_0) > 1$. For any $0 < b < 1$, define the hitting time $\tau_0 = \inf\{t : \upsilon/\gamma_{\min}(\widehat{V}_t) \leq b\}$. Fix $\delta \in (0, 1)$ and suppose both that $f_\psi(x) \leq \kappa x$ for some $\kappa > 0$ and all $x \geq 0$, and $\varphi'(\log(1/\delta))(1 + 2\kappa) < \sqrt{1 - b}$. Then, with probability $1 - \delta$, for all stopping times $\tau \geq \tau_0$,*

$$\|S_\tau\|_{\widehat{V}_\tau^{-1}} \lesssim \left(\frac{\sqrt{1-b}}{\sqrt{1-b} - \varphi'(\log(1/\delta))(1 + 2\kappa)} \right) (\psi^*)^{-1} \left(\frac{\log(\det \widehat{V}_\tau) + \log(1/\delta)}{\sqrt{1 - \gamma_{\max}(U_0)/\gamma_{\min}(\widehat{V}_\tau)}} \right). \tag{23}$$

The hidden constants in (23) are those inherited from (22). We opt to keep the constant

$$\frac{\sqrt{1-b}}{\sqrt{1-b} - \varphi'(\log(1/\delta))(1 + 2\kappa)},$$

in the bound so that the cost of removing the reference to $\|S_\tau\|_{\widehat{V}_\tau^{-1}}$ on the right hand of (22) is made clear.

D Self-normalized 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. (2013, 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 ψ-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. Whitehouse et al., 2023; Abbasi-Yadkori et al., 2011; Lattimore & Szepesvári, 2020).

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

$$N_t^B(\theta;\lambda) = \prod_{k\le t} \exp\left\{\lambda\langle\theta, X_k\rangle - \log \mathbb{E}_{k-1}\exp(\lambda\langle\theta, X\rangle)\right\}, \tag{24}$$

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. (We use the superscript B to signal Bennett/Bernstein). Depending on the assumptions we make on (X_t), we can bound the log-MFG by some function of $\mathbb{E}_{k-1}X_kX_k^\intercal$, which will give rise to a sub-ψ 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 D.1. *Let (X_t) be a stream of random vectors in $\mathbb{R}^d$ with conditional mean 0 obeying $\sup_{\theta\in\mathbb{S}^{d-1}}\langle\theta, X_t\rangle \le b$ almost surely for some constant $b > 0$. Then $S_t = \sum_{j\le t} X_j$ and $V_t = U_0 + \sum_{j\le t}\mathbb{E}_{j-1}X_jX_j^\intercal$ constitute a sub-$\psi_{P,b}$ process for any PSD matrix U_0.*

The proof of Lemma D.1 may be found in Appendix G.7. Next we show how to obtain a sub-ψ process when the random vectors obey a Bernstein condition—the inequality in (25)—in each direction. We note that Bernstein's condition is more general than boundedness, in the sense that bounded random vectors satisfy (25) but the converse is not true in general. The proof of the following result is in Appendix G.8.

Lemma D.2. *Let (X_t) be a stream of random vectors with conditional mean 0. Suppose that for some $c > 0$ and all $\theta \in \mathbb{S}^{d-1}$, $t \ge 1$,*

$$\mathbb{E}_{t-1}[\langle\theta, X_t\rangle^q] \le \frac{q!c^{q-2}}{2}\mathbb{E}_{t-1}\langle\theta, X_t\rangle^2 \quad \textit{for all integers } q \ge 3. \tag{25}$$

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

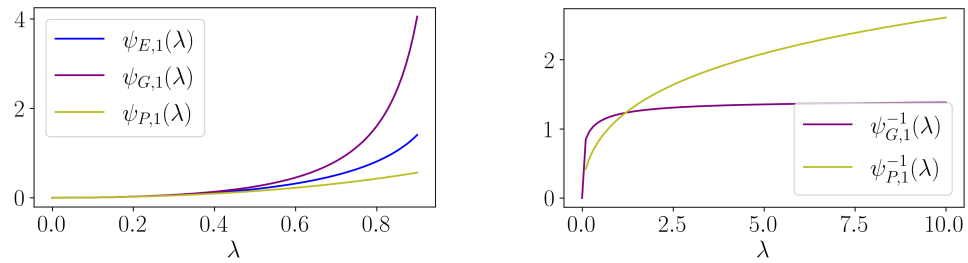
Using these two results in conjunction with Theorem 3.1 we obtain the following line crossing inequalities. Figure 4 plots $\psi_{P,1}^{-1}$ versus $\psi_{G,1}^{-1}$ to give a sense of the constraints on λ in the following result. It also plots $\psi_{G,1}$ versus $\psi_{P,1}$ to indicate how (27) scales compared to (28).

Corollary D.3. *Let (X_t) be a sequence of random vectors with conditional mean 0. Let U_0 be positive definite. Suppose either that (i) $\sup_{\theta\in\mathbb{S}^{d-1}}\langle\theta, X_t\rangle \le b$ for all $t \ge 1$ and some $b > 0$ or (ii) that (25) holds. Let $S_t = \sum_{k\le t} X_k$ and $V_t = \sum_{k\le t}\mathbb{E}_{k-1}[X_kX_k^\intercal]$. Fix any $\delta \in (0,1)$ and let*

$$D_t = \frac{1}{2}\log\left(\frac{\det(V_\tau + U_0)}{\det U_0}\right) + 1 + \log(1/\delta). \tag{26}$$

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 τ,

$$\|S_\tau\|_{(V_\tau+U_0)^{-1}} \le \frac{D_t}{\lambda g_{\psi_{P,b},\tau}(\lambda)} + \frac{g_{\psi_{P,b},\tau}(\lambda)\psi_{P,b}(\lambda)}{\lambda}, \tag{27}$$



(a) (b)

Figure 4: *Left:* Comparison of $\psi_{E,1}$, $\psi_{G,1}$, and $\psi_{P,1}$. As shown in Lemma E.1, $\psi_{G,1}$ 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 λ in Corollary D.3.

where $g_{\psi_{P,b}\tau}$ is defined as in (6). 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 τ,

$$\|S_\tau\|_{(V_\tau+U_0)^{-1}} \leq \frac{D_t}{\lambda g_{\psi_{G,c},\tau}(\lambda)} + \frac{g_{\psi_{G,c},\tau}(\lambda)\psi_{G,c}(\lambda)}{\lambda}. \tag{28}$$

Since the process is sub-gamma when (X_t) satisfies the Bernstein condition in (25), we can in that case apply Theorem 3.4 to obtain the stitched boundary in (8). In particular, by carefully applying the line-crossing inequality above with different values of λ at different times, we are able to apply (28) with λ as roughly $\sqrt{D_t}/(1+c\sqrt{D_t})$, up to constant multipliers on D_t. This gives $\|S_\tau\|_{(V_\tau+U_0)^{-1}}$ scaling like $cD_\tau + \sqrt{D_t}$ with high probability. This combination of linear and sublinear terms is what we expect of a Bernstein bound.

E Stitched Empirical Bernstein Bound

Here we provide the details on the stitched empirical Bernstein bound.

Lemma E.1. *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 λ. Since $h(0) = 0$ this implies that $h(\lambda) \geq 0$ for λ in the desired range. □

As illustrated by Figure 4a, for small λ, $\psi_{E,1}$ and $\psi_{G,1}$ are nearly identical. They separate more as λ grows (note that $\lambda_{\max} = 1$). Using Lemma E.1 and applying Theorem 3.4 with (S_t, V_t) as in (12) gives the following result.

Corollary E.2. *Suppose $\|X_t\|_2 \leq 1/2$. Let $V_t = \sum_{k=1}^t (X_k - \widehat{\mu}_{k-1})(X_k - \widehat{\mu}_{k-1})^\intercal$ and let U_0 be positive definite with $\upsilon = \gamma_{\min}(U_0)$. Fix $\delta \in (0, 1/e]$. Then, with probability $1-\delta$, for all stopping times τ,*

$$\|S_\tau\|_{(V_\tau+U_0)^{-1}} \leq \frac{D_\tau + 1.51\sqrt{D_\tau} + 0.74 \max\left\{\frac{1+\sqrt{1+2\upsilon}}{2\upsilon}, \sqrt{\frac{D_\tau}{2}}\right\}}{H_\tau}, \tag{29}$$

where

$$D_\tau = \frac{1}{2}\log\left(\frac{\det(V_\tau+U_0)}{\det U_0}\right) + 1.5 + 2\log(\log_2(\det(V_\tau+U_0))+1) + \log(1/\delta), \tag{30}$$

and

$$H_\tau = 0 \vee \left(0.42 - 0.86\sqrt{\frac{\gamma_{\max}(U_0)}{\gamma_{\min}(V_\tau+U_0)}}\right). \tag{31}$$

The precise numerical values in the bounds come from instantiating the various constants in Theorem 3.4 with $c = 1$ and simplifying. As with the stitching results in Section 3.2, Corollary E.2 is mostly of theoretical interest to understand the rate of our bounds. In practice we suggest using Corollary 4.1.

F 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. (2011) and de la Peña et al. (2008)—can also be used to prove Theorem B.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-ψ_N process, implying that for all $\theta \in \mathbb{R}^d$,

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

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 G.1 below.) Let ν be a Gaussian with mean 0 and covariance U_0^{-1} and consider the process with increments

$$M_t = \int_{\mathbb{R}^d} \exp\{\langle \theta, S_t\rangle - \psi_N(1)\langle \theta, V_t\theta\rangle\}\nu(\mathrm{d}\theta). \tag{32}$$

To compute M_t, notice that we can write

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

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 ν,

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

Since $\int N_t(\theta)\mathrm{d}\nu(\theta)$ 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 $P(M_\tau \ge 1/\delta) \le \mathbb{E}[M_1]\delta \le \delta$. In other words, with probability $1-\delta$, $\log M_\tau \le \log(1/\delta)$, which translates to

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

which is precisely Theorem B.1.

G Omitted Proofs

G.1 Proof of Theorem B.1

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

$$Z_t^\lambda(\theta) = \exp\left\{\lambda\langle \theta, S_t\rangle - \psi_N(\lambda)\langle \theta, V_t\theta\rangle\right\}. \tag{33}$$

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 θ, 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 (33) 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 τ be a stopping time and consider a Gaussian distribution $\rho = \rho_\tau$ with mean μ_ρ and Σ_ρ. We have

$$\begin{aligned}\int \log Z_\tau(\theta)\mathrm{d}\rho &= \lambda\langle\mu_\rho, S_\tau\rangle - \psi_N(\lambda)\int\langle\theta, V_\tau\theta\rangle\mathrm{d}\rho \\ &= \lambda\langle\mu_\rho, S_\tau\rangle - \psi_N(\lambda)\left(\langle\mu_\rho, V_\tau\mu_\rho\rangle + \mathrm{Tr}(V_\tau\Sigma_\rho)\right).\end{aligned} \tag{34}$$

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

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

Let ν be a Gaussian with mean 0 and covariance Σ_ν. Then

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

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

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

which in our case rearranges to

$$(\lambda - \psi_N(\lambda))\|S_\tau\|^2_{\widehat{V}_\tau^{-1}} \le \psi_N(\lambda)\mathrm{Tr}(\widehat{V}_\tau\Sigma_\rho) - \frac{d}{2} + \frac{1}{2}\log\left(\frac{\det\Sigma_\nu}{\det\Sigma_\rho}\right) + \log(1/\delta). \tag{35}$$

Taking $\Sigma_\rho = (2\psi_N(\lambda)\widehat{V}_\tau)^{-1}$, gives $\psi_N(\lambda)\mathrm{Tr}(\widehat{V}_\tau\Sigma_\rho) = d/2$. Also, $\det\Sigma_\nu/\det\Sigma_\rho = \det\widehat{V}_\tau/\det U_0$. Assuming $\lambda - \psi_N(\lambda) > 0$, (35) thus becomes

$$\|S_\tau\|^2_{(V_\tau+U_0)^{-1}} \le \frac{1}{\lambda - \psi_N(\lambda)}\left(\frac{1}{2}\log\left(\frac{\det(V_\tau+U_0)}{\det U_0}\right) + \log(1/\delta)\right). \tag{36}$$

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

G.2 Proof of Theorem 3.1

First note that if (S_t, V_t) is a sub-ψ process, then so too is $(S_t, V_t + U_0)$ for any PSD U_0. Let $\widehat{V}_t = V_t + U_0$. For the remainder of the proof we will consider the process $(S_t, \widehat{V}_t)$. Set

$$Z_t(\theta) = \exp\left\{\lambda\langle\theta, S_t\rangle - \psi(\lambda)\langle\theta, \widehat{V}_t\theta\rangle\right\}. \tag{37}$$

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. (2025, Lemma 2.10), but let us prove it for completeness. For any $\theta \in \mathbb{B}^d$, let $\phi = \theta/\|\theta\|$ and notice that

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

where the inequality follows from the fact that ψ is super-Gaussian. Since $\lambda\|\theta\| \leq \lambda$, $\psi(\|\theta\|\lambda)$ is well-defined and the final quantity defines a process which is upper bounded by a nonnegative supermartingale. 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 τ be a stopping time and let $\rho = \rho_\tau$ be a distribution over $\mathbb{B}^d$ with mean μ_ρ and covariance Σ_ρ. Then,

$$\begin{aligned}
\int \log Z_\tau(\theta)\mathrm{d}\rho &= \lambda\langle\mu_\rho, S_\tau\rangle - \psi(\lambda)\int\langle\theta, \widehat{V}_\tau\theta\rangle\mathrm{d}\rho \\
&= \lambda\langle\mu_\rho, S_\tau\rangle - \psi(\lambda)\left(\langle\mu_\rho, \widehat{V}_\tau\mu_\rho\rangle + \mathrm{Tr}(\widehat{V}_\tau\Sigma_\rho)\right). \qquad (38)
\end{aligned}$$

For some $\beta > 0$ to be determined later, we take ρ to be the uniform distribution over the ellipsoid with mean $\mu_\rho = \beta\widehat{V}_\tau^{-1}S_\tau$ and shape $A_\rho = \psi^{-1}(\lambda)\widehat{V}_\tau^{-1}$. That is, ρ 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 B.1, both μ_ρ and A_ρ are functions of τ but we suppress this dependence for convenience. The distribution ρ has covariance $\Sigma_\rho = (d+2)^{-1}A_\rho$ and

$$\psi(\lambda)\mathrm{Tr}(\widehat{V}_\tau\Sigma_\rho) = \frac{d}{d+2} \leq 1.$$

With these choices, (38) becomes

$$\int \log Z_\tau(\theta)\mathrm{d}\rho = (\lambda\beta - \beta^2\psi(\lambda))\langle\widehat{V}_\tau^{-1}S_\tau, S_\tau\rangle - \frac{d}{d+2}. \qquad (39)$$

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

$$(\lambda\beta - \beta^2\psi(\lambda))\|S_\tau\|^2_{\widehat{V}_\tau^{-1}} \leq D_{\mathrm{KL}}(\rho\|\nu) + 1 + \log(1/\delta), \qquad (40)$$

for a data-independent prior ν. If ν 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 β is small enough such that $\mathcal{E}_\rho \subseteq \mathcal{E}_\nu$, then

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

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

$$D_{\mathrm{KL}}(\rho\|\nu) = \frac{1}{2}\log\left(\frac{\det\widehat{V}_\tau}{\det U_0}\right). \qquad (41)$$

To ensure that $\mathcal{E}_\nu \subseteq \mathbb{B}^d$ it suffices that the minimum eigenvalue of A_ν^{-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 λ in the statement of the theorem.

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

and A_ν. Put $y = y_\tau = \widehat{V}_\tau^{-1} S_\tau$ and suppose that $\theta^t \widehat{V}_\tau \theta \le \psi^{-1}(\lambda)$. For any $\epsilon > 0$,

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

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 β gives (after taking the positive root of the quadratic equation)

$$\beta^* := \frac{\sqrt{\psi^{-1}(\lambda)} - \sqrt{\alpha_\tau \psi^{-1}(\lambda)}}{\|y\|_{U_0}} = \frac{g_\tau(\lambda)}{\|y\|_{U_0}}. \tag{42}$$

Notice that

$$\begin{aligned}
\|y\|_{U_0}^2 &= S_\tau^t \widehat{V}_\tau^{-1} U_0 \widehat{V}_\tau^{-1} S_\tau = (S_\tau^t \widehat{V}_\tau^{-1/2})(\widehat{V}_\tau^{-1/2} U_0 \widehat{V}_\tau^{-1/2})(\widehat{V}_\tau^{-1/2} S_\tau) \\
&\le \|\widehat{V}_\tau^{-1/2} U_0 \widehat{V}_\tau^{-1/2}\| \cdot S_\tau^t \widehat{V}_\tau^{-1} S_\tau \le \|S_\tau\|_{\widehat{V}_\tau^{-1}}^2,
\end{aligned}$$

where the final inequality follows since $U_0 \preceq \widehat{V}_\tau$ so $\widehat{V}_\tau^{-1/2} U_0 \widehat{V}_\tau^{-1/2} \preceq I$ and

$$\|\widehat{V}_\tau^{-1/2} U_0 \widehat{V}_\tau^{-1/2}\| \le 1.$$

Therefore, setting

$$\beta^{**} := \frac{g_\tau(\lambda)}{\|S_\tau\|_{\widehat{V}_\tau^{-1}}} \le \beta^*,$$

we obtain $(\theta + \beta^{**} y)^t U_0 (\theta + \beta^{**} y) \le h(\epsilon^*, \beta^{**}) \le h(\epsilon^*, \beta^*) = \psi^{-1}(\lambda)$ (where we've used that h is monotonically increasing in β), so (40) holds with $\beta = \beta^{**}$. Noting that

$$(\lambda \beta^{**} - (\beta^{**})^2 \psi(\lambda)) \|S_\tau\|_{\widehat{V}_\tau^{-1}}^2 = \lambda g_\tau(\lambda) \|S_\tau\|_{\widehat{V}_\tau^{-1}} - g_\tau^2(\lambda) \psi(\lambda),$$

and rearranging (40) completes the proof.

G.3 Proof of Theorem 3.4

Theorem 3.4 follows from the following more general result after taking $\eta = 2$ and $\ell(k) = (k+1)^2 \pi^2 / 6$. In this case $\sup_{k \ge 0} \ell(k+2)/\ell(k+1) \le 4$ so $\alpha_{\delta,\eta} \le 4.54$, 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 3.4.

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

$$\alpha_{\delta,\eta} := 1 + \frac{\log(\eta)}{2 + 2\log(1/\delta) - \log(\eta)} + \sup_{k \ge 0} \frac{\ell(k+2)}{\ell(k+1)}, \tag{43}$$

and $\xi(x) = 2/(c + \sqrt{c^2 + 2x})$*. Then, with probability* $1 - \delta$*, for all stopping times* τ*,*

$$\|S_\tau\|_{(V_\tau+U_0)^{-1}} \leq \frac{cD_{\tau,\eta}(U_0, V_\tau) + \sqrt{\frac{\alpha_{\delta,\eta}}{2} D_{\tau,\eta}(U_0, V_\tau)} + \xi(c) \max\left\{\frac{1}{\upsilon\xi(\upsilon)}, \sqrt{\frac{D_{\tau,\eta}(U_0,V_\tau)}{2}}\right\}}{H_c(U_0, V_\tau)}, \tag{44}$$

where

$$D_{\tau,\eta}(U_0, V_\tau) = \frac{1}{2}\log\left(\frac{\det(V_\tau + U_0)}{\det U_0}\right) + 1 + \log(\ell(\log_\eta(\det(V_\tau + U_0)))) + \log(1/\delta), \tag{45}$$

and

$$H_c(U_0, V_t) = 0 \vee \left(\sqrt{\frac{\sqrt{c^2 + 2c} - c}{\sqrt{\alpha_{\delta,\eta}} + c}} - \sqrt{\frac{\gamma_{\max}(U_0)}{\gamma_{\min}(\widehat{V}_t)}\xi(c)}\right) \tag{46}$$

Proof. Let $\widehat{V}_t = V_t + U_0$. Consider $\psi_{G,c}(\lambda)$ for $c > 0$. We consider breaking intrinsic time into geometric epochs, $E_k = \{t : \eta^k \leq \det \widehat{V}_t < \eta^{k+1}\}$ for all $k \geq 0$ and some $\eta > 1$. Recall we are assuming that $U_0 \succeq I_d$, implying 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 $\eta^k \leq \det \widehat{V}_t \leq \eta^{k+1}$ and set

$$D_t = \frac{1}{2}\log\left(\frac{\det \widehat{V}_t}{\det U_0}\right) + 1 + \log(1/\delta_{k(t)}). \tag{47}$$

Note that $k(t)$ is a random variable. By definition, $k(t) \leq \log_\eta \det \widehat{V}_t$, 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 our strategy is to apply Theorem 3.1 with $\delta = \delta_k$ and $\lambda = \lambda_k$, for some δ_k, λ_k to be chosen later. For $g = g_{\psi,t}$, this gives that with probability $1 - \delta_k$, for all $t \in E_k$,

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

where we've used that $g(\lambda_k) \leq \sqrt{\psi_{G,c}^{-1}(\lambda_k)} \leq \sqrt{\xi(c)}$ since $\psi_{G,c}^{-1}(u)$ is increasing in u and $\lambda_k < \lambda_{\max}$. Choose $\delta_k = \delta/\ell(k)$. Then,

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

It remains to choose λ_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

$$\lambda_k = \max\left\{\psi_{G,c}^{-1}(1/\upsilon), \frac{\sqrt{2B_k}}{1+c\sqrt{2B_k}}\right\}, \quad \text{where} \quad \upsilon = \gamma_{\min}(U_0). \tag{48}$$

Observe that $x/(1+cx) < 1/c$ for all $x \geq 0$, so $\lambda_k < \lambda_{\max}$. Therefore, $\lambda_k \in [\psi_{G,c}^{-1}(1/\upsilon), \lambda_{\max})$ and is a legal choice in Theorem 3.1.

Next we want to bound $B_{k(t)}$ and relate it to D_t. Note that B_k is increasing in k, and λ_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 G.2. *Define*

$$\alpha_{\delta,\eta} := 1 + \frac{\log(\eta)}{2 + 2\log(1/\delta) - \log(\eta)} + \sup_k \log(\ell(k+1)/\ell(k)). \tag{49}$$

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

Proof. Write

$$\begin{aligned} B_{k+1} &= \frac{1}{2}\left(\log\left(\frac{\eta^{k+1}}{\det U_0}\right)\right) + 1 + \log(\ell(k+1)/\delta) \\ &= \frac{1}{2}\left(\log(\eta) + \log\left(\frac{\eta^k}{\det U_0}\right)\right) + 1 + \log\left(\frac{\ell(k+1)\ell(k)}{\ell(k)\delta)}\right) \\ &= B_k + \frac{1}{2}\log\eta + \log(\ell(k+1)/\ell(k)) \\ &\leq B_k + \frac{1}{2}\log(\eta) + s, \end{aligned} \tag{50}$$

where $s = \sup_k \log(\ell(k+1)/\ell(k))$. We claim that $\log(\eta) \leq uB_k$ 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 $\eta^k \leq \det \widehat{V}_1 < \eta^{k+1}$. That is, we may assume that

$$\frac{\eta^k}{\det U_0} > \frac{\det \widehat{V}_1}{\eta \det U_0}.$$

Therefore,

$$\begin{aligned} uB_k &= \frac{u}{2}\log\left(\frac{\eta^k}{\det U_0}\right) + u + u\log(1/\delta_k) \\ &\geq \frac{u}{2}\log\left(\frac{\det \widehat{V}_1}{\eta \det U_0}\right) + u + u\log(1/\delta) \\ &\geq \frac{u}{2}\log\left(\frac{1}{\eta}\right) + u + u\log(1/\delta), \end{aligned}$$

where we've used that $\widehat{V}_1 \succeq U_0$. 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.,

$$u \geq \frac{\log(\eta)}{1 + \log(1/\delta) - \log(\eta)/2}, \tag{51}$$

where we are assured that the denominator on the right hand side is greater than 0 since $\eta < (e/\delta)^2$. Set u° to be the right hand side of (51). Then, from (50) we have

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

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. □

Since λ_k is increasing in B_k, Lemma G.2 implies that

$$\frac{\sqrt{2D_t/\alpha_{\delta,\eta}}}{1+c\sqrt{2D_t/\alpha_{\delta,\eta}}} \leq \frac{\sqrt{2B_{k(t)}}}{1+c\sqrt{2B_{k(t)}}} \leq \frac{\sqrt{2D_t}}{1+c\sqrt{2D_t}}. \tag{52}$$

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 λ. Using the lower bound in (52) in the first term and the upper bound in the second, we have

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

Meanwhile, if $\lambda_{k(t)} = \psi_{G,c}^{-1}(1/\upsilon)$, then we may still use the lower bound in (52) and we obtain

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

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

$$\frac{D_t}{\lambda_{k(t)}} + \frac{\xi(c)\psi_{G,c}(\lambda_{k(t)})}{\lambda_{k(t)}} \leq \sqrt{\frac{\alpha_{\delta,\eta}D_t}{2}} + cD_t + \xi(c)\max\left\{\frac{c+\sqrt{c^2+2\upsilon}}{2\upsilon}, \sqrt{\frac{D_t}{2}}\right\}. \tag{53}$$

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

$$\psi_{G,c}^{-1}(\lambda_k) \nearrow \frac{2}{c+\sqrt{c^2+2c}} = \xi(c). \tag{54}$$

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

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

is strictly increasing in u. (54) follows. Further, notice that

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

and using the lower bound in (52),

$$\begin{aligned}
\psi_{G,c}^{-1}(\lambda_k) &= \lambda_k(\sqrt{c^2+2c+2/\lambda_k}-c) \\
&\geq \lambda_k(\sqrt{c^2+2c}-c) \\
&\geq \frac{\sqrt{2D_t/\alpha_{\delta,\eta}}}{1+c\sqrt{2D_t/\alpha_{\delta,\eta}}}(\sqrt{c^2+2c}-c) \\
&= \frac{\sqrt{c^2+2c}-c}{1/\sqrt{2D_t/\alpha_{\delta,\eta}}+c} \\
&\geq \frac{\sqrt{c^2+2c}-c}{\sqrt{\alpha_{\delta,\eta}/\log(1/\delta)}+c} \\
&\geq \frac{\sqrt{c^2+2c}-c}{\sqrt{\alpha_{\delta,\eta}}+c}.
\end{aligned}$$

where we've used that $\log(1/\delta) \geq 1$ since $\delta \leq 1/e$. Therefore, for $t \in E_k$,

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

We have thus shown that

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

which completes the proof. □

G.4 Proof of Lemma C.1

First consider $\psi = \psi_{G,c}$. From Lemma G.3 we know that $\psi'(\lambda^*(a)) = a$, i.e.,

$$\frac{\lambda^*(a)(2-c\lambda^*(a))}{2(1-c\lambda^*(a))^2} = a.$$

Solving for $\lambda^*(a)$ with the quadratic equation gives

$$\lambda^*(a) = \frac{\sqrt{2ac+1}-1}{c\sqrt{2ac+1}},$$

so

$$\psi(\lambda^*(a_n)) = \frac{(\sqrt{2ac+1}-1)^2}{2c^2\sqrt{2ac+1}}.$$

As $a = a_n$ grows, we have $\psi(\lambda^*(a)) = O(\sqrt{a_n}/c^2)$. Meanwhile,

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

so $\psi^{-1}(\lambda^*(a_n)) \nearrow \psi^{-1}(1/c)$. Therefore,

$$f(a_n) \asymp \frac{\sqrt{a_n}}{c^{3/2}} \left(\sqrt{\frac{2}{c + \sqrt{c^2 + 2c}}} - 1 \right).$$

Now, for $\psi = \psi_{E,c}$, setting $\psi'(\lambda^*(a)) = a$ gives

$$\lambda^*(a) = \frac{a}{1 + ca}.$$

Hence

$$\psi(\lambda^*(a)) = \frac{\log(1 + ca)}{c^2} - \frac{a}{c(1 + ca)},$$

and as $a = a_n$ grows with n, we have $\psi(\lambda^*(a_n)) \lesssim \log(ca_n))$. Meanwhile, the inverse of ψ obeys

$$\psi^{-1}(u) = \frac{1 + W_0(-e^{-1-uc^2})}{c}, \tag{55}$$

where W_0 is the principal branch of the W Lambert function, which is increasing, continuous, and satisfies $W_0(z)\exp(W_0(z)) = z$ and has range $z \in [-1/e, \infty)$. Hence

$$\psi^{-1}(\lambda^*(a)) = \frac{1 + W_0(-\exp(-1 - \frac{c^2 a}{1+a}))}{c} \nearrow \frac{1 + W_0(-e^{-1-c})}{c} \leq 1/c.$$

Therefore, $f_\psi(\lambda^*(a)) \lesssim \log(ca_n)(\sqrt{1/c} - 1)$, which completes the argument.

G.5 Proof of Proposition C.2

We begin with some properties of the convex conjugate.

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

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

Then,

(1). $\lambda^*(a)$ is well-defined and unique. Further, for all $a \in \mathrm{Im}(\psi')$, $a = \psi'(\lambda^*(a))$ and $\lim_{a\to\infty} \lambda^*(a) = \lambda_{\max}$,

(2). For all $a \in \mathrm{Im}(\psi')$, $\lambda^*(a) = (\psi')^{-1}(a)$, 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 $a \in \mathrm{Im}(\psi')$, let $f(x) = ax - \psi(x)$. Since ψ is convex, f is concave, so f is maximized at x^* satisfying $a = \psi'(x^*)$ (x^* exists and is unique because $a \in \mathrm{Im}(\psi')$ and ψ' is strictly increasing). But $x^* = \lambda^*(a)$ by definition, so $a = \psi'(\lambda^*(a))$. Now, as $a \to \infty$, $\lambda = \lambda^*(a)$ is chosen to maximize $a - \psi'(\lambda)$, which grows as $\lambda \to \lambda_{\max}$ since ψ' is strictly increasing (since ψ is strictly convex). (2) now also follows: since ψ' is strictly increasing it is invertible, so may invert (1) to get (2). For (3), note that $\psi''' \geq 0$ iff ψ'' is increasing iff ψ' is convex, hence its inverse λ^* is concave (and nonnegative). Therefore, for all x, y, we have $\lambda^*(\alpha x + (1-\alpha)y) \geq \lambda^*\ell(x) + (1-\alpha)\lambda^*(y)$ 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 λ^* 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 3.1. Let $\widehat{V}_t = V_t + U_0$ and let $\upsilon = \gamma_{\min}(U_0)$. For all $\lambda \in [\psi^{-1}(1/\upsilon), \lambda_{\max})$, with probability $1-\delta$, we can rearrange the guarantee to give

$$\lambda\|S_\tau\|_{\widehat{V}_\tau^{-1}} - \psi(\lambda) \leq \frac{D_\tau(\delta)}{g_{\psi,\tau}(\lambda)} + g_{\psi,\tau}(\lambda)\psi(\lambda) - \psi(\lambda), \tag{56}$$

where as usual

$$D_t(\delta) = \frac{1}{2}\log\left(\frac{\det \widehat{V}_t}{\det U_0}\right) + 1 + \log(1/\delta).$$

For $j \in \mathbb{N}$ and some $\eta > 1$ to be determined, consider the event $E_j = \{\eta^j \leq \|S_\tau\|_{\widehat{V}_\tau^{-1}} < \eta^{j+1}\}$. Let $E_0 = \{0 \leq \|S_\tau\|_{\widehat{V}_\tau^{-1}} < \eta\}$. Note that $E_0, E_1, \ldots$ partitions the sample space. Our strategy is to apply Theorem 3.1 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 $\|S_\tau\|_{\widehat{V}_\tau^{-1}}$ since we know how this quantity behaves on E_j. Rewriting (56) with $\delta = \delta_j$ and $\lambda = \lambda_j$ gives

$$\lambda_j\|S_\tau\|_{\widehat{V}_\tau^{-1}} - \psi(\lambda_j) \leq h_\tau(\lambda_j)D_\tau(\delta_j) + \psi(\lambda_j)(g(\lambda_j) - 1), \tag{57}$$

where $g(\lambda) = g_{\psi,\tau}(\lambda)$ and $h_\tau(\lambda) = 1/g(\lambda)$. Conditioning on E_j, we have

$$\begin{aligned}\lambda_j\eta^{j+1} - \psi(\lambda_j) &= \lambda_j\|S_\tau\|_{\widehat{V}_\tau^{-1}} - \psi(\lambda_j) + \lambda_j(\eta^{j+1} - \|S_\tau\|_{\widehat{V}_\tau^{-1}}) \\ &\leq h_\tau(\lambda_j)D_\tau(\delta_j) + \psi(\lambda_j)(g(\lambda_j) - 1) + \lambda_j(\eta^{j+1} - \eta^j).\end{aligned} \tag{58}$$

We want to choose an appropriate λ_j on the event E_j. As in Lemma G.3, let

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

and take

$$\lambda_j = \max\left\{\lambda^*(\eta^{j+1}), \psi^{-1}(1/\upsilon)\right\}. \tag{59}$$

Let us suppose for now that $\lambda_j = \lambda^*(\eta^{j+1})$ so that $\lambda_j\eta^{j+1} - \psi(\lambda_j) = \psi^*(\eta^{j+1})$ (by definition of ψ^*). In this case we have

$$\psi(\lambda_j)(g(\lambda_j) - 1) \leq \psi(\lambda_j)\left(\sqrt{\psi^{-1}(\lambda_j)} - 1\right) = f(\eta^{j+1}),$$

where we recall that

$$f(x) = f_\psi(x) = \psi(\lambda^*(x))\left(\sqrt{\psi^{-1}(\lambda^*(x))} - 1\right).$$

Then we can rearrange (58) to read

$$\psi^*(\|S_\tau\|_{\widehat{V}_\tau^{-1}}) \leq \psi^*(\eta^{j+1}) \leq h_\tau(\lambda_j)D_\tau(\delta_j) + \eta^j\lambda_j(\eta - 1) + f(\eta^{j+1}), \tag{60}$$

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

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

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

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

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

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

For such η (60) implies

$$\begin{aligned}\psi^*(\|S_\tau\|_{\widehat{V}_\tau^{-1}}) &\le h_\tau(\lambda_j)D_\tau(\delta_j) + \frac{\psi^*(\eta^j)}{2} + f(\eta^{j+1}) \\ &\le h_\tau(\lambda_j)D_\tau(\delta_j) + \frac{\psi^*(\|S_\tau\|_{\widehat{V}_\tau^{-1}})}{2} + f(\eta\|S_\tau\|_{\widehat{V}_\tau^{-1}}),\end{aligned}$$

whence $\psi^*(\|S_\tau\|_{\widehat{V}_\tau^{-1}}) \le 2h_\tau(\lambda_j)D_\tau(\delta_j) + 2f(2\|S_\tau\|_{\widehat{V}_\tau^{-1}})$, i.e.,

$$\|S_\tau\|_{\widehat{V}_\tau^{-1}} \le (\psi^*)^{-1}\left(2h_\tau(\lambda_j)D_\tau(\delta_j) + 2f(2\|S_\tau\|_{\widehat{V}_\tau^{-1}})\right). \tag{61}$$

Now we return to the scenario when $\lambda_j = \psi^{-1}(1/\upsilon)$ in (59). In this case, we have

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

We have thus shown that

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

where B_τ is the piecewise boundary

$$B_\tau = \begin{cases} (\psi^*)^{-1}\left(2h_\tau(\lambda^*(\eta^{j(\tau)+1})D_\tau(\delta_{j(\tau)}) + 2f(2\|S_\tau\|_{\widehat{V}_\tau^{-1}})\right), & \text{if } \lambda^*(\eta^{j(\tau)+1}) \ge \psi^{-1}(1/\upsilon), \\ \frac{h_\tau(\psi^{-1}(1/\upsilon))D_\tau(\delta_j)+g(\psi^{-1}(1/\upsilon))/\upsilon}{\psi^{-1}(1/\upsilon)}, & \text{otherwise,}\end{cases} \tag{62}$$

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

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

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

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

Therefore, for large enough t, as long as $\psi^{-1}(1/\upsilon) < \lambda_{\max}$ (guaranteed by assumption), 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 $1-\delta$,

$$\|S_\tau\|_{V_\tau^{-1}} \le B_\tau \lesssim (\psi^*)^{-1}\left(h_\tau(\lambda^*(\eta^{j(\tau)+1}))D_\tau(\delta_{j(\tau)}) + f(2\|S_\tau\|_{V_\tau^{-1}})\right). \tag{63}$$

It remains to bound $h_\tau(\lambda^*(\eta^{j(\tau)+1}))$. As was done in the proof of Proposition 3.4, notice that $\alpha_t \le \gamma_{\max}(U_0)/\gamma_{\min}(\widehat{V}_t)$. Then,

$$\begin{aligned} g_{\psi_\tau}(\lambda_j) &= \sqrt{\psi^{-1}(\lambda_j)} - \sqrt{\alpha_t \psi^{-1}(\lambda_j)} \\ &\ge \sqrt{\psi^{-1}(\lambda_j^*)}\left(1 - \sqrt{\frac{\gamma_{\max}(U_0)}{\gamma_{\min}(\widehat{V}_\tau)}}\right) \\ &\ge \sqrt{\psi^{-1}(\psi^{-1}(1/\upsilon))}\left(1 - \sqrt{\frac{\gamma_{\max}(U_0)}{\gamma_{\min}(\widehat{V}_\tau)}}\right). \end{aligned}$$

Hence

$$h_\tau(\lambda^*(\eta^{j(\tau)+1})) \lesssim \frac{1}{\sqrt{1-\gamma_{\max}(U_0)/\gamma_{\min}(\widehat{V}_\tau)}}.$$

Returning to (63), we have that with probability $1-\delta$,

$$\begin{aligned} \|S_\tau\|_{V_\tau^{-1}} &\lesssim (\psi^*)^{-1}\left(\frac{\log(\det V_\tau) + \log\left(\frac{\ell(j(\tau))}{\delta}\right)}{\sqrt{1-\gamma_{\max}(U_0)/\gamma_{\min}(\widehat{V}_\tau)}} + f(2\|S_\tau\|_{\widehat{V}_\tau^{-1}})\right) \\ &\lesssim (\psi^*)^{-1}\left(\frac{\log(\det V_\tau) + \log(\log_\eta(\|S_\tau\|_{\widehat{V}_\tau^{-1}})/\delta)}{\sqrt{1-\gamma_{\max}(U_0)/\gamma_{\min}(\widehat{V}_\tau)}} + f(2\|S_\tau\|_{\widehat{V}_\tau^{-1}})\right), \end{aligned}$$

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

G.6 Proof of Corollary C.3

Since ψ is CGF-like, so too is ψ^* (Whitehouse et al., 2025, Proposition A.1). In particular, ψ^* is strictly convex and increasing, implying that φ is strictly increasing and concave. It follows that for all x, y,

$$\varphi(x+y) - \varphi(x) \le \varphi'(x)y. \tag{64}$$

(A concave function is bounded by its first order Taylor approximation.) Let $\widehat{V}_t = V_t + U_0$ and set

$$A_\tau = \frac{\log(\det \widehat{V}_\tau) + \log(1/\delta)}{\sqrt{1-\upsilon/\gamma_{\min}(\widehat{V}_\tau)}}, \quad B_\tau = \frac{\log\log(\|S_\tau\|_{\widehat{V}_\tau^{-1}})}{\sqrt{1-\upsilon/\gamma_{\min}(\widehat{V}_\tau)}} + f(2\|S_\tau\|_{\widehat{V}_\tau^{-1}}).$$

Now, for $\tau > \tau_0 = \inf\{t : \upsilon/\gamma_{\min}(\widehat{V}_t) \le b\}$ we have $1-\upsilon/\gamma_{\min}(\widehat{V}_\tau) \ge 1-b$. For such τ therefore, we have

$$B_\tau \le \frac{\log\log(\|S_\tau\|_{V_\tau^{-1}})}{\sqrt{1-1/\upsilon}} + f(2\|S_\tau\|_{\widehat{V}_\tau^{-1}}) \le \frac{\|S_\tau\|_{V_\tau^{-1}}(1+2\kappa)}{\sqrt{1-b}},$$

using the assumption that $0 < b < 1$. Taking $x = A_\tau$ and $y = B_\tau$ in (64), we obtain that the right hand side of (22) is bounded as

$$\varphi(A_\tau + B_\tau) \le \varphi(A_\tau) + \varphi'(A_\tau)B_\tau,$$

where we've used that $A_\tau \geq \log(1/\delta)$ for $\tau \geq \tau_0$. Since ψ^* is strictly increasing and convex, $(\psi^*)'$ is also increasing. Therefore the term $(\psi^*)'(\varphi(x))$ is increasing in x, and

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

where the equalities follows from the chain rule. Hence, with probability $1-\delta$, whenever $\tau \geq \tau_0$,

$$\|S_\tau\|_{\widehat{V}_\tau^{-1}} \leq \varphi(A_\tau + B_\tau) \leq \varphi(A_\tau) + \frac{\varphi'(\log(1/\delta))\|S_\tau\|_{\widehat{V}_\tau^{-1}}(1+2\kappa)}{\sqrt{1-b}}.$$

Rearranging gives the desired result.

G.7 Proof of Lemma D.1

Our goal is to upper bound the term $\log \mathbb{E}_{t-1}\exp(\lambda\langle\theta, X\rangle)$ in (24). Since $\log x \leq x-1$ for all $x>0$, we have

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

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

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

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

G.8 Proof of Lemma D.2

In the proof of Lemma D.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 (25) 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

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

is upper bounded by $N_t^B(\theta)$ in (24) and hence is itself a nonnegative supermartingale. As in the proof of Lemma D.1, noticing that we can rewrite $N_t^{\mathrm{Bern}}(\theta)$ as $N_t^{\mathrm{Bern}}(\theta) = \exp\{\lambda\langle\theta, S_t\rangle - \psi_{G,c}(\lambda)\langle\theta, V_t\theta\rangle\}$ for $V_t = \sum_{k\leq t}\mathbb{E}_{k-1}X_kX_k^\intercal$ shows that (S_t, V_t) is sub-$\psi_{G,c}$, and adding any PSD matrix to V_t maintains this property.

H Simulation Details

Figure 1 In both figures we take $d = 10$, $U_0 = I_d$, and $V_t = U_0 + \sum_{k \le t} X_k X_k^\intercal$ 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 $d = 20$ and $c = 1$. For the first 100 time steps, we make a full rank update in order to sufficiently grow the minimum eigenvalue. 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,

$$V_t = V_{t-1} + \sum_{j \le k} u_k u_k^\intercal,$$

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 (the left panel) we use $k = 1$, for moderate growth (middle panel) we use $k = 10$ and for large growth (right panel) we use $k = 19 = d - 1$ (nearly a full-rank update).

We compare Theorem 3.4 to Corollary 2 in Whitehouse et al. (2025), 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 $V_t \succeq \upsilon I_d$ for all t, then, with probability $1 - \delta$, for all t,

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

where

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

We take $\beta = 2$, $\eta = 2$, and $\epsilon = 1/2$, as suggested in their paper. In all plots, we omit the first 100 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, and one from which the bound of Whitehouse et al. (2025) does not suffer. Note that $\psi_{G,1}^{-1}(1) \approx 0.73$, so all our choices of λ are legal.

Figure 3. We take $d = 20$ and $U_0 = 2 \cdot I_d$. For $k \in \{2, 5, 10\}$, we generate X_t by letting the first k components be random uniform noise, and then normalizing the vector such that $\|X_t\|_2 \le 1/2$. We take $\widehat{\mu}_t$ to be the empirical mean of the first t observations, i.e., $\widehat{\mu}_t = t^{-1}\sum_{k \le t} X_k$.

A per the discussion after Corollary 4.1, when instantiating our bound we split U_0 into Γ and U_0'. For these experiments we take $\Gamma = 1.5U_0$, $U_0' = 0.5U_0$. We then apply the result with the process $(S_t, \Gamma + \sum_{k \le t}(X_k - \widehat{\mu}_{k-1})(X_k - \widehat{\mu}_{k-1})^\top$ with the fixed matrix U_0' in place of U_0.

We compare Corollary 4.1 with the empirical Bernstein bound of Whitehouse et al. (2025), which is given by (13). However, for the purposes of the experiments, we use the following version of the bound, stated by Whitehouse et al. (2025) immediately after Theorem 4.1: For any $u > 0$, with probability $1 - \delta$, for all stopping times τ,

$$\|S_\tau\|_{(V_\tau + U_0)^{-1}} \le 2\sqrt{L(V_\tau)} + \frac{2L(V_\tau)}{\gamma_{\min}(V_\tau + U_0)}, \tag{65}$$

where

$$L(V_t) = \log\left(\frac{2}{\delta}\right) + \log\left(h\left(\log_2\left(\frac{\gamma_{\max}(V_t + U_0)}{\gamma_{\min}(U_0)}\right)\right)\right) + \log\left(2\sqrt{\kappa_t} \cdot N_{d-1}\left(\frac{1/4}{\sqrt{\kappa_t}}\right)\right),$$

Despite being a relaxation of their main theorem for particular (unstated) hyperparameters, we find that it is actually tighter for the values of the hyperparameters that we tried, and more stable overall.