Probability Surveys
Vol. 17 (2020) 257–317
ISSN: 1549-5787
https://doi.org/10.1214/18-PS321

Time-uniform Chernoff bounds via nonnegative supermartingales

Steven R. Howard[*]

Department of Statistics
University of California, Berkeley
Berkeley, California 94720, USA
e-mail: stevehoward@berkeley.edu

Aaditya Ramdas

Departments of Statistics and Machine Learning
Carnegie Mellon University
Pittsburgh, PA 15213, USA
e-mail: aramdas@stat.cmu.edu

Jon McAuliffe

Department of Statistics
University of California, Berkeley
Berkeley, California 94720, USA
e-mail: jonmcauliffe@berkeley.edu

Jasjeet Sekhon[†]

Departments of Statistics and Political Science
University of California, Berkeley
Berkeley, California 94720, USA
e-mail: sekhon@berkeley.edu

Abstract: We develop a class of exponential bounds for the probability that a martingale sequence crosses a time-dependent linear threshold. Our key insight is that it is both natural and fruitful to formulate exponential concentration inequalities in this way. We illustrate this point by presenting a single assumption and theorem that together unify and strengthen many tail bounds for martingales, including classical inequalities (1960–80) by Bernstein, Bennett, Hoeffding, and Freedman; contemporary inequalities (1980–2000) by Shorack and Wellner, Pinelis, Blackwell, van de Geer, and de la Peña; and several modern inequalities (post-2000) by Khan, Tropp, Bercu and Touati, Delyon, and others. In each of these cases, we give the strongest and most general statements to date, quantifying the time-uniform concentration of scalar, matrix, and Banach-space-valued martingales, under a variety of nonparametric assumptions in discrete and continuous time. In doing so, we bridge the gap between existing line-crossing inequalities, the sequential probability ratio test, the Cramér-Chernoff method, self-normalized processes, and other parts of the literature.

[*]Supported by Office of Naval Research grant N00014-17-1-2176.
[†]Supported by Office of Naval Research grants N00014-15-1-2367, N00014-17-1-2176.

MSC 2010 subject classifications: Primary 60E15, 60G17; secondary 60F10, 60B20.

Keywords and phrases: Exponential concentration inequalities, nonnegative supermartingale, line crossing probability.

Received November 2018.

Contents

1. Introduction

Concentration inequalities play an important role in probability and statistics, giving non-asymptotic tail probability bounds for random variables or suprema of random processes. In this paper, we consider a method to bound the probability that a martingale ever crosses a time-dependent linear threshold. We were motivated by the fact that such bounds are the key ingredient in many sequential inference procedures. We argue, however, that this formulation is materially better for the development of exponential concentration inequalities, even in some non-sequential settings. We give a master assumption and theorem which handle all of these cases, in discrete and continuous time, for scalar-valued, matrix-valued, and smooth Banach-space-valued martingales. By unifying and organizing dozens of results, we illustrate how these results relate to one another and highlight the specific ingredients contributed by each author. Our improvements to existing results come in the form of weakened assumptions, extension of fixed-time or finite-horizon bounds to infinite-horizon uniform bounds, and improved exponents.

Our main results are presented in full generality in the following section. To motivate these results, we first contrast a small handful of well-known, concrete results from the exponential concentration literature; see Section 1.2 for a more detailed overview of the literature we draw upon. Throughout the paper, most of our results are presented for filtered probability spaces, and we use $\mathbb{E}_t$ to denote expectation conditional on the underlying filtration $\mathcal{F}_t$ at time t. For any discrete-time process $(Y_t)_{t\in\mathbb{N}}$, we write $\Delta Y_t \coloneqq Y_t - Y_{t-1}$ for the increments. Finally, we write $\mathcal{H}^d$ for the space of $d \times d$ Hermitian matrices. The relation $A \preceq B$ denotes the semidefinite order on $\mathcal{H}^d$, while $\gamma_{\max} : \mathcal{H}^d \to \mathbb{R}$ denotes the maximum eigenvalue map.

Example 1. Unless indicated otherwise, let $(S_t)_{t=0}^{\infty}$ be a real-valued martingale with respect to a filtration $(\mathcal{F}_t)_{t=0}^{\infty}$, with $S_0 = 0$.

(a) Three of the earliest and most well-known results for exponential concentration are attributed to Bernstein, Bennett, and Hoeffding. Assume the increments (ΔS_t) are independent, and let $v_t \coloneqq \sum_{i=1}^{t} \mathbb{E}(\Delta S_i)^2$. We present Bernstein's inequality (Bernstein, 1927) in a widely used form (e.g., Boucheron et al., 2013, Corollary 2.11): if, for some fixed $m \in \mathbb{N}$ and $c > 0$, the increments satisfy the moment condition $\sum_{i=1}^{m} \mathbb{E}(\Delta S_t)^k \leq$

$\frac{k!}{2}c^{k-2}v_m$ for all integers $k \geq 3$, then for any $x > 0$, we have

$$\mathbb{P}\left(S_m \geq x\right) \leq \exp\left\{-\frac{x^2}{2(v_m + cx)}\right\}. \tag{1.1}$$

Bernstein's moment condition is easily seen to be satisfied if the increments are bounded. Bennett (1962, eq. 8b) improved Bernstein's result for bounded increments: if $\Delta S_t \leq 1$ for all t, then for any $x > 0$ and $m \in \mathbb{N}$, we have

$$\mathbb{P}\left(S_m \geq x\right) \leq \left(\frac{v_m}{x + v_m}\right)^{x+v_m} e^x. \tag{1.2}$$

Finally, Hoeffding (1963, eq. 2.3) gave a simplified result for increments bounded from above and below: if $|\Delta S_t| \leq 1$ for all t, then for any $x > 0$ and $m \in \mathbb{N}$, we have

$$\mathbb{P}\left(S_m \geq x\right) \leq e^{-x^2/2m}. \tag{1.3}$$

(b) Blackwell (1997, Theorem 1): if $|\Delta S_t| \leq 1$ for all t, then for any $a, b > 0$, we have

$$\mathbb{P}(\exists t \in \mathbb{N} : S_t \geq a + bt) \leq e^{-2ab}. \tag{1.4}$$

Relative to Hoeffding's inequality, Blackwell removes the assumption of independent increments, although this possibility was noted by Hoeffding himself (Hoeffding, 1963, p. 18). More importantly, Blackwell replaces the event $\{S_m \geq x\}$ for fixed time m with the time-uniform event $\{\exists t \in \mathbb{N} : S_t \geq a + bt\}$. To see that Blackwell's result recovers and strengthens that of Hoeffding, set $a = x/2$, $b = x/2m$ and note that Blackwell's uniform bound recovers Hoeffding's bound at time $t = m$, so that Blackwell obtains the same probability bound for a larger event.

(c) Freedman (1975, Theorem 1.6): if $|\Delta S_t| \leq 1$ for all t, then writing $V_t \coloneqq \sum_{i=1}^{t} \mathrm{Var}\left(\Delta S_i \mid \mathcal{F}_{i-1}\right)$, then for any $x, m > 0$, we have

$$\mathbb{P}\left(\exists t \in \mathbb{N} : V_t \leq m \text{ and } S_t \geq x\right) \leq \left(\frac{m}{x + m}\right)^{x+m} e^x. \tag{1.5}$$

Similar to Bernstein's and Bennett's inequalities, but unlike those of Hoeffding and Blackwell, Freedman's inequality measures time in terms of a predictable quantity, the accumulated conditional variance V_t, rather than simply the number of observations t. Freedman's inequality bounds the deviations of (S_t) uniformly over time, but only up to the finite time horizon implied by $V_t \leq m$.

(d) de la Peña (1999, Theorem 6.2, eq. 6.4): if the increments are conditionally symmetric, that is, $\Delta S_t \sim -\Delta S_t \mid \mathcal{F}_{t-1}$ for all t, then letting $V_t = \sum_{i=1}^{t} \Delta S_i^2$, for any $\alpha \geq 0$ and $\beta, x, m > 0$ we have

$$\mathbb{P}\left(\exists t \in \mathbb{N} : V_t \geq m \text{ and } \frac{S_t}{\alpha + \beta V_t} \geq x\right) \leq \exp\left\{-x^2\left(\frac{\beta^2}{2m} + \alpha\beta\right)\right\}. \tag{1.6}$$

A remarkable feature of this result is that we measure time via the adapted quantity V_t. Unlike Freedman's inequality, which uses the true conditional variance to measure time, de la Peña's inequality relies only on empirical quantities. In further contrast to Freedman's inequality, de la Peña's bound holds uniformly over $V_t \geq m$ rather than $V_t \leq m$, and we bound the deviations of the self-normalized process $S_t/(\alpha + \beta V_t)$.

(e) Tropp (2012, Theorem 6.2): departing from the above results for real-valued martingales, here we begin with a martingale $(Y_t)_{t\in\mathbb{N}}$ taking values in $\mathcal{H}^d$. Assume that the increments ΔY_t are independent and, for some fixed $c > 0$ and $\mathcal{H}^d$-valued sequence $(W_t)_{t\in\mathbb{N}}$, the moments of the increments satisfy $\mathbb{E}\left(\Delta S_t^k \mid \mathcal{F}_{t-1}\right) \preceq \frac{k!}{2} c^{k-2} \Delta W_t$ for all t and all $k \geq 2$. Then, writing $S_t = \gamma_{\max}(Y_t)$ and $V_t = \gamma_{\max}(W_t)$, for any $x > 0$ and $t \geq 1$, we have

$$\mathbb{P}\left(S_t \geq x\right) \leq d \cdot \exp\left\{-\frac{x^2}{2(V_t + cx)}\right\}. \tag{1.7}$$

This elegant result extends Bernstein's inequality to the matrix setting. Note the prefactor of d that appears when we bound the deviations of the maximum eigenvalue of a $d \times d$ matrix-valued process.

(f) Finally, we recall a textbook result for Brownian motion (e.g., Durrett, 2017, Exercise 7.5.2): if $(S_t)_{t\in(0,\infty)}$ is a standard Brownian motion, then for any $a, b > 0$, we have

$$\mathbb{P}(\exists t \in (0,\infty) : S_t \geq a + bt) = e^{-2ab}. \tag{1.8}$$

The result closely resembles Blackwell's inequality for discrete-time martingales with bounded increments, but here we have an equality.

Clearly, these results have much in common with each other and with myriad other results from the exponential concentration literature. Examining the proofs, we find many shared ingredients which are now well known: the notions of sub-Gaussian and sub-exponential random variables, the Cramér-Chernoff method, the large-deviations supermartingale, and so on. Nonetheless, there are enough differences among the results and their proofs to leave us wondering whether these results are merely similar in appearance, or whether they are all special cases of some underlying, general argument.

In this paper, we provide a framework that formally unifies the above results along with many others. Our framework consists of two pieces. First, we crystallize the notion of a *sub-ψ process* (Definition 1), a sufficient condition general enough to encompass a broad set of results not previously treated together, yet specific enough to derive a useful set of equivalent concentration inequalities. This definition provides a convenient categorization of exponential concentration results into sub-Bernoulli, sub-Gaussian, sub-Poisson, sub-exponential, and sub-gamma bounds. Second, we give a generalization of the Cramér-Chernoff argument, Theorem 1. This result yields strengthened versions of many existing inequalities and illustrates equivalences among different forms of exponential

bounds. For example, Theorem 1 strengthens both "Freedman-style" inequalities such as (1.5) and "de la Peña-style" inequalities such as (1.6) to hold uniformly over all time, and in these strengthened forms, the two styles of inequalities are shown to be equivalent, as depicted in Figure 1. We remark that the seminal works from which these examples are drawn, like others referenced below, include many other important contributions, and our claims about Theorem 1 refer only to the particular inequalities cited from each work.

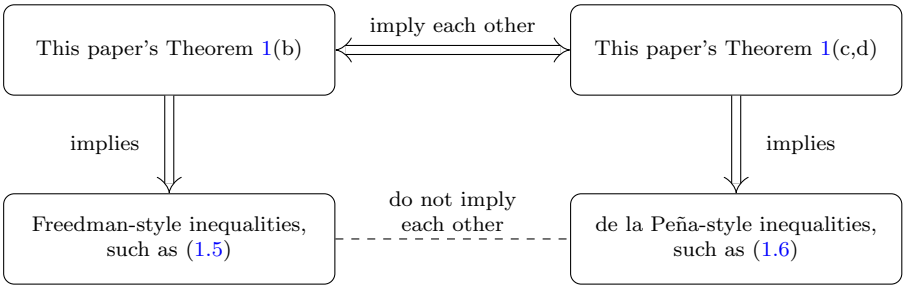

Fig 1: This paper's Theorem 1 implies both Freedman-style inequalities such as (1.5) and de la Peña-style inequalities such as (1.6). Refer also to Figures 4 and 5 for visualizations of these implications.

Once the framework is in place, the proof of the main result follows using tools from classical large-deviation theory (Dembo and Zeitouni, 2010). We construct a nonnegative supermartingale as in Freedman (1975), and we obtain a bound on its entire trajectory using Ville's maximal inequality (Ville, 1939). We invoke Tropp's ideas (Tropp, 2011) to extend the results to the matrix setting. The equivalences that follow from optimizing linear bounds are obtained using convex analysis (Rockafellar, 1970). By drawing together various proof ingredients from different sources, we elucidate previously unrecognized or understated connections. For example, we demonstrate how self-normalized matrix inequalities follow easily upon combining ideas from the literature on self-normalized processes with those from matrix concentration.

1.1. Paper organization

Section 2 lays out our framework for exponential line-crossing inequalities. Specifically, we formally state Definition 1 and Theorem 1 that together describe a general formulation of the Cramér-Chernoff method. After stating Theorem 1, we give a quick overview of existing results which can be recovered in our framework and the improvements thus obtained. A short proof of our master theorem comes next, and following some remarks, we provide three illustrative examples.

Sections 3 and 4 are devoted to a catalog of important results from the literature which fit into our framework, often yielding results which are stronger than those originally published. In Section 3, we consider the maximum-eigenvalue

process of a matrix-valued martingale and enumerate useful sufficient conditions for such a process to be sub-ψ, collecting and in some cases generalizing a variety of ingenious results from the literature. Section 4 examines various instantiations of our master theorem, obtaining corollaries by combining one of the sufficient conditions from Section 3 with one of the four equivalent conclusions of Theorem 1. These illustrate how our framework recovers and strengthens existing exponential concentration results. We discuss sharpness, another geometrical insight, and future work in Section 5. Proofs of most results are in Section 6.

1.2. Historical context

To aid the reader, we give here some historical context for the existing results discussed below. This is not intended to be a comprehensive history of the literature on exponential concentration, and we focus on the specific results discussed in Section 4, giving pointers to further references as appropriate.

The Cramér-Chernoff method takes its name from the works of Cramér (1938) and Chernoff (1952). Both of these authors were concerned with a precise characterization of the asymptotic decay of tail probabilities beyond the regime in which the central limit theorem applies; Cramér provided the first proof of such a "large deviation principle", while Chernoff gave a more general formulation and placed more emphasis on the non-asymptotic upper bound which is our focus. These results spawned a vast literature on large deviation principles, with the goal of giving sharp upper and lower bounds on the limiting exponential decay of certain probabilities under a sequence of measures; see Dembo and Zeitouni (2010) for an excellent presentation of this literature. Our focus, on non-asymptotic upper bounds for nonparametric classes of distributions, is rather different, though such upper bounds often make an appearance in proofs of large deviation principles.

Bernstein was perhaps the earliest proponent of the sort of exponential tail bounds that are the focus of this paper, having proposed his famous inequality in 1911, according to Prokhorov (1995); see also Craig (1933), Uspensky (1937, ch. 10, ex. 12–14, pp. 204–205) and Bernstein (1927), though the last source appears rather inaccessible. The modern theory of exponential concentration began to take shape in the 1960's, as (using the terminology of this paper, from Section 3) Bennett (1962) improved Bernstein's sub-gamma inequality to sub-Bernoulli and sub-Poisson ones for random variables bounded from above. Hoeffding (1963) gave alternative sub-Bernoulli and sub-Gaussian bounds for random variables bounded from both above and below. For further references on this line of work, see Boucheron et al. (2013), whose treatment of the Cramér-Chernoff method has been invaluable in formulating our own framework, as well as McDiarmid (1998).

Godwin (1955, p. 936) reports that Bernstein generalized his inequality to dependent random variables. Hoeffding (1963, pp. 17–18) considered the generalization of his sub-Bernoulli and sub-Gaussian bounds to martingales and the possibility of finite-horizon uniform inequalities based on Doob's maximal

inequality; the martingale generalization was later explored by Azuma (1967). Freedman (1975) extended Bennett's sub-Poisson bound to martingales, giving a uniform bound subject to a maximum value of the predictable quadratic variation of the martingale. This "Freedman-style" bound has been generalized to other settings in many subsequent works (de la Peña, 1999; Khan, 2009; Tropp, 2011; Fan et al., 2015). Chen (2012a,b) has considered the extension of Chernoff-style bounds to hold uniformly over time for scalar-valued martingales in a manner similar to our line-crossing inequalities, including a condition similar to our sub-ψ definition; our formulation further encompasses matrix-valued processes and self-normalized inequalities.

The extension of these methods to matrix-valued processes, via control of the matrix moment-generating function, originated with Ahlswede and Winter (2002). The method was refined by Christofides and Markström (2007), Oliveira (2010a,b) and then by Tropp (2011, 2012), whose influential treatment synthesized and improved upon past work, generalizing many scalar exponential inequalities to operator-norm inequalities for matrix martingales. We have incorporated Tropp's formulation into our framework, and we focus on his theorem statements for our matrix bound statements. See Tropp (2015) for a recent exposition and further references.

There is a long history of investigation of the concentration of Student's t-statistic under non-normal sampling. Efron (1969) gives many references to early work. He also shows, by making use of Hoeffding's sub-Gaussian bound, that the equivalent self-normalized statistic $(\sum_i X_i)/\sqrt{\sum_i X_i^2}$ satisfies a 1-sub-Gaussian tail bound whenever the X_i satisfy a symmetry condition, a result he attributes to Bahadur and Eaton (Efron, 1969, p. 1284). Starting with Logan et al. (1973), there has been a great deal of work on limiting distributions and large deviation principles for self-normalized statistics; see Shao (1997) and references therein. In terms of exponential tail bounds, de la Peña (1999) explored general conditions for bounding the deviations of a martingale, introduced new decoupling techniques (cf. de la Peña and Giné, 1999), and showed that any martingale with conditionally symmetric increments satisfies a self-normalized sub-Gaussian bound with no integrability condition. This work laid the foundation for the type of self-normalized exponential inequalities which we explore in this paper. These methods were extended by de la Peña et al. (2000, 2004), which introduced a general supermartingale "canonical assumption" that is a key precursor of our sub-ψ condition, and initiated a flurry of subsequent activity on self-normalized exponential inequalities (cf. de la Peña et al., 2007; de la Peña, Klass and Lai, 2009). We note in particular inequality (3.9) of de la Peña et al. (2001), which gives an infinite-horizon boundary-crossing inequality based on a mixture extension of their canonical assumption, as well as the multivariate inequalities (3.24) (for a t-statistic) and (3.29) (for general mixture boundaries) given by de la Peña, Klass and Lai (2009). Bercu and Touati (2008) gave a self-normalized sub-Gaussian bound without symmetry by incorporating the conditional quadratic variation, requiring only finite second moments, and some ingenious further extensions have been given by Delyon (2009), Fan et al. (2015), and Bercu et al. (2015), many of which we include in our collection of

sufficient conditions for a process to be sub-ψ (Section 3.2). See de la Peña, Lai and Shao (2009) and Bercu et al. (2015) for further references.

Ville's maximal inequality for nonnegative supermartingales, the technical underpinning of Theorem 1, originates with Ville (1939, p. 101). It is commonly attributed to Doob, though Doob acknowledged Ville's priority extensively in his works, e.g., Doob (1940, pp. 458–460). Mazliak and Shafer (2009) contains further historical discussion and sources.

2. Main results

Let $(S_t)_{t\in\mathcal{T}\cup\{0\}}$ be a real-valued process adapted to an underlying filtration $(\mathcal{F}_t)_{t\in\mathcal{T}\cup\{0\}}$, where either $\mathcal{T} = \mathbb{N}$ for discrete-time processes or $\mathcal{T} = (0,\infty)$ for continuous-time processes. In continuous time, we assume $(\mathcal{F}_t)$ satisfies the "usual hypotheses", namely, that it is right-continuous and complete, and we assume (S_t) is càdlàg; see, e.g., Protter (2005). In a statistical setting, we may think of (S_t) as a summary statistic accumulating over time, for example a cumulative sum of observations, whose deviations from zero we would like to bound under some null hypothesis. In this setting, a bound on the deviations of (S_t) holding uniformly over time can be used to construct an appropriate sequential hypothesis test, a special case of which is Wald's sequential probability ratio test discussed in Section 4.6. We first explain our key condition on (S_t), the sub-ψ condition. We then state, prove, and interpret our master theorem, followed by some more detailed examples of its application.

2.1. The sub-ψ condition

Our key condition on (S_t) is stated in terms of two additional objects. The first object is a real-valued, nondecreasing process $(V_t)_{t\in\mathcal{T}\cup\{0\}}$, also adapted to $(\mathcal{F}_t)$ (and càdlàg in the continuous-time case). It is an "accumulated variance" process which serves as a measure of *intrinsic time*, an appropriate quantity to control the deviations of S_t from zero (Blackwell and Freedman, 1973). The second object is a function $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}$, reminiscent of a cumulant-generating function, which quantifies the relationship between S_t and V_t. The simplest case is when S_t is a cumulative sum of i.i.d., real-valued, mean-zero random variables with distribution F, in which case we take $V_t = t$ and let $\psi(\lambda) = \log \int e^{\lambda x}\, \mathrm{d}F(x)$ be the CGF of F. Our key condition requires that S_t is unlikely to grow too quickly relative to intrinsic time V_t; it generalizes developments from Freedman (1975); de la Peña et al. (2004); Tropp (2011); Chen (2012b), and others.

Definition 1 (Sub-ψ process)**.** Let $(S_t)_{t\in\mathcal{T}\cup\{0\}}$ and $(V_t)_{t\in\mathcal{T}\cup\{0\}}$ be two real-valued processes adapted to an underlying filtration $(\mathcal{F}_t)_{t\in\mathcal{T}\cup\{0\}}$ with $S_0 = V_0 = 0$ a.s. and $V_t \geq 0$ a.s. for all $t \in \mathcal{T}$. For a function $\psi : [0, \lambda_{\max}) \to \mathbb{R}$ and a scalar $l_0 \in [1,\infty)$, we say (S_t) is l_0*-sub-ψ with variance process* (V_t) if, for each $\lambda \in [0, \lambda_{\max})$, there exists a supermartingale $(L_t(\lambda))_{t\in\mathcal{T}\cup\{0\}}$ with respect to $(\mathcal{F}_t)$ such that $L_0(\lambda) \leq l_0$ a.s. and

$$\exp\{\lambda S_t - \psi(\lambda)V_t\} \leq L_t(\lambda) \text{ a.s. for all } t \in \mathcal{T}. \tag{2.1}$$

For given ψ and l_0, we write $\mathbb{S}_\psi^{l_0}$ for the class of pairs of l_0-sub-ψ processes (S_t, V_t):

$$\mathbb{S}_\psi^{l_0} := \{(S_t, V_t) : (S_t) \text{ is } l_0\text{-sub-}\psi \text{ with variance process } (V_t)\}. \tag{2.2}$$

We often say simply that a process is sub-ψ, omitting l_0 from our terminology for simplicity. All examples considered in this paper fit into three cases for the value of l_0: $l_0 = 1$, when deriving one-sided bounds on scalar martingales; $l_0 = 2$, when deriving bounds on the norm of certain Banach-space-valued martingales; or $l_0 = d$, when deriving bounds on the maximum-eigenvalue process of a $d \times d$ matrix-valued martingale. Also, though we often speak of a process (S_t) being sub-ψ, the sub-ψ condition formally applies to the pair (S_t, V_t) and not to the process (S_t) alone, so that meaningful statements are always made in the context of a specific intrinsic time process (V_t).

Definition 1 may at first defy intuition. We can motivate it from several angles:

- Suppose S_t is a scalar-valued martingale whose deviations we wish to bound uniformly over time. We might like to apply Ville's maximal inequality (see Section 2.3), but must first transform S_t into a *nonnegative* supermartingale. It is natural to consider the exponential transform $e^{\lambda S_t}$ for some $\lambda > 0$, which immediately yields a submartingale. Our task, then, is to find some appropriate ψ and (V_t) which "pull down" the submartingale so that the process $\exp\{\lambda S_t - \psi(\lambda)V_t\}$ is a supermartingale. Intuitively, the exponential process $\exp\{\lambda S_t - \psi(\lambda)V_t\}$ measures how quickly S_t has grown relative to intrinsic time V_t, and the free parameter λ determines the relative emphasis placed on the tails of the distribution of S_t, i.e., on the higher moments. Larger values of λ exaggerate larger movements in S_t, and ψ captures how much we must correspondingly exaggerate V_t.
- Consider again the simple case in which S_t is a cumulative sum of i.i.d. draws from a distribution F over the reals with mean zero and CGF $\psi(\lambda) < \infty$ for $\lambda \in [0, \lambda_{\max})$. Then, setting $V_t = t$, we may take $L_t(\lambda)$ equal to the exponential process $\exp\{\lambda S_t - \psi(\lambda)t\}$, which is a martingale in this case, so that the defining inequality of Definition 1 is an equality. The exponential process may be interpreted as the likelihood ratio in an exponential family containing F with sufficient statistic S_t. See Example 2 for a more detailed exposition of this setting and Section 4.6 for more on the connection with exponential families.
- Alternatively, we may begin from the martingale method for concentration inequalities (Hoeffding, 1963; Azuma, 1967; McDiarmid, 1998; Raginsky and Sason, 2012, section 2.2), itself based on the classical Cramér-Chernoff method (Cramér, 1938; Chernoff, 1952; Boucheron et al., 2013, section 2.2). The martingale method starts from an assumption such as $\mathbb{E}\left(e^{\lambda(X_t - \mathbb{E}(X_t \mid \mathcal{F}_{t-1}))} \,\middle|\, \mathcal{F}_{t-1}\right) \leq e^{\psi(\lambda)\sigma_t^2}$ for all $t \geq 1$ and $\lambda \in [0, \lambda_{\max})$. When $\psi(\lambda) = \lambda^2/2$ and $\lambda_{\max} = \infty$ (and the condition holds for $\lambda < 0$ as well), this is the definition of a conditionally sub-Gaussian random variable with variance parameter σ_t^2. When $\psi(\lambda) = \lambda^2/(2(1 - c\lambda))$ and

$\lambda_{\max} = 1/c$, we have the definition of a random variable which is conditionally sub-gamma on the right tail with variance parameter σ_t^2 and scale parameter c (Boucheron et al., 2013). Writing $S_t \coloneqq \sum_{i=1}^t (X_i - \mathbb{E}_{i-1} X_i)$ and $V_t \coloneqq \sum_{i=1}^t \sigma_i^2$, the process $\exp\{\lambda S_t - \psi(\lambda) V_t\}$ is then a supermartingale for each $\lambda \in \mathbb{R}$. For example, if $\Delta S_t \in [a_t, b_t]$ for all t, then (S_t) is 1-sub-ψ with $\psi(\lambda) = \lambda^2/2$ on $\lambda \in [0, \infty)$, and $V_t = \sum_{i=1}^t \left(\frac{b-a}{2}\right)^2$; this fact underlies Example 1(a,b). Or, if $S_t \leq 1$ for all t, then (S_t) is 1-sub-ψ with $\psi(\lambda) = e^\lambda - \lambda - 1$ on $\lambda \in [0, \infty)$, a fact which leads to Example 1(c).

- Unlike the martingale method assumption, Definition 1 allows (V_t) to be adapted rather than predictable, which leads to a variety of self-normalized inequalities (de la Peña, 1999; de la Peña et al., 2004; de la Peña, Lai and Shao, 2009; Bercu et al., 2015; Fan et al., 2015), for example yielding bounds on the deviation of a martingale in terms of its quadratic variation. In this context, Definition 1 is closely related to the "canonical assumption" of de la Peña et al. (2004, eq. 1.6), which requires that $\exp\{\lambda S_t - \Phi(\lambda V_t)\}$ is a supermartingale for certain nonnegative, strictly convex functions Φ. We have found it more useful to separate the second term into $\psi(\lambda) V_t$, though both formulations yield interesting results. For example, if $\Delta S_t \sim -\Delta S_t \mid \mathcal{F}_{t-1}$, then (S_t) is 1-sub-ψ with $\psi(\lambda) = \lambda^2/2$ over $\lambda \in [0, \infty)$, and $V_t = \sum_{i=1}^t \Delta S_t^2$, from which we may obtain Example 1(d).
- Also in contrast to de la Peña et al. (2004), we allow the exponential process to be merely upper bounded by a supermartingale, rather than being a supermartingale itself; this permits us to handle bounds on the maximum eigenvalue process of a matrix-valued martingale, using techniques from Tropp (2011). For example, under the conditions of Example 1(e), the maximum eigenvalue process (S_t) is d-sub-ψ with $\psi(\lambda) = \lambda^2/[2(1 - c\lambda)]$ on $\lambda \in [0, 1/c)$. In this case, the exponential process $\exp\{\lambda S_t - \psi(\lambda) V_t\}$ is not a supermartingale, but is upper bounded by the trace-exponential supermartingale $\operatorname{tr} \exp\{\lambda Y_t - \psi(\lambda) W_t\}$. The initial value of this trace-exponential process is $l_0 = d$, which leads to the pre-factor of d in the bound (1.7).

Section 3 collects a variety of sufficient conditions from the literature for a process to be sub-ψ, including all of the examples given above. These conditions illustrate the broad applicability of Definition 1 in nonparametric settings, i.e., those which restrict the distribution of (S_t) to some infinite-dimensional class, for example all processes with bounded increments, or with increments having finite variance. Even in such nonparametric cases, ψ is still a CGF of some distribution in all of our examples, though this is not required for the most basic conclusion of Theorem 1. Indeed, the full force of Theorem 1 comes into effect only when ψ satisfies certain properties which hold for CGFs of zero-mean, non-constant random variables (Jorgensen, 1997, Theorem 2.3):

Definition 2. A real-valued function ψ with domain $[0, \lambda_{\max})$ is called *CGF-like* if it is strictly convex and twice continuously differentiable with $\psi(0) =$

$\psi'(0_+) = 0$ and $\sup_{\lambda\in[0,\lambda_{\max})} \psi(\lambda) = \infty$. For such a function we define

$$\bar{b} = \bar{b}(\psi) := \sup_{\lambda\in[0,\lambda_{\max})} \psi'(\lambda) \in (0,\infty]. \tag{2.3}$$

In many typical cases we have $\lambda_{\max} = \infty$ and $\bar{b} = \infty$. With Definitions 1 and 2 in place, we are ready to set up and state our main result in the following section.

2.2. The master theorem

To state our main theorem on general exponential line-crossing inequalities, we will make use of the following transforms of ψ:

$$\text{The Legendre-Fenchel transform: } \psi^\star(u) := \sup_{\lambda\in[0,\lambda_{\max})} [\lambda u - \psi(\lambda)], \quad \text{for } u \geq 0. \tag{2.4}$$

$$\text{The "decay" transform: } D(u) := \sup\left\{\lambda \in (0,\lambda_{\max}) : \frac{\psi(\lambda)}{\lambda} \leq u\right\}, \quad \text{for } u \geq 0. \tag{2.5}$$

$$\text{The "slope" transform: } \mathfrak{s}(u) := \frac{\psi(\psi^{\star\prime}(u))}{\psi^{\star\prime}(u)}, \quad \text{for } u \in (0,\bar{b}). \tag{2.6}$$

In the definition of $D(u)$, we take the supremum of the empty set to equal zero instead of the usual $-\infty$. For $u > 0$, this case can arise in general, but not when ψ is CGF-like. Note that $D(u)$ can also be infinite. We call $D(u)$ the "decay" transform because it determines the rate of exponential decay of the upcrossing probability bound in Theorem 1(a) below. We call $\mathfrak{s}(u)$ the "slope" transform because it gives the slope of the linear boundary in Theorem 1(b); this is defined only when ψ is CGF-like. Defining $\mathfrak{s}(0) = 0$ and $\mathfrak{s}(\bar{b}) = \bar{b}$ when $\bar{b} < \infty$, we find that $\mathfrak{s}(u)$ is continuous, strictly increasing, and $0 \leq \mathfrak{s}(u) < u$ on $u \in [0,\bar{b})$ (see Lemma 2). We do not know of other references for the slope transform, or other situations where it arises naturally. Table 2 gives examples of these transforms for some common ψ functions.

Our main theorem has four parts, each of which facilitates comparisons with a particular related literature, as we discuss in Section 4. Recall Definition 1 of the class $\mathbb{S}_\psi^{l_0}$ of l_0-sub-ψ processes, and the underlying filtration $(\mathcal{F}_t)$ to which processes (S_t) and (V_t) are adapted.

Theorem 1. *Fix any $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}$ and $l_0 \in [1,\infty)$.*

(a) For any $a, b > 0$, we have

$$\sup_{(S_t,V_t)\in\mathbb{S}_\psi^{l_0}} \mathbb{P}\left(\exists t \in \mathcal{T} : S_t \geq a + bV_t \mid \mathcal{F}_0\right) \leq l_0 \exp\left\{-aD(b)\right\}. \tag{2.7}$$

Additionally, whenever ψ is CGF-like, the following three statements are equivalent to statement (a).

(b) For any $m > 0$ and $x \in (0, m\bar{b})$, we have

$$\sup_{(S_t,V_t)\in\mathbb{S}_\psi^{l_0}} \mathbb{P}\left(\exists t \in \mathcal{T} : S_t \geq x + \mathfrak{s}\left(\frac{x}{m}\right)\cdot(V_t - m) \,\middle|\, \mathcal{F}_0\right) \leq l_0 \exp\left\{-m\psi^\star\left(\frac{x}{m}\right)\right\}. \quad (2.8)$$

(c) For any $m > 0$ and $x \in (0, \bar{b})$, we have

$$\sup_{(S_t,V_t)\in\mathbb{S}_\psi^{l_0}} \mathbb{P}\left(\exists t \in \mathcal{T} : \frac{S_t}{V_t} \geq x - \left(\frac{x - \mathfrak{s}(x)}{V_t}\right)\cdot(V_t - m) \,\middle|\, \mathcal{F}_0\right) \leq l_0 \exp\left\{-m\psi^\star(x)\right\}. \quad (2.9)$$

(d) For any $m \geq 0$, $x > 0$ and $b > 0$, we have (below we take $m\bar{b} = \infty$ whenever $\bar{b} = \infty$)

$$\sup_{(S_t,V_t)\in\mathbb{S}_\psi^{l_0}} \mathbb{P}\left(\exists t \in \mathcal{T} : V_t \geq m \text{ and } S_t \geq x + b(V_t - m) \mid \mathcal{F}_0\right) \leq \begin{cases} l_0 \exp\left\{-(x - (b\wedge\bar{b})m)D(b)\right\}, & x > m\bar{b} \text{ or } \mathfrak{s}\left(\frac{x}{m}\right) > b \\ l_0 \exp\left\{-m\psi^\star\left(\frac{x}{m}\right)\right\}, & x \leq m\bar{b} \text{ and } \mathfrak{s}\left(\frac{x}{m}\right) \leq b. \end{cases} \quad (2.10)$$

We give a straightforward proof in Section 2.3 that uses only Ville's maximal inequality for nonnegative supermartingales (Ville, 1939) and elementary convex analysis. Theorem 1 can be seen to unify and strengthen many known exponential bounds, showing that we lose nothing in going from a fixed-time to a uniform bound. This includes classical inequalities by Hoeffding (Corollary 1a), Bennett and Freedman (Corollary 1b), and Bernstein (Corollary 1c), along with their matrix extensions due to Tropp and Mackey et al. (Corollary 1a-c); discrete-time scalar line-crossing inequalities due to Blackwell (Corollaries 4 and 5) and Khan (Section 4.2); self-normalized bounds due to de la Peña (Corollaries 6 and 7), Delyon (Corollary 8), Bercu and Touati (Corollary 8), and Fan (Corollary 9); bounds for martingales in smooth Banach spaces due to Pinelis (Corollary 10); continuous-time bounds due to Shorack and Wellner (Corollary 11) and van de Geer (Corollary 11); and Wald's sequential probability ratio test (Corollary 12). Visualizations of how the bounds of Theorem 1 relate to Freedman's and de la Peña's inequalities are provided in Figures 4 and 5. For convenience, Table 1 lists the existing results we recover and our corresponding corollaries, along with ways in which our analysis strengthens conclusions.

For the remainder of the paper after Section 2.3, we will assume $\mathcal{F}_0$ is the trivial σ-field and omit from our notation the conditioning on $\mathcal{F}_0$ in the results of Theorem 1 and its corollaries.

2.3. Proof of Theorem 1

Throughout the proof, we write $\mathbb{P}_0(\cdot)$ for the conditional probability $\mathbb{P}(\cdot \mid \mathcal{F}_0)$. Ville's maximal inequality for nonnegative supermartingales (Ville, 1939) is the

TABLE 1

Some existing results which are strengthened by Theorem 1, as detailed in Section 4. For clarity, we enumerate the different ways in which we strengthen or generalize existing results with the following mnemonics:

[A] Assumptions: we recover the result under weaker conditions on the distributional or dependence structure of the process.

[B] Boundary: we strengthen the result by replacing a fixed-time bound or a finite-horizon constant uniform boundary with an infinite-horizon linear uniform boundary which is everywhere at least as strong (i.e., low) as the fixed-time or finite-horizon bound.

[C] Continuous time: we extend a discrete-time result to include continuous time.

[D] Dimension: we extend a result for scalar process to one for $\mathcal{H}^d$-valued processes, recovering the scalar result at $d = 1$.

[E] Exponent: we improve the exponent in the result's probability bound.

Existing result	Our result	[A]	[B]	[C]	[D]	[E]
Bernstein (1927)	Corollary 1(c)		✓	✓	✓	
Bennett (1962, eq. 8b)	Corollary 1(b)	✓	✓	✓	✓	
Hoeffding (1963, Theorem 2)	Corollary 1(a)	✓	✓		✓	
Freedman (1975, Theorem 1.6)	Corollary 1(b)		✓	✓	✓	
Shorack and Wellner (1986, App. B, Ineq. 1)	Corollary 11(b)		✓			
Pinelis (1994, Theorems 3.4, 3.5)	Corollary 10		✓			
van de Geer (1995, Lemma 2.2)	Corollary 11(c)		✓		✓	
Blackwell (1997, Theorem 1)	Corollary 4(a)	✓		✓	✓	
Blackwell (1997, Theorem 2)	Corollary 5				✓	
Blackwell (1997, Theorem 2)	Corollary 4(b)	✓		✓	✓	
de la Peña (1999, Theorems 6.1, 1.2B (eq. 1.5))	Corollary 6		✓	✓	✓	
de la Peña (1999, Theorem 6.2)	Corollary 7			✓	✓	✓
Bercu and Touati (2008, Theorem 2.1)	Corollary 8		✓		✓	✓
Delyon (2009, Theorem 4)	Corollary 8		✓		✓	
Khan (2009, Theorem 4.2)	Theorem 1(b)		✓	✓	✓	
Khan (2009, Theorem 4.3)	Theorem 1(d)			✓	✓	✓
Tropp (2011, Theorem 1.2)	Corollary 1(b)		✓			
Tropp (2012, Theorem 1.3)	Corollary 1(a)		✓			✓
Tropp (2012, Theorem 1.4)	Corollary 1(c)		✓			
Mackey et al. (2014, Corollary 4.2)	Corollary 1(a)	✓	✓			

foundation of all uniform bounds in this paper. It is an infinite-horizon uniform extension of Markov's inequality:

Lemma 1 (Ville's inequality)**.** *If $(L_t)_{t\in\mathcal{T}\cup\{0\}}$ is a nonnegative supermartingale with respect to the filtration $(\mathcal{F}_t)_{t\in\mathcal{T}\cup\{0\}}$, then for any $a > 0$, we have*

$$\mathbb{P}_0\left(\exists t \in \mathcal{T} : L_t \geq a\right) \leq \frac{L_0}{a}. \tag{2.11}$$

For completeness, we give an elementary proof of Lemma 1 in Section 6.1. Applying Ville's inequality to Definition 1 gives, for any $(S_t, V_t) \in \mathbb{S}_\psi^{l_0}$, $\lambda \in (0, \lambda_{\max})$, and $z \in \mathbb{R}$,

$$\begin{aligned}\mathbb{P}_0\left(\exists t \in \mathcal{T} : \exp\left\{\lambda S_t - \psi(\lambda)V_t\right\} \geq e^z\right)& \\ \leq \mathbb{P}_0\left(\exists t \in \mathcal{T} : L_t \geq e^z\right) \leq L_0 e^{-z} \leq l_0 e^{-z}.&\end{aligned} \tag{2.12}$$

To derive Theorem 1(a) from (2.12), fix $a, b > 0$ and choose $\lambda \in [0, \lambda_{\max})$ such that $\psi(\lambda) \leq b\lambda$, supposing for the moment that some such value of λ exists. Then

$$\mathbb{P}_0\left(\exists t \in \mathcal{T} : S_t \geq a + bV_t\right) = \mathbb{P}_0\left(\exists t \in \mathcal{T} : \exp\left\{\lambda S_t - b\lambda V_t\right\} \geq e^{a\lambda}\right) \tag{2.13}$$

$$\leq \mathbb{P}_0\left(\exists t \in \mathcal{T} : \exp\left\{\lambda S_t - \psi(\lambda) V_t\right\} \geq e^{a\lambda}\right) \tag{2.14}$$

$$\leq l_0 e^{-a\lambda}, \tag{2.15}$$

applying (2.12) in the last step. This bound holds for all choices of λ in the set $\{\lambda \in [0, \lambda_{\max}) : \psi(\lambda)/\lambda \leq b\}$, so to minimize the final bound, we take the supremum over this set, recovering the stated bound $l_0 e^{-aD(b)}$ by the definition of $D(b)$. If no value $\lambda \in [0, \lambda_{\max})$ satisfies $\psi(\lambda) \leq b\lambda$, then $D(b) = 0$ by definition, so that the bound holds trivially. This shows that Definition 1 implies Theorem 1(a).

To complete the proof we will show that the four parts of Theorem 1 are equivalent whenever ψ is CGF-like. We repeatedly use the well-known fact about the Legendre-Fenchel transform that $\psi'^{-1}(u) = \psi^{\star\prime}(u)$ for $0 < u < \bar{b}$, which follows by differentiating the identity $\psi^\star(u) = u\psi'^{-1}(u) - \psi(\psi'^{-1}(u))$. We also require some simple facts about $\psi(\lambda)/\lambda$:

Lemma 2. *Suppose ψ is CGF-like with domain $[0, \lambda_{\max})$.*

(i) $\psi(\lambda)/\lambda < \psi'(\lambda)$ *for all* $\lambda \in (0, \lambda_{\max})$.
(ii) $\lambda \mapsto \psi(\lambda)/\lambda$ *is continuous and strictly increasing on* $\lambda > 0$.
(iii) $\inf_{\lambda \in (0, \lambda_{\max})} \psi(\lambda)/\lambda = \lim_{\lambda \downarrow 0} \psi(\lambda)/\lambda = 0$.
(iv) $\sup_{\lambda \in (0, \lambda_{\max})} \psi(\lambda)/\lambda = \lim_{\lambda \uparrow \lambda_{\max}} \psi(\lambda)/\lambda = \bar{b}$.
(v) $\psi(D(b))/D(b) = b$ *for any* $b \in (0, \bar{b})$, *that is,* $D(b)$ *is the inverse of* $\psi(\lambda)/\lambda$.
(vi) $\mathfrak{s}(u)$ *is continuous, strictly increasing, and* $0 < \mathfrak{s}(u) < u$ *for all* $u \in (0, \bar{b})$.

Proof of Lemma 2. To see (i), write $\psi(\lambda) = \int_0^\lambda \psi'(t)\,\mathrm{d}t < \lambda\psi'(\lambda)$, where the inequality follows since ψ is strictly convex so that ψ' is strictly increasing. For (ii), the function is continuous because ψ is continuous, and differentiating reveals it to be strictly increasing by part (i). L'Hôpital's rule implies (iii) along with the assumptions $\psi(\lambda) = \psi'(\lambda) = 0$ at $\lambda = 0$, and implies (iv) along with the CGF-like assumption $\sup_\lambda \psi(\lambda) = \infty$, which means $\psi(\lambda) \uparrow \infty$ as $\lambda \uparrow \lambda_{\max}$ since ψ is convex. Part (v) follows from the definition of $D(\cdot)$ and parts (ii), (iii) and (iv). To obtain (vi), note that $\mathfrak{s}$ is the composition of $\lambda \mapsto \psi(\lambda)/\lambda$ with $\psi^{\star\prime}$. Both of these are continuous and strictly increasing, the former by part (ii) and the latter since $\psi^{\star\prime} = \psi'^{-1}$ and ψ' is continuous and strictly increasing by the CGF-like assumption. As $u \downarrow 0$, we have $\psi^{\star\prime}(u) = \psi'^{-1}(u) \downarrow 0$, so $\mathfrak{s}(u) \downarrow 0$ since $\psi(0) = \psi'(0_+) = 0$. Likewise, if $\bar{b} < \infty$, then as $u \uparrow \bar{b}$, $\psi^{\star\prime}(u) \uparrow \lambda_{\max}$ and $\mathfrak{s}(u) \uparrow \bar{b}$. Hence $\mathfrak{s}(u)$ is continuous as defined. Next, note that $\psi(u) > 0$ for $u > 0$ since ψ is strictly convex with $\psi(0) = \psi'(0_+) = 0$, and $\psi^{\star\prime}(u) = \psi'^{-1}(u) > 0$ since $\psi'(\lambda)$ increases from zero at $\lambda = 0$ to $\bar{b}$ as $\lambda \uparrow \lambda_{\max}$. Hence $\mathfrak{s}(u) > 0$ for $u > 0$. Finally, use part (i) to write $\mathfrak{s}(u) = \psi(\psi^{\star\prime}(u))/\psi^{\star\prime}(u) < \psi'(\psi^{\star\prime}(u)) = u$, using the fact that $\psi^{\star\prime}(u) = \psi'^{-1}(u)$ for $u \in (0, \bar{b})$. □

Lemma 2 allows us to prove the equivalences among the parts of Theorem 1 as follows.

- $(a) \Rightarrow (b)$: Fix $m > 0$ and $x \in (0, m\bar{b})$. Any line with slope $b \in (0, x/m)$ and intercept $x - bm$ passes through the point (m, x) in the (V_t, S_t) plane, and part (a) yields

$$\mathbb{P}_0\left(\exists t \in \mathcal{T} : S_t \geq x + b(V_t - m)\right) \leq l_0 \exp\left\{-(x - bm)D(b)\right\} \tag{2.16}$$

$$= l_0 \exp\left\{-m\left(\frac{x}{m} \cdot D(b) - \psi(D(b))\right)\right\} \tag{2.17}$$

  using Lemma 2(v) in the second step. Now we choose the slope b to minimize the probability bound. The unconstrained optimizer $b_\star$ satisfies $\psi'(D(b_\star)) = x/m$, and a solution is guaranteed to exist by our restriction on x. This solution is given by $D(b_\star) = \psi'^{-1}(x/m) = \psi^{\star\prime}(x/m)$. Hence $b_\star = \mathfrak{s}(x/m)$ using Lemma 2(v) and the definition of $\mathfrak{s}(\cdot)$. Lemma 2(vi) shows $0 < b_\star < x/m$, verifying that $b_\star$ is feasible. Identify the Legendre-Fenchel transformation $\psi^\star(x/m) = (x/m)D(b_\star) - \psi(D(b_\star))$ to complete the proof of part (b).
- $(b) \Rightarrow (c)$: Fix $m > 0$ and $x \in (0, \bar{b})$ and observe that

$$\mathbb{P}_0\left(\exists t \in \mathcal{T} : \frac{S_t}{V_t} \geq x - \left(\frac{x - \mathfrak{s}(x)}{V_t}\right) \cdot (V_t - m)\right)$$

$$= \mathbb{P}_0\left(\exists t \in \mathcal{T} : S_t \geq mx + \mathfrak{s}(x) \cdot (V_t - m)\right). \tag{2.18}$$

  Now applying part (b) with values m and mx yields part (c).
- $(c) \Rightarrow (a)$: Fix $a, b > 0$. Suppose first that $b < \bar{b}$, and set $x = \psi'(D(b))$ and $m = a/(x - \mathfrak{s}(x))$. Recalling $\psi^{\star\prime} = \psi'^{-1}$ we see that $\mathfrak{s}(x) = \psi(D(b))/D(b) = b$ by Lemma 2(v). Also, Lemma 2(vi) shows that $m > 0$. Now apply part (c) to obtain

$$\mathbb{P}_0\left(\exists t \in \mathcal{T} : S_t \geq a + bV_t\right) \leq l_0 \exp\left\{-a \cdot \frac{\psi^\star(x)}{x - \mathfrak{s}(x)}\right\} \tag{2.19}$$

$$= l_0 \exp\left\{-a \cdot \frac{\psi^\star(x) \cdot \psi^{\star\prime}(x)}{x\psi^{\star\prime}(x) - \psi(\psi^{\star\prime}(x))}\right\}. \tag{2.20}$$

  Recognizing the Legendre-Fenchel transform in the denominator of the final exponent, we see that the probability bound equals $l_0 \exp\left\{-a\psi^{\star\prime}(x)\right\}$. Again using $\psi^{\star\prime}(x) = \psi'^{-1}(x) = D(b)$ yields part (a).
  If instead $b \geq \bar{b}$, then the above argument yields

$$\mathbb{P}_0\left(\exists t \in \mathcal{T} : S_t \geq a + bV_t\right) \leq \inf_{b' < \bar{b}} \mathbb{P}_0\left(\exists t \in \mathcal{T} : S_t \geq a + b'V_t\right)$$

$$\leq l_0 \exp\left\{-a \sup_{b' < \bar{b}} D(b')\right\}. \tag{2.21}$$

  But $\sup_{b' < \bar{b}} D(b') = \lambda_{\max} = D(b)$ from the definition of $D(\cdot)$.

- $(a) \Rightarrow (d)$: Fix $m \geq 0$ and $x, b > 0$. Observe that $\{\exists t \in \mathcal{T} : V_t \geq m,\, S_t \geq x + b(V_t - m)\} \subseteq \{\exists t \in \mathcal{T} : S_t \geq x' + b'(V_t - m)\}$ for any $0 < x' \leq x$ and $0 < b' \leq b$, so part (a) yields

$$\mathbb{P}_0\left(\exists t \in \mathcal{T} : V_t \geq m,\, S_t \geq x + b(V_t - m)\right) \leq l_0 \exp\left\{-(x' - b'm)D(b')\right\} \tag{2.22}$$

  for any (x', b') in the feasible set $\{x' \in (0, x], b' \in (0, b] : x' > mb'\}$. If $x > m\bar{b}$, then $(x, b \wedge \bar{b})$ is feasible; note that $D(b \wedge \bar{b}) = D(b)$ by the definition of $D(\cdot)$. If $x \leq m\bar{b}$ and $b < s(x/m)$, then by Lemma 2(vi) and the definition $\mathfrak{s}(\bar{b}) \coloneqq \bar{b}$, we have $b < x/m$, so (x, b) is feasible and $b \leq \bar{b}$. Combining these two cases, we have

$$\mathbb{P}_0\left(\exists t \in \mathcal{T} : V_t \geq m,\, S_t \geq x + b(V_t - m)\right) \leq l_0 \exp\left\{-(x - (b \wedge \bar{b})m)D(b)\right\} \tag{2.23}$$

  whenever $x > m\bar{b}$ or $b < s(x/m)$, proving the first case in (2.10). On the other hand, if $x \leq m\bar{b}$ and $s(x/m) \leq b$, then $(x', s(x'/m))$ is feasible for any $x' < x$, by Lemma 2(vi). This yields

$$\mathbb{P}_0\left(\exists t \in \mathcal{T} : V_t \geq m,\, S_t \geq x + b(V_t - m)\right) \leq l_0 \exp\left\{-m\psi^\star\left(\frac{x'}{m}\right)\right\} \tag{2.24}$$

  as in part (b). We minimize the probability bound over $x' < x$, noting that $\sup_{x' < x} \psi^\star(x'/m) = \psi^\star(x/m)$ since $\psi^\star$ is increasing (as ψ is CGF-like) and closed (Rockafellar, 1970, Theorem 12.2). This proves the second case in (2.10).
- $(d) \Rightarrow (a)$: set $m = 0$ and $x = a$ to recover part (a). □

It is worth noting here that, unlike the proofs of Freedman (1975), Khan (2009), Tropp (2011), and Fan et al. (2015), we do not explicitly construct a stopping time in our proof. While an optional stopping argument is hidden within the proof of Ville's inequality, the underlying stopping time here is different from that in the aforementioned citations.

2.4. Interpreting the theorem

It is instructive to think of the parts of Theorem 1 as statements about the process (V_t, S_t) or $(V_t, S_t/V_t)$ in $\mathbb{R}^2$. Many of our results are better understood via this geometric intuition. Specifically, Figure 2 illustrates the following points:

- Theorem 1(a) takes a given line $a + bV_t$ and bounds its S_t-upcrossing probability.
- Theorem 1(b) takes a point (m, x) in the (V_t, S_t)-plane and, out of the infinitely many lines passing through it, chooses the one which yields the tightest upper bound on the corresponding S_t-upcrossing probability.

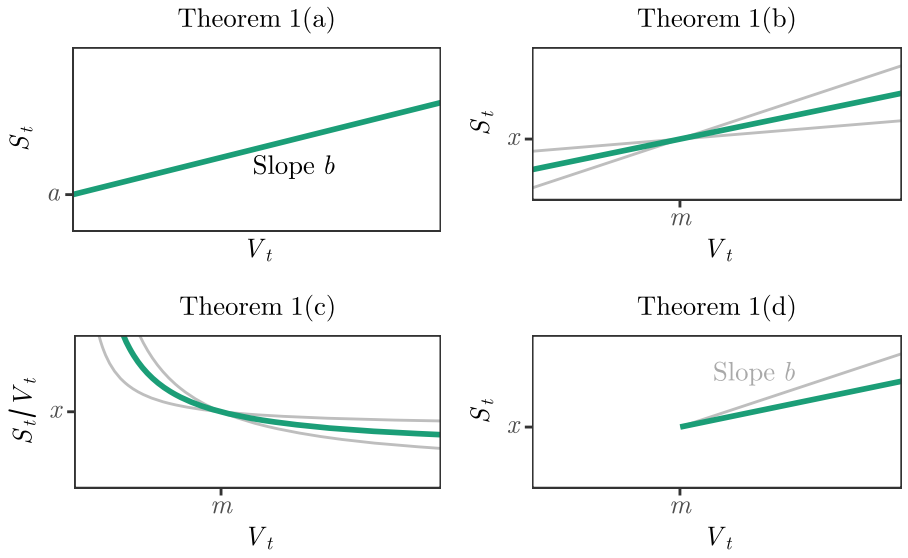

Fig 2: Illustration of the equivalent statements of Theorem 1, as described in the text.

- Theorem 1(c) is like part (b), but instead of looking at S_t, we look at S_t/V_t, fix a point (m, x) in the $(V_t, S_t/V_t)$-plane, and choose from among the infinitely many curves $b + a/V_t$ passing through it to minimize the probability bound.
- The intuition for Theorem 1(d) is as follows. If we want to bound the upcrossing probability of the line $(x-bm)+bV_t$ on $\{V_t \geq m\}$, we can clearly obtain a conservative bound from Theorem 1(a) with $a = x - bm$. This yields the first case in (2.10). However, we can also apply Theorem 1(b) with the values m, x, obtaining a bound on the upcrossing probability for a line which passes through the point (m, x) in the (V_t, S_t)-plane, and this line yields the minimum possible probability bound among all lines passing through (m, x). If the slope of this line, $\mathfrak{s}(x/m)$, is less than b, then this optimal probability bound is conservative for the upcrossing probability over the original line $x + b(V_t - m)$ on $\{V_t \geq m\}$. This gives the second case in (2.10), which is guaranteed to be at least as small as the bound in the first case when $\mathfrak{s}(x/m) \leq b$.

We make some additional remarks below:

- We extend bounds for discrete-time scalar-valued processes to include both discrete-time matrix-valued processes and continuous-time scalar-valued processes, but we do not handle continuous-time matrix-valued processes, as this seems to require further technical developments beyond the scope of this paper (see Bacry et al. (2018) for one approach to exponential

bounds in this case). We write [C or D] when discussing extensions to existing results to emphasize this fact (see Table 1).
- Most of this paper is concerned with right-tail bounds, hence the restriction to $\lambda \geq 0$ in Definition 1. It is understood that identical techniques yield left-tail bounds upon verifying that Definition 1 holds for $(-S_t)$.
- The purpose of excluding ψ being CGF-like from Definition 1 is to separate the truth of statement (a), which follows solely from Definition 1, from its equivalence to (b), (c), and (d), which follows from ψ being CGF-like.

2.5. Three simple examples

We illustrate some simple instantiations of our theorem with three examples: a sum of coin flips, a discrete-time concentration inequality for random matrices, and a continuous-time scalar Brownian motion. These examples make use of several results from Section 3 describing conditions under which a process is sub-ψ; such results may be taken for granted on a first reading.

Example 2 (Coin flipping)**.** Suppose $X_i \overset{\text{iid}}{\sim} \text{Ber}(p)$, and let $S_t = \sum_{i=1}^{t}(X_i - p)$ denote the centered sum. The CGF of each increment of S_t, scaled by $1/[p(1-p)]$, is $\psi_B(\lambda) \coloneqq [p(1-p)]^{-1} \log \mathbb{E} \exp\{\lambda(X_i - p)\} = [p(1-p)]^{-1} \log(pe^{(1-p)\lambda} + (1-p)e^{-p\lambda})$, so that $\lambda_{\max} = \infty$ and $\bar{b} = 1/p$. One may directly check the martingale property to confirm that $L_t(\lambda) \coloneqq \exp\{\lambda S_t - \psi_B(\lambda)p(1-p)t\}$ is a martingale for any λ, so that (S_t) is 1-sub-ψ_B with $V_t = p(1-p)t$. Then, for any $t_0 \in \mathbb{N}$ and $x \in (0, (1-p)t_0)$, setting $m = p(1-p)t_0$ in Theorem 1(b) yields

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t \geq x + p(1-p)\mathfrak{s}_B\left(\frac{x}{p(1-p)t_0}\right) \cdot (t - t_0)\right) \tag{2.25}$$

$$\leq \exp\left\{-t_0 \text{KL}\left(p + \frac{x}{t_0} \,\middle\|\, p\right)\right\} \tag{2.26}$$

$$= \left[\left(\frac{p}{p + x/t_0}\right)^{p+x/t_0} \left(\frac{1-p}{1-p-x/t_0}\right)^{1-p-x/t_0}\right]^{t_0}. \tag{2.27}$$

Here KL denotes the Bernoulli Kullback-Leibler divergence,

$$\text{KL}(q \,\|\, p) = q \log\left(\frac{q}{p}\right) + (1-q)\log\left(\frac{1-q}{1-p}\right). \tag{2.28}$$

It takes some algebra to obtain this KL as the Legendre-Fenchel transform of ψ_B; in Table 2 we summarize all such transforms used in this paper. The final expression is Equation (2.1) of Hoeffding (1963), but here we have a bound not just for the deviation of S_m above its expectation at the fixed time m, but for the upper deviations of S_t for all $t \in \mathbb{N}$, simultaneously. We can use this to sequentially test a hypothesis about p, or to construct a sequence of confidence intervals for p possessing a coverage guarantee holding uniformly over unbounded time.

The slope transform $\mathfrak{s}_B(u)$ for ψ_B, given in Table 2, is unwieldy. To derive a more analytically convenient bound, we use the fact that $p(1-p)\psi_B(\lambda) \le \lambda^2/8$ for all $\lambda \ge 0$; see the proof of Proposition 2, part 2. Hence $\exp\{\lambda S_t - \lambda^2 t/8\} \le L_t(\lambda)$ with L_t defined as above, so (S_t) is also 1-sub-ψ with $\psi(\lambda) = \lambda^2/8$ and $V_t = t$. Now Theorem 1(b) yields

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t \ge x + \frac{x}{2m} \cdot (t-m)\right) \le \exp\left\{-\frac{2x^2}{m}\right\}. \tag{2.29}$$

This is equivalent to Blackwell's line-crossing inequality (1.4), and in the form (2.29) it is clear that it recovers Hoeffding's inequality at the fixed time $t = m$. Instead of using $p(1-p)\psi_B(\lambda) \le \lambda^2/8$, we might alternatively use $\psi_B(\lambda) \le (1-2p)^{-2}(e^{(1-2p)\lambda} - (1-2p)\lambda - 1)$; see the proof of Proposition 2, part 3. This will yield a uniform extension of Bennett's inequality (1.2) which improves upon Hoeffding's inequality substantially for values of p near zero and one. We will see other examples of such "sub-Poisson" bounds below.

Example 3 (Covariance estimation for a spiked random vector ensemble)**.** The estimation of a covariance matrix via an i.i.d. sample is a common application of exponential matrix concentration, starting with Rudelson (1999). See also Vershynin (2012), Gittens and Tropp (2011), Tropp (2015), and Koltchinskii and Lounici (2017) for more recent treatments; this particular example is drawn from Wainwright (2017). Let $d \ge 2$ and consider $\mathbb{R}^d$-valued, mean-zero observations $X_i = \sqrt{d}\xi_i e_{U_i}$, where $\xi_i \overset{\text{iid}}{\sim}$ Rademacher, $(e_k)_{k=1}^d$ are the standard basis vectors and $U_i \overset{\text{iid}}{\sim} \text{Unif}\{1,\dots,d\}$. What can we say about the concentration of the sample covariance matrix $\widehat{\Sigma}_t \coloneqq t^{-1}\sum_{i=1}^t X_i X_i^T$ around the true covariance I_d, the $d \times d$ identity matrix? Let $\gamma_{\max}(A)$ denote the maximum eigenvalue of a matrix A. We have $\gamma_{\max}(X_i X_i^T - I_d) = d - 1$ always, and $\mathbb{E}(X_i X_i^T - I_d)^2 = \left(\frac{(d-1)^2}{d}\right) I_d$. Hence Fact 1(c) shows that $S_t = t\gamma_{\max}(\widehat{\Sigma}_t - I_d)$ is d-sub-ψ with variance process $V_t = \frac{(d-1)^2 t}{d}$, where

$$\psi(\lambda) = \frac{e^{(d-1)\lambda} - (d-1)\lambda - 1}{(d-1)^2} \le \frac{\lambda^2}{2(1-(d-1)\lambda/3)}. \tag{2.30}$$

Here the inequality holds for all $\lambda \in [0, 3/(d-1))$ as demonstrated in the proof of Proposition 2, part 5. Applying Theorem 1(c) with ψ equal to the final expression in (2.30), we obtain, after some algebra, for any $x, m > 0$,

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \gamma_{\max}\left(\widehat{\Sigma}_t - I_d\right) \ge x\left(\frac{1 + \frac{m}{t}\sqrt{1 + 2x/3(d-1)}}{1 + \sqrt{1 + 2x/3(d-1)}}\right)\right) \le d\exp\left\{-\frac{mx^2}{2(d-1)\left[(d-1)/d + x/3\right]}\right\}. \tag{2.31}$$

At the fixed time $t = m$, this implies

$$\gamma_{\max}\left(\widehat{\Sigma}_m - I_d\right) \le \sqrt{\frac{2(d-1)^2\log(d/\alpha)}{dm}} + \frac{2(d-1)\log(d/\alpha)}{3m} \tag{2.32}$$

with probability at least $1-\alpha$, a known fixed-sample result (Wainwright, 2017). However, as above, (2.31) gives a bound on the upper deviations of $\widehat{\Sigma}_t$ for all $t \in \mathbb{N}$ simultaneously. Such a bound enables, for example, sequential hypothesis tests concerning the true covariance matrix.

Example 4 (Line-crossing for Brownian motion)**.** Let $(S_t)_{t\in[0,\infty)}$ denote standard Brownian motion. It is a standard fact that the process $\exp\left\{\lambda S_t - \lambda^2 t/2\right\}$ is a martingale, so that (S_t) is 1-sub-ψ with $\psi(\lambda) = \lambda^2/2$ and $V_t = t$. In this case, Theorem 1 says that, for any $a, b > 0$,

$$\mathbb{P}\left(\exists t \in (0,\infty) : S_t \geq a + bt\right) \leq e^{-2ab}, \tag{2.33}$$

a well-known line-crossing bound for Brownian motion, which in fact holds with equality (Durrett, 2017, Exercise 7.5.2).

3. Sufficient conditions for sub-ψ processes

Much of the power of Definition 1 comes from the array of sufficient conditions for it which have been discovered under diverse, nonparametric conditions. In this section, we define some standard ψ functions and collect a broad set of conditions from the literature for a process (S_t) to be sub-ψ with one of these functions, summarized in Tables 3 and 4. In other words, we collect here some families of process pairs (S_t, V_t) which are contained within $\mathbb{S}_\psi^{l_0}$ for standard choices of ψ. All discrete-time results in this paper use $S_t = \gamma_{\max}(Y_t)$ where $(Y_t)_{t\in\mathbb{N}}$ is a martingale taking values in $\mathcal{H}^d$, with the exception of Section 4.4, which deals with martingales in abstract Banach spaces. Typically, setting $d = 1$ recovers the corresponding known scalar result exactly. We note also that our results for Hermitian matrices extend directly to rectangular matrices using Hermitian dilations (Tropp, 2012), as we illustrate in Corollary 2.

3.1. Five useful ψ functions

We define five particular ψ functions corresponding to five sub-ψ cases: the sub-Gaussian case in Hoeffding's inequality, the "sub-gamma" case corresponding to Bernstein's inequality, the sub-Poisson case from Bennett's and Freedman's inequalities, and the sub-exponential and sub-Bernoulli cases which are used in several other existing bounds. The ψ functions and corresponding transforms for these five cases are summarized in Table 2, while Figure 3 summarizes relationships among these cases, with Proposition 2 containing the formal statements. Recall $\bar{b} = \sup_{\lambda\in[0,\lambda_{\max})} \psi'(\lambda)$ from Definition 2, and note that we take $1/0 = \infty$ by convention in the expressions for $\lambda_{\max}$ and $\bar{b}$ below.

1. We say (S_t) is *sub-Bernoulli* with range parameters $g, h > 0$ when it is sub-$\psi_{B,g,h}$ for some suitable variance process (V_t), where

$$\psi_{B,g,h}(\lambda) \coloneqq \frac{1}{gh}\log\left(\frac{ge^{h\lambda} + he^{-g\lambda}}{g+h}\right) \quad \text{for } 0 \leq \lambda < \infty = \lambda_{\max}, \tag{3.1}$$

which is the scaled CGF of a mean-zero random variable taking values $-g$ and h. Here $\bar{b} = 1/g$.

2. We say (S_t) is *sub-Gaussian* when it is sub-ψ_N for some suitable variance process (V_t), where
$$\psi_N(\lambda) := \lambda^2/2 \quad \text{for } 0 \le \lambda < \infty = \lambda_{\max}. \tag{3.2}$$
Here $\bar{b} = \infty$.
3. We say (S_t) is *sub-Poisson* with scale parameter $c \in \mathbb{R}$ when it is sub-$\psi_{P,c}$ for some suitable variance process (V_t), where
$$\psi_{P,c}(\lambda) := \frac{e^{c\lambda} - c\lambda - 1}{c^2} \quad \text{for } 0 \le \lambda < \infty = \lambda_{\max}. \tag{3.3}$$
By taking the limit, we define $\psi_{P,0} = \psi_N$. Here $\bar{b} = |c \wedge 0|^{-1}$.
4. We say (S_t) is *sub-gamma* with scale parameter $c \in \mathbb{R}$ when it is sub-$\psi_{G,c}$ for some suitable variance process (V_t), where
$$\psi_{G,c}(\lambda) := \frac{\lambda^2}{2(1 - c\lambda)} \quad \text{for } 0 \le \lambda < \frac{1}{c \vee 0} = \lambda_{\max}, \tag{3.4}$$
Here $\bar{b} = |2c \wedge 0|^{-1}$.
5. We say (S_t) is *sub-exponential* with scale parameter $c \in \mathbb{R}$ when it is sub-$\psi_{E,c}$ for some suitable variance process (V_t), where
$$\psi_{E,c}(\lambda) := \frac{-\log(1 - c\lambda) - c\lambda}{c^2}, \quad \text{for } 0 \le \lambda < \frac{1}{c \vee 0} = \lambda_{\max}. \tag{3.5}$$
By taking the limit, we define $\psi_{E,0} = \psi_N$. Here $\bar{b} = |c \wedge 0|^{-1}$.

We will typically write ψ_B, ψ_P, ψ_G, and ψ_E, omitting the range or scale parameters from the notation when they are clear from the context. Whenever Definition 1 holds, one can multiply ψ by any positive constant and divide V_t by the same constant without affecting (2.1); all of the above functions are scaled so that $\psi''(0_+) = 1$ by convention. We follow the definition of sub-gamma from Boucheron et al. (2013), despite the somewhat inconsistent terminology: unlike the other four cases, ψ_G is not the CGF of a gamma-distributed random variable. It is convenient for a number of reasons: it includes ψ_N as a special case, it gives a useful upper bound for ψ_P (see Proposition 2 part 5, below), it falls naturally out of the use of a Bernstein condition on higher moments to bound the CGF, and it is simple enough to permit analytically tractable results for the slope and decay transforms and the various bounds to follow. We remark also that our definition of sub-exponential in terms of the CGF of the exponential distribution follows that of Boucheron et al. (2013, Exercise 2.22), but differs from another well-known definition which says that the CGF is bounded by $\lambda^2/2$ for λ in some neighborhood of zero. The two are equivalent up to appropriate choice of constants, as detailed in Appendix E.

The sub-gamma and sub-exponential functions $\psi_{G,c}$ and $\psi_{E,c}$ possess the following universality property, which we prove in Section 6.2.

TABLE 2

Summary of common ψ functions and related transforms. KL denotes the Bernoulli Kullback-Leibler divergence, $KL(q \,\|\, p) = q \log\left(\frac{q}{p}\right) + (1-q)\log\left(\frac{1-q}{1-p}\right)$. For the gamma and exponential cases, the domain of ψ is bounded by $\lambda_{\max} = 1/(c \vee 0)$; for the other three cases, $\lambda_{\max} = \infty$. For the Bernoulli, Poisson, and exponential cases, a closed-form expression for $D(u)$ is not available, but we give lower bounds based on Proposition 2; $\varphi(g,h)$ is defined in (3.7).

Name	$\psi(\lambda)$	Legendre-Fenchel transform $\psi^\star(u)$	Slope transform $\mathfrak{s}(u)$	Decay transform $D(u)$
Bernoulli $\psi_B, \mathfrak{s}_B, D_B$	$\frac{1}{gh}\log\left(\frac{ge^{h\lambda}+he^{-g\lambda}}{g+h}\right)$	$\frac{1}{gh}\mathrm{KL}\left(\frac{g(1+hu)}{g+h}\,\middle\|\,\frac{g}{g+h}\right)$	$\frac{h\log(1-gu)^{-1}-g\log(1+hu)}{gh(\log(1-gu)^{-1}+\log(1+hu))}$	$\geq \frac{2ghu}{\varphi(g,h)}$
Gaussian/ normal $\psi_N, \mathfrak{s}_N, D_N$	$\lambda^2/2$	$u^2/2$	$u/2$	$2u$
Poisson $\psi_P, \mathfrak{s}_P, D_P$	$\frac{e^{c\lambda}-c\lambda-1}{c^2}$	$\frac{(1+cu)\log(1+cu)-cu}{c^2}$	$\frac{cu-\log(1+cu)}{c\log(1+cu)}$	$\geq \frac{2u}{1+2cu/3}$
"Gamma" $\psi_G, \mathfrak{s}_G, D_G$	$\frac{\lambda^2}{2(1-c\lambda)}$	$\frac{u^2}{1+cu+\sqrt{1+2cu}}$	$\frac{u}{1+\sqrt{1+2cu}}$	$\frac{2u}{1+2cu}$
Exponential $\psi_E, \mathfrak{s}_E, D_E$	$\frac{\log(1-c\lambda)^{-1}-c\lambda}{c^2}$	$\frac{cu-\log(1+cu)}{c^2}$	$\frac{(1+cu)\log(1+cu)-cu}{c^2u}$	$\geq \frac{2u}{1+2cu}$

Proposition 1. *For any twice-differentiable $\psi : [0, \lambda_{\max}) \to \mathbb{R}$ with $\psi(0) = \psi'(0_+) = 0$, there exist constants $a, c > 0$ such that $\psi(\lambda) \leq a\psi_{G,c}(\lambda)$ for all $\lambda \in [0, \lambda_{\max})$. Likewise, there exists constants $\tilde{a}, \tilde{c} > 0$ such that $\psi(\lambda) \leq \tilde{a}\psi_{E,\tilde{c}}(\lambda)$ for all $\lambda \in [0, \lambda_{\max})$.*

In particular, this means that if $S_t = \sum_{i=1}^t X_i$ for any zero-mean, i.i.d. sequence (X_i) satisfying $\mathbb{E}e^{\lambda X_1} < \infty$ for some $\lambda > 0$, then (S_t) is sub-gamma and sub-exponential with appropriate scale constants and variance process V_t proportional to t. Furthermore, any process that is sub-ψ with a CGF-like ψ function is also sub-gamma and sub-exponential with appropriate scaling of the variance process by a constant.

3.2. Conditions for sub-ψ processes

In Tables 3 and 4, we summarize a variety of standard and novel conditions for a process (S_t) to be sub-ψ. Fact 1 and Lemma 3 contain discrete-time results, while results for continuous time are in Fact 2. We let I_d denote the $d \times d$ identity matrix. For a process $(Y_t)_{t\in\mathcal{T}}$, $[Y]_t$ denotes the quadratic variation and $\langle Y\rangle_t$ the conditional quadratic variation; in discrete time, $[Y]_t := \sum_{i=1}^t \Delta Y_i^2$ and $\langle Y\rangle_t := \sum_{i=1}^t \mathbb{E}_{i-1}\Delta Y_i^2$. We extend a function $f : \mathbb{R} \to \mathbb{R}$ on the real line to an operator $f : \mathcal{H}^d \to \mathcal{H}^d$ on the space of Hermitian matrices in the standard way: if $A \in \mathcal{H}^d$ has the spectral decomposition $U\Lambda U^\star$ where Λ is diagonal with elements $\lambda_1, \ldots, \lambda_d$, then $f(A) = Uf(\Lambda)U^\star$ where $f(\Lambda)$ is diagonal with elements $f(\lambda_1), \ldots, f(\lambda_d)$. In particular, the absolute value function extends to $\mathcal{H}^d$ by taking absolute values of the eigenvalues, while $[Y_+]_t := \sum_{i=1}^t \max(0, \Delta Y_i)^2$ and $\langle Y_-\rangle_t := \sum_{i=1}^t \mathbb{E}_{i-1}\min(0, \Delta Y_i)^2$ operate by truncating the eigenvalues.

In the discrete-time case, we have the following known results.

TABLE 3

Summary of sufficient conditions for a real-valued, discrete- or continuous-time martingale (S_t) to be sub-ψ with the given variance process. We use the shorthand $\mu_t^k \coloneqq \mathbb{E}_{t-1}(\Delta S_t)^k$ and $|\mu|_t^k \coloneqq \mathbb{E}_{t-1}|\Delta S_t|^k$. In starred cases (*), the first moment $\mathbb{E}_{i-1}\Delta S_i$ need not exist, so (S_t) need not be a martingale. See Facts 1 and 2 and Lemma 3 for details of each case. "⇒ Hoeffding I" indicates that the variance process (V_t) for Hoeffding-KS is smaller. "SN" is short for "self-normalized".

	Condition	ψ	V_t
Discrete time, one-sided			
Bernoulli II	$\Delta S_t \le h, \mathbb{E}\Delta S_t^2 \le gh$	ψ_B	ght
Bennett	$\Delta S_t \le c$	ψ_P	$\langle S\rangle_t$
Bernstein	$\mu_t^k \le \frac{k!}{2}c^{k-2}\mu_t^2$	ψ_G	$\langle S\rangle_t$
*Heavy on left	$\mathbb{E}_{t-1}T_a(\Delta S_t) \le 0,\ \forall a > 0$	ψ_N	$[S]_t$
Bounded below	$\Delta S_t \ge -c$	ψ_E	$[S]_t$
Discrete time, two-sided			
Parametric	$\Delta S_t \overset{\text{iid}}{\sim} F$	$\log \mathbb{E}e^{\lambda \Delta S_1}$	t
Bernoulli I	$-g \le \Delta S_t \le h$	ψ_B	ght
Hoeffding-KS	$-g_t \le \Delta S_t \le h_t$	ψ_N	$\sum_{i=1}^t \varphi(g_i, h_i)$
⇒ Hoeffding I	$-g_t \le \Delta S_t \le h_t$	ψ_N	$\sum_{i=1}^t \left(\frac{g_i+h_i}{2}\right)^2$
*Symmetric	$\Delta S_t \sim -\Delta S_t \mid \mathcal{F}_{t-1}$	ψ_N	$[S]_t$
SN I	$\mathbb{E}_{t-1}\Delta S_t^2 < \infty$	ψ_N	$([S]_t + 2\langle S\rangle_t)/3$
SN II	$\mathbb{E}_{t-1}\Delta S_t^2 < \infty$	ψ_N	$([S_+]_t + \langle S_-\rangle_t)/2$
Cubic SN	$\mathbb{E}_{t-1}\lvert\Delta S_t\rvert^3 < \infty$	ψ_G	$[S]_t + \sum_{i=1}^t \lvert\mu\rvert_i^3$
Continuous time, one-sided			
Bennett	$\Delta S_t \le c$	ψ_P	$\langle S\rangle_t$
Bernstein	$W_{m,t} \le \frac{m!}{2}c^{m-2}V_t$	ψ_G	V_t
Continuous time, two-sided			
Lévy	$\mathbb{E}e^{\lambda S_1} < \infty$	$\log \mathbb{E}e^{\lambda S_1}$	t
Continuous paths	$\Delta S_t \equiv 0$	ψ_N	$\langle S\rangle_t$

TABLE 4

Summary from Fact 1 and Lemma 3 of sufficient conditions for an $\mathcal{H}^d$-valued, discrete-time martingale (Y_t) to have a sub-ψ maximum eigenvalue process $S_t = \gamma_{\max}(Y_t)$ with variance process $V_t = \gamma_{\max}(Z_t)$. We use the shorthand $\mu_t^k \coloneqq \mathbb{E}_{t-1}(\Delta S_t)^k$ and $|\mu|_t^k \coloneqq \mathbb{E}_{t-1}|\Delta S_t|^k$. In the symmetric* case, $\mathbb{E}_{i-1}\Delta Y_i$ need not exist, so (Y_t) need not be a martingale. "⇒ Hoeffding I" indicates that (V_t) for Hoeffding-KS is smaller. "SN" is short for "self-normalized".

	Condition	ψ	Z_t
Discrete time, one-sided			
Bernoulli II	$\Delta Y_t \preceq hI_d, \mathbb{E}\Delta Y_t^2 \preceq ghI_d$	ψ_B	$ghtI_d$
Bennett	$\Delta Y_t \preceq cI_d$	ψ_P	$\langle Y\rangle_t$
Bernstein	$\mu_t^k \preceq \frac{k!}{2}c^{k-2}\mu_t^2$	ψ_G	$\langle Y\rangle_t$
Bounded below	$\Delta Y_t \succeq -cI_d$	ψ_E	$[Y]_t$
Discrete time, two-sided			
Bernoulli I	$-gI_d \preceq \Delta Y_t \preceq hI_d$	ψ_B	$ghtI_d$
Hoeffding-KS	$-G_tI_d \preceq \Delta Y_t \preceq H_tI_d$	ψ_N	$\sum_{i=1}^t \varphi(G_i, H_i)I_d$
⇒ Hoeffding I	$-G_tI_d \preceq \Delta Y_t \preceq H_tI_d$	ψ_N	$\sum_{i=1}^t \left(\frac{G_i+H_i}{2}\right)^2 I_d$
Hoeffding II	$\Delta Y_t^2 \preceq A_t^2$	ψ_N	$\sum_{i=1}^t A_i^2$
*Symmetric	$\Delta Y_t \sim -\Delta Y_t \mid \mathcal{F}_{t-1}$	ψ_N	$[Y]_t$
SN I	$\mathbb{E}_{t-1}\Delta Y_t^2 < \infty$	ψ_N	$([Y]_t + 2\langle Y\rangle_t)/3$
SN II	$\mathbb{E}_{t-1}\Delta Y_t^2 < \infty$	ψ_N	$([Y_+]_t + \langle Y_-\rangle_t)/2$
Cubic SN	$\mathbb{E}_{t-1}\lvert\Delta Y_t\rvert^3 < \infty$	ψ_G	$[Y]_t + \sum_{i=1}^t \lvert\mu\rvert_i^3$

Fact 1. Let $(Y_t)_{t\in\mathbb{N}}$ be any $\mathcal{H}^d$-valued martingale, and let $S_t \coloneqq \gamma_{\max}(Y_t)$ for $t \in \mathbb{N}$. In all cases we set $l_0 = d$.

(a) (Scalar parametric) If $d = 1$ and S_t is a cumulative sum of i.i.d., real-valued random variables, each of which is mean zero with known CGF $\psi(\lambda)$ that is finite on $\lambda \in [0, \lambda_{\max})$, then (S_t) is sub-ψ with variance process $V_t = t$.

(b) (Bernoulli I) If $-gI_d \preceq \Delta Y_t \preceq hI_d$ a.s. for all $t \in \mathbb{N}$, then (S_t) is sub-Bernoulli with variance process $V_t = ght$ and range parameters g, h (Hoeffding, 1963; Tropp, 2012).

(c) (Bennett) If $\Delta Y_t \preceq cI_d$ a.s. for all $t \in \mathbb{N}$ for some $c > 0$, then (S_t) is sub-Poisson with variance process $V_t = \gamma_{\max}(\langle Y\rangle_t)$ and scale parameter c (Bennett, 1962; Hoeffding, 1963; Tropp, 2012).

(d) (Bernstein) If $\mathbb{E}_{t-1}(\Delta Y_t)^k \preceq (k!/2)c^{k-2}\mathbb{E}_{t-1}(\Delta Y_t)^2$ for all $t \in \mathbb{N}$ and $k = 2, 3, \ldots$, then (S_t) is sub-gamma with variance process $V_t = \gamma_{\max}(\langle Y\rangle_t)$ and scale parameter c (Bernstein, 1927; Tropp, 2012; Boucheron et al., 2013).

(e) (Heavy on left) Let $T_a(y) \coloneqq (y \wedge a) \vee -a$ for $a > 0$ denote the truncation of y. If $d = 1$ and

$$\mathbb{E}_{t-1}T_a(\Delta Y_t) \le 0 \quad \text{for all } a > 0, t \in \mathbb{N}, \tag{3.6}$$

then (S_t) is sub-Gaussian with variance process $V_t = \gamma_{\max}([Y]_t)$. A random variable satisfying (3.6) is called *heavy on left*, and (Y_t) need not be a martingale in this case (Bercu and Touati, 2008; Delyon, 2015; Bercu et al., 2015). For example, the centered versions of the exponential, gamma, Pareto, log-normal, Poisson ($\lambda \in \mathbb{N}$), Bernoulli ($p < 1/2$) and geometric ($0 < p < 1$) distributions are known to be heavy on left. When $-\Delta Y_t$ satisfies (3.6) we say ΔY_t is *heavy on right*.

In addition to the above known results, we provide the following extensions of known scalar results to matrices.

Lemma 3. *Let $(Y_t)_{t\in\mathbb{N}}$ be any $\mathcal{H}^d$-valued martingale, and let $S_t \coloneqq \gamma_{\max}(Y_t)$ for $t \in \mathbb{N}$. In all cases we set $l_0 = d$.*

(a) (Bernoulli II) If, for all $t \in \mathbb{N}$, $\Delta Y_t \preceq hI_d$ a.s. and $\mathbb{E}\Delta Y_t^2 \preceq ghI_d$, then (S_t) is sub-Bernoulli with variance process $V_t = ght$.

(b) (Hoeffding-KS) If $-G_tI_d \preceq \Delta Y_t \preceq H_tI_d$ a.s. for all $t \in \mathbb{N}$ for some real-valued, predictable sequences (G_t) and (H_t), then (S_t) is sub-Gaussian with variance process $V_t = \sum_{i=1}^t \varphi(G_i, H_i)$, where

$$\varphi(g, h) \coloneqq \begin{cases} \frac{h^2-g^2}{2\log(h/g)}, & g < h \\ gh, & g \ge h. \end{cases} \tag{3.7}$$

(c) (Hoeffding I) If $-G_tI_d \preceq \Delta Y_t \preceq H_tI_d$ a.s. for all $t \in \mathbb{N}$ for some real-valued, predictable sequences (G_t) and (H_t), then (S_t) is sub-Gaussian with variance process $V_t = \sum_{i=1}^t (G_i + H_i)^2/4$.

(d) *(Conditionally symmetric) If* $\Delta Y_t \sim -\Delta Y_t \mid \mathcal{F}_{t-1}$ *for all* $t \in \mathbb{N}$*, then* (S_t) *is sub-Gaussian with variance process* $V_t = \gamma_{\max}([Y]_t)$*. Here,* ΔY_t *need not be integrable, so* (Y_t) *need not be a martingale.*

(e) *(Bounded from below) If* $\Delta Y_t \succeq -cI_d$ *a.s. for all* $t \in \mathbb{N}$ *for some* $c > 0$*, then* (S_t) *is sub-exponential with variance process* $V_t = \gamma_{\max}([Y]_t)$ *and scale parameter* c*.*

(f) *(General self-normalized I) If* $\mathbb{E}_{t-1}\Delta Y_t^2$ *is finite for all* $t \in \mathbb{N}$*, then* (S_t) *is sub-Gaussian with variance process* $V_t = \gamma_{\max}([Y]_t + 2\langle Y\rangle_t)/3$*.*

(g) *(General self-normalized II) If* $\mathbb{E}_{t-1}\Delta Y_t^2$ *is finite for all* $t \in \mathbb{N}$*, then* (S_t) *is sub-Gaussian with variance process* $V_t = \gamma_{\max}([Y_+]_t + \langle Y_-\rangle_t)/2$*.*

(h) *(Hoeffding II) If* $\Delta Y_t^2 \preceq A_t^2$ *a.s. for all* $t \in \mathbb{N}$ *for some* $\mathcal{H}^d$*-valued predictable sequence* (A_t)*, then* (S_t) *is sub-Gaussian with variance process* $V_t = \gamma_{\max}(\sum_{i=1}^t A_i^2)$*.*

(i) *(Cubic self-normalized) If* $\mathbb{E}_{t-1}|\Delta Y_t|^3$ *is finite for all* $t \in \mathbb{N}$*, then* (S_t) *is sub-gamma with variance process* $V_t = \gamma_{\max}\left([Y]_t + \sum_{i=1}^t \mathbb{E}_{i-1}|\Delta Y_i|^3\right)$ *and scale parameter* $c = 1/6$*.*

The proof of the above lemma can be found in Section 6.5. Case (a) is a straightforward extension of Bennett's condition for upper-bounded random variables with bounded variance to matrices with upper-bounded eigenvalues and bounded matrix variance (Bennett, 1962, p. 42). Cases (b) and (c) are similar extensions of Hoeffding's sub-Gaussian conditions for bounded random variables to matrices with bounded eigenvalues (Hoeffding, 1963, Theorems 1 and 2; Kearns and Saul, 1998; Bercu et al., 2015, Theorem 2.49). In the conditionally symmetric case (d), we can achieve control without any moment or boundedness assumptions by defining V_t in terms of observed rather than expected squared deviations; this is known for $d = 1$ (de la Peña, 1999, Lemma 6.1; Bercu et al., 2015), allowing exponential concentration for distributions like Cauchy. In the lower-bounded increments case (e), we have a self-normalized complement to the Bennett-style bound, a result known for $d = 1$ (Fan et al., 2015, Lemma 4.1). For the square-integrable martingale cases (f, g), we achieve control for a broad class of processes by incorporating the conditional variance and the observed squared deviations, as known for $d = 1$ (Delyon, 2009, Theorem 4; Bercu et al., 2015). The Hoeffding-like case (h) follows from the self-normalized bounds, highlighting a connection implicit in the proof of Corollary 4.2 of Mackey et al. (2014). The third moment bound (i) is similar to a fixed-sample bound given by Fan et al. (2015, Corollary 2.2).

In the continuous-time, scalar case we have the following sufficient conditions for a local martingale (S_t) to be sub-ψ. Here we always assume (S_t) is càdlàg, $\Delta S_t \coloneqq S_t - S_{t-}$ denotes the jumps of S, $[S]_t$ denotes the quadratic variation, and $\langle S\rangle_t$ is the conditional quadratic variation, the compensator of $[S]_t$.

Fact 2. Here $\mathcal{T} = (0, \infty)$ and $d = 1$, and we set $l_0 = 1$.

(a) (Lévy process) If (S_t) is a Lévy process which is a martingale with the CGF $\psi(\lambda) = \log \mathbb{E}e^{\lambda S_1} < \infty$ for all $\lambda \in [0, \lambda_{\max})$, then (S_t) is sub-ψ with variance process $V_t = t$. See, e.g., Papapantoleon (2008, Proposition 10.2).

(b) (Continuous Bennett) If (S_t) is a local martingale with $\Delta S_t \leq c$ for all t a.s., then (S_t) is sub-Poisson with scale parameter c and variance process $V_t = \langle S \rangle_t$ (Lepingle, 1978, p. 157).

(c) (Continuous Bernstein) Suppose (S_t) is a locally square integrable martingale: let $W_{2,t} = \langle S \rangle_t$, and for $m = 3, 4, \ldots$ let $W_{m,t}$ be the compensator of the process $\sum_{u \leq t} |\Delta S_u|^m$. If, for some $c > 0$ and predictably measurable, càdlàg, nondecreasing process (V_t), it holds that $W_{m,t} \leq \frac{m!}{2} c^{m-2} V_t$ for all $m \geq 2$, then (S_t) is sub-gamma with scale parameter c and variance process V_t (van de Geer, 1995, implicit in the proof of Lemma 2.2).

(d) (Continuous paths) If (S_t) is a local martingale with a.s. continuous paths, then (S_t) is sub-Gaussian with variance process $V_t = \langle S \rangle_t$. This may be seen as a special case of (c), or a limiting case of (b).

3.3. Implications between sub-ψ conditions

In many settings, a process (S_t) may satisfy Definition 1 with several different choices of ψ and (V_t). Choosing a smaller ψ function will lead to tighter bounds in Theorem 1, but in some cases one may opt for a larger ψ function to achieve analytical or computational convenience. It is clear that making ψ uniformly larger retains the sub-ψ property, since the exponential process $\exp\{\lambda S_t - \psi(\lambda) V_t\}$ can only become smaller. It is therefore useful to characterize relationships among the above sub-ψ conditions, so that, after invoking one of the sufficient conditions given in Section 3.2, one may invoke Theorem 1 with a different, more convenient ψ function.

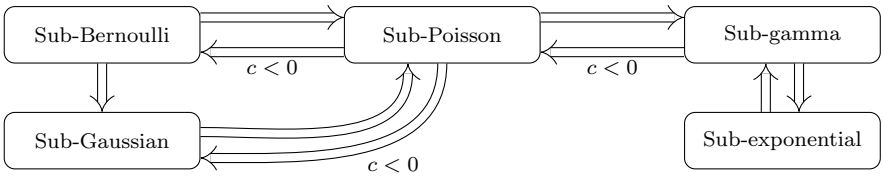

Fig 3: Each arrow indicates that any process satisfying the source sub-ψ condition, subject to a restriction on the scale parameter c, also satisfies the destination sub-ψ condition with appropriately scaled variance process. See Table 5 and Proposition 2 for details.

Note that ψ_G, ψ_P and ψ_E are nondecreasing in c for all values of $\lambda \geq 0$, so that if a process is sub-ψ with scale c for any of these ψ functions, then it is sub-ψ for any scale $c' > c$ as well. Similarly, ψ_B is nonincreasing in g and nondecreasing in h. Table 5 and Proposition 2 fully characterize all implications among sub-ψ conditions, as illustrated in Figure 3. These follow from inequalities of the form $\psi_1 \leq a\psi_2$, some of which are based on standard arguments; see Section 6.3.

Proposition 2. *For each row in Table 5, if (S_t) is sub-ψ_1 with variance process (V_t), and the given restrictions are satisfied, then (S_t) is also sub-ψ_2 with variance process (aV_t). Furthermore, when we allow only scaling of V_t by a constant,*

TABLE 5

If (S_t) is sub-ψ_1 with variance process (V_t), subject to the given restriction, then (S_t) is also sub-ψ_2 with variance process (aV_t). $\varphi(g,h)$ is defined in (3.7). See Proposition 2 for details.

	ψ_1	ψ_2	a	Restriction
(1)	$\psi_{B,g,h}$	ψ_N	$\frac{\varphi(g,h)}{gh}$	
(2)	$\psi_{B,g,h}$	ψ_N	$\frac{(g+h)^2}{4gh}$	
(3)	$\psi_{B,g,h}$	$\psi_{P,h-g}$	1	
(4)	ψ_N	$\psi_{P,0}$	1	
(5)	$\psi_{P,c}$	$\psi_{G,c/3}$	1	
(6)	$\psi_{G,c}$	$\psi_{E,3c/2}$	1	
(7)	$\psi_{E,c}$	$\psi_{G,c}$	1	$c \geq 0$
(8)	$\psi_{E,c}$	$\psi_{G,c/2}$	1	$c < 0$
(9)	$\psi_{G,c}$	$\psi_{P,2c}$	1	$c < 0$
(10)	$\psi_{P,c}$	ψ_N	1	$c < 0$
(11)	$\psi_{P,c}$	$\psi_{B,-c,h}$	1	$c < 0$, any $h > 0$

these capture all possible implications among the five sub-ψ conditions defined above, and the given constants are the best possible (in the case of row (2), the constant $(g+h)^2/4gh$ is the best possible of the form k/gh where k depends only on the total range $g+h$).

4. Applications of Theorem 1

Here, we illustrate how Theorem 1 recovers or strengthens a wide variety of existing results. Most results in this section follow immediately upon combining one of the sufficient conditions from Fact 1, Lemma 3, or Fact 2 with Theorem 1, and we omit proof details in many cases. As a rough plan, we first discuss classical Cramér-Chernoff and Freedman-style bounds and then Blackwell's line crossing inequalities. After discussing de la Peña-style self-normalized bounds and Pinelis' Banach-space inequalities, we end by exhibiting some continuous time results and mention connections to the sequential probability ratio test.

4.1. Fixed-time Cramér-Chernoff and Freedman-style uniform bounds

In the discrete-time, scalar setting, a simple sufficient condition for a process (S_t) to be 1-sub-ψ with variance process (V_t) is that

$$\mathbb{E}_{t-1} \exp\left\{\lambda \Delta S_t - \psi(\lambda)\Delta V_t\right\} \leq 1, \quad \forall t, \tag{4.1}$$

which is the standard assumption for a martingale-method Cramér-Chernoff inequality, typically with (V_t) predictable (McDiarmid, 1998; Chung and Lu, 2006; Boucheron et al., 2013). When (V_t) is deterministic, the fixed-time Cramér-Chernoff method gives, for fixed x and m,

$$\mathbb{P}(S_m \geq x) \leq \exp\left\{-V_m \psi^\star\left(\frac{x}{V_m}\right)\right\}, \tag{4.2}$$

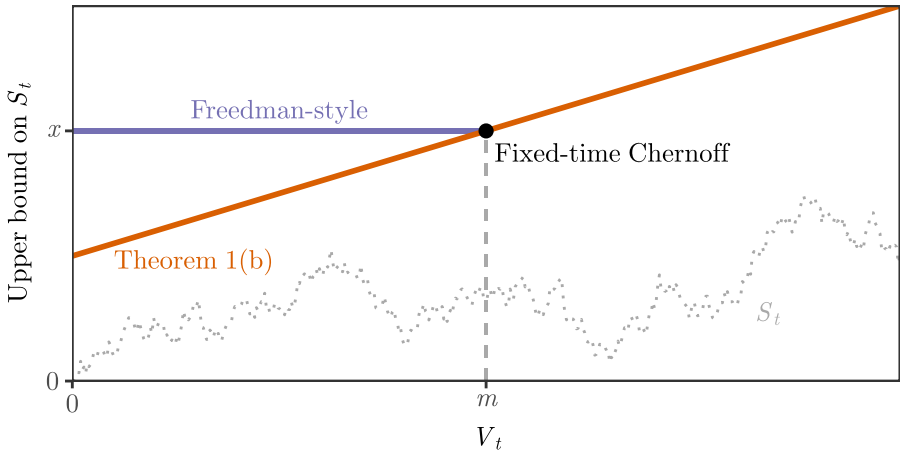

Fig 4: Comparison of (*i*) fixed-time Cramér-Chernoff bound (4.2), which bounds the deviations of S_m at a fixed time m; (*ii*) "Freedman-style" constant uniform bound (4.3), which bounds the deviations of S_t for all t such that $V_t \leq m$, with a constant boundary equal in value to the fixed-time Cramér-Chernoff bound; and (*iii*) linear uniform bound from Theorem 1(b), which bounds the deviations of S_t for all $t \in \mathbb{N}$, with a boundary growing linearly in V_t. Each bound gives the same tail probability and thus implies the preceding one.

so Theorem 1(b) is a uniform *extension* of the Cramér-Chernoff inequality, losing nothing at the fixed time m [B; C or D]. For random (V_t), a stopping time argument due to Freedman (1975) extends this to the uniform bound

$$\mathbb{P}(\exists t \in \mathcal{T} : S_t \geq x \text{ and } V_t \leq m) \leq \exp\left\{-m\psi^\star\left(\frac{x}{m}\right)\right\}. \tag{4.3}$$

When (V_t) is deterministic, analogous uniform bounds can be obtained from Doob's maximal inequality for submartingales, as in Hoeffding (1963, eq. 2.17). Theorem 1 strengthens this "Freedman-style" inequality [B; C or D], since it yields tighter bounds for all times t such that $V_t < m$, and also extends the inequality to hold for all times t with $V_t > m$, as illustrated by Figure 4.

Tropp (2011, 2012) extends the scalar Cramér-Chernoff approach to random matrices via control of the matrix moment-generating function, giving matrix analogues of Hoeffding's, Bennett's, Bernstein's and Freedman's inequalities. Following this approach, Theorem 1 gives corresponding strengthened versions of these inequalities for matrix-valued processes [B].

We summarize explicit results below for three well-known special cases reviewed in Example 1(a): Hoeffding's sub-Gaussian inequality for observations bounded from above and below, with variance process depending only on the radius of the interval of boundedness (Hoeffding, 1963); Bennett's sub-Poisson inequality for observations bounded from above, with variance process depending on the true variance of the observations (Bennett, 1962); and Bernstein's

sub-gamma inequality for observations satisfying a bound on growth of higher moments, also with a variance process depending on the true variance (Bernstein, 1927). In each case below, we recover the standard, fixed-sample result at $V_t = m$. Recall the definitions of $\mathfrak{s}_P, \psi_P^\star, \mathfrak{s}_G, \psi_G^\star$ from Table 2.

Corollary 1. *(a) Suppose $(Y_t)_{t\in\mathbb{N}}$ is an $\mathcal{H}^d$-valued martingale whose increments satisfy $\Delta Y_t^2 \preceq A_t^2$ a.s. for all t for some $\mathcal{H}^d$-valued, predictable sequence (A_t). Let $S_t \coloneqq \gamma_{\max}(Y_t)$, and let either*

$$V_t \coloneqq \frac{1}{2}\gamma_{\max}\left(\langle Y\rangle_t + \sum_{i=1}^t A_i^2\right) \quad \text{or} \quad V_t \coloneqq \gamma_{\max}\left(\sum_{i=1}^t A_i^2\right). \tag{4.4}$$

Then for any $x, m > 0$, we have

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t \geq x + \frac{x}{2m}(V_t - m)\right) \leq d\exp\left\{-\frac{x^2}{2m}\right\}. \tag{4.5}$$

This strengthens Hoeffding's inequality (Hoeffding, 1963) [A,B,D] and its matrix analogues in Tropp (2012, Theorem 7.1) [B,E] and Mackey et al. (2014, Corollary 4.2) [A,B].

(b) Suppose $(Y_t)_{t\in\mathbb{N}}$ is an $\mathcal{H}^d$-valued martingale satisfying $\gamma_{\max}(\Delta Y_t) \leq c$ a.s. for all t. Let $S_t \coloneqq \gamma_{\max}(Y_t)$ and $V_t \coloneqq \gamma_{\max}(\langle Y\rangle_t)$. Then for any $x, m > 0$, we have

$$\begin{aligned}\mathbb{P}\left(\exists t \in \mathbb{N} : S_t \geq x + \mathfrak{s}_P\left(\frac{x}{m}\right)\cdot(V_t - m)\right) &\leq d\exp\left\{-m\psi_P^\star\left(\frac{x}{m}\right)\right\} \\ &\leq d\exp\left\{-\frac{x^2}{2(m + cx/3)}\right\}. \end{aligned}\tag{4.6}$$

This strengthens Bennett's and Freedman's inequalities (Bennett, 1962; Freedman, 1975) [B; C or D] for scalars and the corresponding matrix bounds from Tropp (2011, 2012) [B].

(c) Suppose (S_t) is l_0-sub-gamma with variance process (V_t) and scale parameter c. Then for any $x, m > 0$, we have

$$\begin{aligned}\mathbb{P}\left(\exists t \in \mathcal{T} : S_t \geq x + \mathfrak{s}_G\left(\frac{x}{m}\right)\cdot(V_t - m)\right) &\leq l_0\exp\left\{-m\psi_G^\star\left(\frac{x}{m}\right)\right\} \\ &\leq l_0\exp\left\{-\frac{x^2}{2(m + cx)}\right\}. \end{aligned}\tag{4.7}$$

This strengthens Bernstein's inequality (Bernstein, 1927) [B; C or D], along with the matrix Bernstein inequality (Tropp, 2012) [B].

Case (a) is a consequence of Lemma 3(g); see also Corollary 8, which uses $V_t = \frac{1}{2}\gamma_{\max}([Y_+]_t + \langle Y_-\rangle_t)$. The first setting of V_t in case (a) follows from the bound $[Y_+]_t \preceq \sum_{i=1}^t A_i^2$, and further upper bounding $\langle Y_-\rangle_t \preceq \sum_{i=1}^t A_i^2$ yields the second setting of V_t. As is well known, the Hoeffding-style bound in part (a) and the Bennett-style bound in part (b) are not directly comparable: V_t may be

smaller in part (b), but $\psi_P^\star \leq \psi_N^\star$, so neither subsumes the other. We remark that $\psi_P^\star(u) \geq \frac{u}{2c}\operatorname{arcsinh}\left(\frac{cu}{2}\right)$, so the Bennett-style inequality in part (b) is an improvement on the inequality of Prokhorov (1959) for sums of independent random variables, as noted by Hoeffding (1963), as well as its extension to martingales in de la Peña (1999).

As an example of the Hermitian dilation technique for extending bounds on Hermitian matrices to bounds for rectangular matrices, we give a bound for rectangular matrix Gaussian and Rademacher series, following Tropp (2012); here $\|A\|_{\mathrm{op}}$ denotes the largest singular value of A. The proof is in Section 6.6.

Corollary 2. *Consider a sequence $(B_t)_{t\in\mathbb{N}}$ of fixed matrices with dimension $d_1 \times d_2$, and let $(\epsilon_t)_{t\in\mathbb{N}}$ be a sequence of independent standard normal or Rademacher variables. Let $S_t := \|\sum_{i=1}^t \epsilon_i B_i\|_{op}$ and*

$$V_t := \max\left\{\|\sum_{i=1}^t B_i B_i^\star\|_{op}, \|\sum_{i=1}^t B_i^\star B_i\|_{op}\right\}. \tag{4.8}$$

Then for any $x, m > 0$, we have

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t \geq x + \frac{x}{2m}(V_t - m)\right) \leq (d_1 + d_2)\exp\left\{-\frac{x^2}{2m}\right\}. \tag{4.9}$$

This strengthens Corollary 4.2 of Tropp (2012) [B].

4.2. Line-crossing inequalities

Before giving specific results in this section, we start with simplified versions of Theorem 1(d) which are useful for recovering existing results. The probability bound in (4.10) is merely an analytically simplified upper bound on that from Theorem 1(d). We prove the following in Section 6.7.

Corollary 3. *If (S_t) is l_0-sub-ψ with variance process (V_t) and ψ is CGF-like, then for any $m \geq 0$, $x > 0$ and $b \in (0, \bar{b})$, we have*

$$\mathbb{P}\left(\exists t \in \mathcal{T} : V_t \geq m \text{ and } S_t \geq x + b(V_t - m)\right) \leq l_0 \exp\left\{-m\psi^\star(b) - (x - bm)\psi^{\star\prime}(b)\right\}. \tag{4.10}$$

In particular, for $m > 0$, we have

$$\mathbb{P}\left(\exists t \in \mathcal{T} : V_t \geq m \text{ and } S_t \geq bV_t\right) \leq l_0 \exp\left\{-m\psi^\star(b)\right\}. \tag{4.11}$$

In fitting with the approach of this paper, Theorem 1(d) and Corollary 3 bound the upcrossing probability on $\{V_t \geq m\}$ using the results of Theorem 1(a,b) and a geometric argument. It may seem naive and wasteful to bound a line-crossing probability on $\{V_t \geq m\}$ using a bound which applies for $\{V_t > 0\}$. The literature includes a handful of results bounding line-crossing probabilities

on $\{V_t \geq m\}$ which appear to give bounds tighter than what Theorem 1 offers, by making more direct use of the intrinsic-time condition (Blackwell, 1997; Khan, 2009). Below we demonstrate that this is not true: we give several special cases of Theorem 1(d) and Corollary 3 which improve upon existing results.

Corollary 4. *Suppose (S_t) is l_0-sub-gamma with variance process (V_t) and scale parameter c.*

(a) For any $a, b > 0$, we have

$$\mathbb{P}\left(\exists t \in \mathcal{T} : S_t \geq a + bV_t\right) \leq l_0 \exp\left\{-\frac{2ab}{1+2cb}\right\}. \tag{4.12}$$

When $\mathcal{T} = \mathbb{N}$, $c = 0$ and $d = 1$ this strengthens Theorem 1 of Blackwell (1997) [A; C or D], which is written for discrete-time scalar processes with bounded increments.

(b) For any $m, b > 0$, we have

$$\begin{aligned}\mathbb{P}\left(\exists t \in \mathcal{T} : V_t \geq m \text{ and } S_t \geq bV_t\right) &\leq l_0 \exp\left\{-m\psi_G^\star(b)\right\} \\ &\leq l_0 \exp\left\{-\frac{b^2 m}{2(1+cb)}\right\}. \end{aligned} \tag{4.13}$$

When $\mathcal{T} = \mathbb{N}$, $c = 0$ and $d = 1$ this strengthens the second bound in Theorem 2 of Blackwell (1997) [A; C or D], which is written for discrete-time scalar processes with bounded increments.

In discrete time, as presented in Fact 1, for a process with bounded increments we may construct both sub-Bernoulli and sub-Gaussian bounds. The sub-Bernoulli case, in combination with (4.11), yields the following:

Corollary 5. *Suppose $(Y_t)_{t\in\mathbb{N}}$ is an $\mathcal{H}^d$-valued martingale satisfying $\|\Delta Y_t\|_{op} \leq 1$ a.s. for all $t \in \mathbb{N}$. Then for any $b \in [0,1]$ and $m \geq 1$, we have*

$$\mathbb{P}\left(\exists t \in \mathbb{N} : t \geq m \text{ and } \gamma_{\max}(Y_t) \geq bt\right) \leq \left[(1+b)^{(1+b)}(1-b)^{(1-b)}\right]^{-m/2}. \tag{4.14}$$

This strengthens the first bound in Theorem 2 of Blackwell (1997) [D].

Theorems 4.1-4.3 of Khan (2009) are closest in form to our main results and represent key precedents to our framework. The simplified bound (4.10) recovers Khan's Theorem 4.3 [C or D], while Theorem 1(d) improves the exponent [E]. Our Theorem 1(b) gives a strengthened version of Khan's "Freedman-style" Theorem 4.2 [B; C or D]. Khan's Theorem 4.1 is not strictly comparable to our work since it involves an initial condition on *nominal* time, $t \geq t_0$, rather than on intrinsic time, $V_t \geq m$, but when V_t is deterministic, then our Theorem 1(d) is tighter [B; C or D; E].

4.3. Self-normalized uniform bounds

Collectively, de la Peña (1999); de la Peña et al. (2000, 2004, 2007); de la Peña, Klass and Lai (2009); and de la Peña, Lai and Shao (2009) give a wide variety

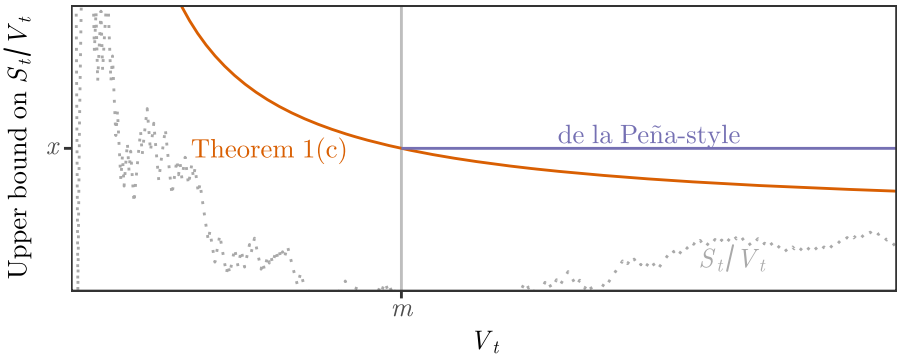

Fig 5: Comparison of our decreasing boundary from Theorem 1(c), as in inequality (4.16), to a "de la Peña-style" constant uniform bound as in inequality (4.15), which bounds the deviations of S_t/V_t for all t such that $V_t \geq m$ with a constant boundary.

of sufficient conditions for the exponential process $\exp\{\lambda S_t - \psi(\lambda)V_t\}$ to be a supermartingale in both discrete- and continuous-time settings. They formulate their bounds for ratios involving S_t in the numerator and V_t in the denominator, as in Theorem 1(c), and often specify initial-time conditions, as in Theorem 1(d). In this section we draw some comparisons between Theorem 1 and their results. As a first example, consider the boundary of Theorem 1(c) for the ratio S_t/V_t, strictly decreasing towards the asymptotic level $\mathfrak{s}(x)$. In particular, at time $V_t = m$ the boundary equals x, so Theorem 1(c) strengthens various theorems of de la Peña (1999) and de la Peña et al. (2007) which use a constant boundary after time $V_t = m$ [B; C or D]; for example, Theorem 1.2B, eq. 1.5 of de la Peña (1999) states that

$$\mathbb{P}\left(\exists t \geq 1 : V_t \geq m \text{ and } \frac{S_t}{V_t} \geq x\right) \leq \exp\{-m\psi_G^\star(x)\} \tag{4.15}$$

for scalar processes (S_t) which are 1-sub-gamma with variance process (V_t). As before, we give explicit results for special cases.

Corollary 6. *Suppose (S_t) is l_0-sub-gamma with variance process (V_t) and scale parameter c. Then for any $x, m > 0$, we have*

$$\begin{aligned}\mathbb{P}\left(\exists t \in \mathcal{T} : \frac{S_t}{V_t} \geq \mathfrak{s}_G(x)\left(1 + \frac{m\sqrt{1+2cx}}{V_t}\right)\right) &\leq l_0 \exp\{-m\psi_G^\star(x)\} \\ &\leq l_0 \exp\left\{-\frac{mx^2}{2(1+cx)}\right\}. \end{aligned} \tag{4.16}$$

This strengthens eq. 1.5 from Theorem 1.2B of de la Peña (1999) [B; C or D].

In the sub-Gaussian case (obtained at $c = 0$), the above bound simplifies to

$$\mathbb{P}\left(\exists t \in \mathcal{T} : \frac{S_t}{V_t + m} \geq x\right) \leq l_0 \exp\left\{-2mx^2\right\}. \tag{4.17}$$

This strengthens Theorem 2.1 of de la Peña et al. (2007) and Theorem 6.1 of de la Peña (1999) [B, C or D].

Recall that $\mathfrak{s}_G(x) = x/(1 + \sqrt{1 + 2cx})$, so for the boundary in (4.16), we have $\mathfrak{s}_G(x)(1 + m\sqrt{1 + 2cx}/V_t) \leq x$ for all $V_t \geq m$ with equality at $V_t = m$. Corollary 6(a), therefore, gives the same probability bound as (4.15) for a larger crossing event. Figure 5 visualizes this relationship.

More generally, when we normalize by $\alpha + \beta V_t$ and include an initial time condition $V_t \geq m$, Theorem 1(d) and Corollary 3 become the following:

Corollary 7. *If (S_t) is l_0-sub-ψ with variance process (V_t), where ψ is CGF-like and $\bar{b} = \infty$, then for any $\beta, x > 0$ and $\alpha, m \geq 0$ with at least one of $\alpha, m > 0$, we have*

$$\mathbb{P}\left(\exists t \in \mathcal{T} : V_t \geq m \text{ and } \frac{S_t}{\alpha + \beta V_t} \geq x\right) \tag{4.18}$$

$$\leq \begin{cases} l_0 \exp\left\{-\alpha x D(\beta x)\right\}, & \beta x \leq \mathfrak{s}\left(\frac{x(\alpha+\beta m)}{m}\right) \\ l_0 \exp\left\{-m\psi^\star\left(\frac{x(\alpha+\beta m)}{m}\right)\right\}, & \beta x \geq \mathfrak{s}\left(\frac{x(\alpha+\beta m)}{m}\right) \end{cases} \tag{4.19}$$

$$\leq l_0 \exp\left\{-m\psi^\star(\beta x) - \alpha x \psi^{\star\prime}(\beta x)\right\}. \tag{4.20}$$

In the case (S_t) is sub-Gaussian, for any $\beta, x > 0$ and $\alpha, m \geq 0$ with at least one of $\alpha, m > 0$, we have

$$\mathbb{P}\left(\exists t \in \mathcal{T} : V_t \geq m \text{ and } \frac{S_t}{\alpha + \beta V_t} \geq x\right)$$

$$\leq \exp\left\{-x^2\left(2\alpha\beta + \frac{(\beta m - \alpha)^2 1_{\alpha \leq \beta m}}{2m}\right)\right\}, \tag{4.21}$$

taking $0/0 = 0$ on the right-hand side when $m = 0$. With Lemma 3(d), this improves eq. 6.4 from Theorem 6.2 of de la Peña (1999) [C or D; E].

A defining feature of self-normalized bounds is that they involve a variance process (V_t) constructed with the squared observations themselves rather than just conditional variances or constants. Such normalization can be found in common statistical procedures such as the t-test. Furthermore, it allows for Gaussian-like concentration while reducing or eliminating moment conditions. Lemma 3 gives several extensions of well-known conditions for scalar sub-Gaussian concentration of self-normalized processes. As one particular special case, Lemma 3(f) and (g) yield general self-normalized uniform bounds for any discrete-time, square-integrable, $\mathcal{H}^d$-valued martingale, building upon breakthrough results obtained for scalar processes by Bercu, Touati and Delyon:

Corollary 8. *Suppose $(Y_t)_{t\in\mathbb{N}}$ is an $\mathcal{H}^d$-valued martingale with $\mathbb{E}Y_t^2 < \infty$ for all $t \in \mathbb{N}$. Let $S_t \coloneqq \gamma_{\max}(Y_t)$ and either*

$$V_t \coloneqq \frac{1}{2}\gamma_{\max}([Y_+]_t + \langle Y_- \rangle_t) \quad \text{or} \quad V_t \coloneqq \frac{1}{3}\gamma_{\max}([Y]_t + 2\langle Y \rangle_t). \tag{4.22}$$

Then for any $x, m > 0$, we have

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \frac{S_t}{V_t + m} \geq x\right) \leq d \exp\left\{-2mx^2\right\}. \tag{4.23}$$

This strengthens eq. 20 from Theorem 4 of Delyon (2009) [B,D], Theorem 2.1 of Bercu and Touati (2008) [B,D,E], and an implicit self-normalized bound of Mackey et al. (2014, Corollary 4.2) [B].

Corollary 8 is remarkable for the fact that it gives Gaussian-like concentration with only the existence of second moments for the increments. If the increments have conditionally symmetric distributions, one may instead apply Lemma 3(d) to achieve Gaussian-like concentration without existence of any moments, as discovered by de la Peña (1999) and illustrated in the following example.

Example 5 (Cauchy increments)**.** Let $(\Delta S_t)_{t\in\mathbb{N}}$ be i.i.d. standard Cauchy random variables (symmetric about zero). Lemma 3(d) shows that (S_t) is sub-Gaussian with variance process $V_t = [S]_t$. Corollary 6 yields, for any $x, m > 0$,

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \frac{S_t}{[S]_t + m} \geq x\right) \leq \exp\left\{-2mx^2\right\}. \tag{4.24}$$

For another example, Lemma 3(i) gives a self-normalized bound involving third rather than second moments:

Corollary 9. *Suppose $(Y_t)_{t\in\mathbb{N}}$ is an $\mathcal{H}^d$-valued martingale with $\mathbb{E}\left|Y_t\right|^3$ finite for all $t \in \mathbb{N}$. Let $S_t \coloneqq \gamma_{\max}(Y_t)$ and $V_t \coloneqq \gamma_{\max}([Y]_t + \sum_{i=1}^t \mathbb{E}_{i-1}(\Delta Y_i)_-^3)$. Then for $\mathfrak{s}_G$ and $\psi_G^\star$ using $c = 1/6$, we have for any $x, m > 0$,*

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t \geq x + \mathfrak{s}_G\left(\frac{x}{m}\right) \cdot (V_t - m)\right) \leq d \exp\left\{-m\psi_G^\star\left(\frac{x}{m}\right)\right\} \tag{4.25}$$

$$\leq d \exp\left\{-\frac{x^2}{2(m + x/6)}\right\}. \tag{4.26}$$

This is a uniform alternative to Corollary 2.2 of Fan et al. (2015) [B,D].

Note the exponent in (4.26) is different from that in Fan et al. (2015); neither strictly dominates the other. Also note that, unlike the classical Bernstein bound, neither of Corollaries 8 and 9 assume existence of moments of all orders.

4.4. Martingales in smooth Banach spaces

The applications presented thus far allow us to uniformly bound the operator norm deviations of a sequence of random Hermitian matrices. A different

approach is due to Pinelis (1992, 1994), who gave an innovative approach to exponential tail bounds in abstract Banach spaces. We describe how this approach can be incorporated into our framework. For this section, let $(Y_t)_{t\in\mathbb{N}}$ be a martingale with respect to $(\mathcal{F}_t)$ taking values in a separable Banach space $(\mathcal{X}, \|\cdot\|)$. We can use Pinelis's device to uniformly bound the process $(\Psi(Y_t))$ for any function $\Psi : \mathcal{X} \to \mathbb{R}$ which satisfies the following smoothness property:

Definition 3 (Pinelis, 1994)**.** A function $\Psi : \mathcal{X} \to \mathbb{R}$ is called $(2, D)$-*smooth* for some $D > 0$ if, for all $x, v \in \mathcal{X}$, we have

$$\Psi(0) = 0 \tag{4.27a}$$

$$|\Psi(x + v) - \Psi(x)| \le \|v\| \tag{4.27b}$$

$$\Psi^2(x + v) - 2\Psi^2(x) + \Psi^2(x - v) \le 2D^2\|v\|^2. \tag{4.27c}$$

A Banach space is called $(2, D)$-smooth if its norm is $(2, D)$-smooth; in such a space we may take $\Psi(\cdot) = \|\cdot\|$ to uniformly bound the deviations of a martingale. In this case, observe that property (4.27a) is part of the definition of a norm, property (4.27b) is the triangle inequality, and property (4.27c) can be seen to hold with $D = 1$ for the norm induced by the inner product in any Hilbert space, regardless of the (possibly infinite) dimensionality of the space. Note also that setting $x = 0$ shows that $D \ge 1$ whenever $\Psi(\cdot) = \|\cdot\|$. Finally, observe that if we write $f(x) = \Psi^2(x)$, then we may equivalently replace condition (4.27c) by $f(tx + (1-t)y) \ge tf(x) + (1-t)f(y) - D^2t(1-t)\|x - y\|^2$, a perhaps more familiar definition of smoothness.

Corollary 10. *Consider a martingale $(Y_t)_{t\in\mathbb{N}}$ taking values in a separable Banach space $(\mathcal{X}, \|\cdot\|)$. Let the function $\Psi : \mathcal{X} \to \mathbb{R}$ be $(2, D)$-smooth and define $D_\star \coloneqq 1 \vee D$.*

(a) Suppose $\|\Delta Y_t\| \le c_t$ a.s. for all $t \in \mathbb{N}$ for some constants $(c_t)_{t\in\mathbb{N}}$, and let $V_t \coloneqq \sum_{i=1}^t c_i^2$. Then for any $x, m > 0$, we have

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \Psi(Y_t) \ge x + \frac{D_\star^2 x}{2m}(V_t - m)\right) \le 2\exp\left\{-\frac{x^2}{2D_\star^2 m}\right\}. \tag{4.28}$$

This strengthens Theorem 3.5 from Pinelis (1994) [B].

(b) Suppose $\|\Delta Y_t\| \le c$ a.s. for all $t \in \mathbb{N}$ for some constant c, and let $V_t \coloneqq \sum_{i=1}^t \mathbb{E}_{i-1}\|\Delta Y_i\|^2$. Then for any $x, m > 0$, we have

$$\begin{aligned}&\mathbb{P}\left(\exists t \in \mathbb{N} : \Psi(Y_t) \ge x + D_\star^2 \mathfrak{s}_P\left(\frac{x}{m}\right) \cdot (V_t - m)\right) \\ &\quad\le 2\exp\left\{-D_\star^2 m \psi_P^\star\left(\frac{x}{D_\star^2 m}\right)\right\} \le 2\exp\left\{-\frac{x^2}{2(D_\star^2 m + cx/3)}\right\}.\end{aligned} \tag{4.29}$$

This strengthens Theorem 3.4 from Pinelis (1994) [B].

We prove this result in Section 6.8. As before, the Hoeffding-style bound in part (a) and the Bennett-style bound in part (b) are not directly comparable: V_t may be smaller in part (b), but the exponent is also smaller.

We briefly highlight some of the strengths and limitations of this approach. Since the Euclidean l_2-norm is induced by the standard inner product in $\mathbb{R}^d$, Corollary 10 gives a dimension-free uniform bound on the l_2-norm deviations of a vector-valued martingale in $\mathbb{R}^d$ which exactly matches the form for scalars. Compare this to bounds based on the operator norm of a Hermitian dilation: the bound of Tropp (2012) includes dimension dependence [B,E] while the bound of Minsker (2017, Corollary 4.1) incurs an extra constant factor of 14 [B,E]. Our bounds extend to martingales taking values in sequence space $\left\{(a_i)_{i\in\mathbb{N}} : \sum_i |a_i|^2 < \infty\right\}$ or function space $L^2[0,1]$, and we may instead use the l_p norm, $p \geq 2$, in which case $D = \sqrt{p-1}$. These cases follow from Pinelis (1994, Proposition 2.1).

Similarly, Corollary 10 gives dimension-free uniform bounds for the Frobenius-norm deviations of a matrix-valued martingale. This extends to martingales taking values in a space of Hilbert-Schmidt operators on a separable Hilbert space, with deviations bounded in the Hilbert-Schmidt norm; compare Minsker (2017, §3.2), which gives operator-norm bounds. The method of Corollary 10 does not extend directly to operator-norm bounds because the operator norm is not $(2, D)$-smooth for any D: for a simple illustration in $\mathcal{H}^2$, consider $x = aI_2$ and $v = \operatorname{diag}\{b, -b\}$, so that $\|x+v\|_{\text{op}}^2 + \|x-v\|_{\text{op}}^2 - 2\|x\|_{\text{op}}^2 = 2b^2 + 4ab$ and condition (4.27c) cannot be satisfied. However, Corollary 10 does apply to the matrix Schatten p-norm for $p < \infty$, using $D = \sqrt{p-1}$, and this holds for rectangular matrices as well (Ball et al., 1994).

4.5. Continuous-time processes

While Corollaries 1, 4, 6, and 7 already generalize results known in discrete time to new results for continuous-time martingales [C], here we summarize a few more useful bounds explicitly for continuous-time processes which follow from Theorem 1 and the conditions of Fact 2, making use of the novel strategies devised by Shorack and Wellner (1986) and van de Geer (1995). These results use the conditional quadratic variation $\langle S\rangle_t$. We remind the reader that $[S]_t = \langle S\rangle_t = t$ for Brownian motion, and the first equality holds more generally for martingales with continuous paths, while for a Poisson process with rate one, $\langle S\rangle_t = t$ but $[S]_t = S_t$.

Corollary 11. *Let $(S_t)_{t\in(0,\infty)}$ be a real-valued process.*

(a) If (S_t) is a locally square-integrable martingale with a.s. continuous paths, then for any $a, b > 0$, we have

$$\mathbb{P}\left(\exists t \in (0,\infty) : S_t \geq a + b\langle S\rangle_t\right) \leq e^{-2ab}. \tag{4.30}$$

If $\langle S\rangle_t \uparrow \infty$ as $t \uparrow \infty$, then the probability upper bound holds with equality. This recovers as a special case the standard line-crossing probability for Brownian motion (e.g., Durrett, 2017, Exercise 7.5.2).

(b) *If (S_t) is a local martingale with $\Delta S_t \le c$ for all t, then for any $x, m > 0$, we have*

$$\mathbb{P}\left(\exists t \in (0,\infty) : S_t \ge x + \mathfrak{s}_P\left(\frac{x}{m}\right) \cdot (\langle S \rangle_t - m)\right) \le \exp\left\{-m\psi_P^\star\left(\frac{x}{m}\right)\right\} \le \exp\left\{-\frac{x^2}{2(m + cx/3)}\right\}. \quad (4.31)$$

This strengthens Appendix B, Inequality 1 of Shorack and Wellner (1986) [B].

(c) *If (S_t) is any locally square-integrable martingale satisfying the Bernstein condition of Fact 2(c) for some predictable process (V_t), then for any $x, m > 0$, we have*

$$\mathbb{P}\left(\exists t \in (0,\infty) : S_t \ge x + \mathfrak{s}_G\left(\frac{x}{m}\right) \cdot (V_t - m)\right) \le \exp\left\{-m\psi_G^\star\left(\frac{x}{m}\right)\right\} \le \exp\left\{-\frac{x^2}{2(m + cx)}\right\}. \quad (4.32)$$

This strengthens Lemma 2.2 of van de Geer (1995) [B,E].

Clearly, Corollary 11(b) applies to centered Poisson processes with $c = 1$. Of course, one can also apply Fact 2(a) for general Lévy processes, obtaining the same bound (4.31). The point of Corollary 11(b) is that any local martingale with bounded jumps obeys this inequality, and so concentrates like a centered Poisson process in this sense. Barlow et al. (1986, §4) describe further exponential supermartingales obtained for continuous-time processes using the quadratic variation, and derive "Freedman-style" self-normalized bounds; incorporating these cases into our framework would be interesting future work.

4.6. Exponential families and the sequential probability ratio test

It is well known that the likelihood ratio $f_{1,t}(X_1^t)/f_{0,t}(X_1^t)$ is a martingale under the null hypothesis that $X_1^t \sim f_{0,t}$. Then Ville's inequality gives a sequential test with valid type I error, equivalent to an open-ended sequential probability ratio test (SPRT, Wald, 1945), in which we stop when the likelihood ratio exceeds an upper threshold, but not when it drops below any lower threshold. In the one-parameter exponential family case, we obtain a simple analytical result which is equivalent to Theorem 1, as we detail below.

Suppose $(X_t)_{t\in\mathbb{N}}$ are i.i.d. from a one-parameter exponential family with natural parameter θ and log-partition function A, so that X_t has density $f_\theta(x) = h(x)\exp\{\theta T(x) - A(\theta)\}$. Let $S_t = \sum_{i=1}^t T(X_i)$. An open-ended SPRT testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_0 + \lambda$ stops to reject H_0 as soon as the likelihood ratio $L_t = \exp\{\lambda S_t - [A(\theta_0 + \lambda) - A(\theta_0)]t\}$ exceeds the threshold $\alpha^{-1} > 1$.

Corollary 12. *This one-sided SPRT has type I error rate no greater than α: $\mathbb{P}_{\theta_0}(\exists t \in \mathbb{N} : L_t \ge \alpha^{-1}) \le \alpha$.*

This standard fact follows easily from Theorem 1 because $L_t \geq A$ if and only if $S_t \geq (\log A)/\lambda + \psi(\lambda)t/\lambda$, where $\psi(\lambda) = A(\theta_0 + \lambda) - A(\theta_0)$, the CGF of $T(X_i)$ at $\theta = \theta_0$. Hence the rejection boundary for the SPRT is equivalent to the linear boundary of Theorem 1. In light of this, we may interpret the above sub-Gaussian, sub-Poisson, sub-exponential and sub-Bernoulli bounds as open-ended SPRTs for i.i.d. observations from these exponential families. The fact that such tests are also valid for testing various nonparametric classes of distributions, as outlined in Section 3, illustrates how our framework provides nonparametric generalizations of the SPRT. For example, if one wants to test the mean of a bounded distribution, our framework suggests that one apply an SPRT for Bernoulli or Poisson observations, for example. It has long been known that the normal SPRT bound can be applied to sequential problems involving any i.i.d. sequence of sub-Gaussian observations (Darling and Robbins, 1967; Robbins, 1970). Our work expands the breadth of nonparametric sequential problems amenable to such methods and deepens the connection between exponential concentration inequalities and sequential testing procedures.

5. Discussion and extensions

This section is divided into three parts. We first discuss the sharpness of the derived bounds. Then, building further on the geometric intuition of the paper, we point out an interesting geometric relationship between fixed-sample exponential bounds and our uniform bounds. We end by discussing directions for future work.

5.1. When is Theorem 1 sharp?

In the discrete-time, sub-Gaussian case $\psi = \psi_N$ and $l_0 = 1$, Theorem 1(a) is sharp: for any $a, b > 0$,

$$\sup_{(S_t, V_t) \in \mathbb{S}^1_{\psi_N}} \mathbb{P}\left(\exists t \in \mathbb{N} : S_t \geq a + bV_t \mid \mathcal{F}_0\right) = e^{-2ab}. \tag{5.1}$$

In fact, this can be achieved by rescaling any sum of i.i.d. observations with finite variance, which we prove in Section 6.9 as a corollary of Theorem 2 of Robbins and Siegmund (1970):

Corollary 13. *Suppose $(X_t)_{t\in\mathbb{N}}$ are i.i.d. mean zero with variance $\sigma^2 < \infty$. Let $S_t = \sum_{i=1}^t X_i$. Let $S_t^{(m)} := S_t/\sqrt{m}$ and $V_t^{(m)} := t\sigma^2/m$. Then for any $a, b > 0$,*

$$\lim_{m\to\infty} \mathbb{P}\left(\exists t \in \mathbb{N} : S_t^{(m)} \geq a + bV_t^{(m)}\right) = e^{-2ab}. \tag{5.2}$$

The following more general sandwich relation, which we prove in Section 6.10, quantifies the looseness in Theorem 1(a) and gives a sufficient condition for the probability bound to be exact. This condition involves the "overshoot" of the process S_t over the line $a+bV_t$, a quantity which has been studied extensively in

the context of sequential testing (Siegmund, 1985). The upper bound in equation (5.3) below is a restatement of Theorem 1(a); only the lower bound is new.

Proposition 3. *Consider real-valued processes* (S_t), (V_t) *and a CGF-like function* ψ. *Fix* $a > 0, b \in (0, \bar{b})$ *and suppose*

1. $M_t := \exp\{D(b)S_t - \psi(D(b))V_t\}$ *is a martingale with* $M_0 \equiv 1$ *(rather than just upper bounded by a supermartingale, as Definition 1 requires),*
2. $S_t - bV_t \to -\infty$ *as* $t \uparrow \infty$ *a.s., and*
3. *For some* $\epsilon \geq 0$, $S_\tau \leq a + bV_\tau + \epsilon$ *a.s. on* $\{\tau < \infty\}$, *where* $\tau := \inf\{t \in \mathcal{T} : S_t \geq a + bV_t\}$.

Then we have

$$e^{-\epsilon D(b)} \leq \frac{\mathbb{P}\left(\exists t \in \mathcal{T} : S_t \geq a + bV_t\right)}{\exp\{-aD(b)\}} \leq 1. \tag{5.3}$$

In particular, if the conditions of Proposition 3 hold with $\epsilon = 0$, then the probability bounds in Theorem 1 parts (a), (b) and (c) hold with equality. In the continuous-time case with (S_t) a continuous martingale, these conditions often hold with $\psi = \psi_N$ and $V_t = [S]_t$. We give details for the following result in Section 6.11; see Protter (2005, Theorem III.44) for more on Kazamaki's criterion:

Corollary 14. *Suppose* $(S_t)_{t\in(0,\infty)}$ *is a continuous martingale with* $S_0 = 0$ *and* $[S]_t \uparrow \infty$ *a.s. satisfying Kazamaki's criterion:* $\sup_T \mathbb{E}e^{S_T/2} < \infty$, *where the supremum is taken over all bounded stopping times* T. *Then* $\mathbb{P}(\exists t \in (0,\infty) : S_t \geq a + bV_t) = e^{-2ab}$.

In the discrete-time case with i.i.d. observations bounded above by ϵ a.s. and having CGF ψ, the conditions of Proposition 3 hold, setting $V_t = t$. Hence the probability bound in Theorem 1(a) can be made arbitrarily close to exact by taking b sufficiently small relative to ϵ, and similarly for parts (b) and (c). So Theorem 1 is sharp in the sense that for any such process, the probability bound is arbitrarily close to exact for some choice of (a, b) or (x, m). To see the connection with Corollary 13, rewrite (5.2) to keep the processes S_t and $V_t = t\sigma^2$ fixed and take limits with respect to a, b:

$$\lim_{m\to\infty} \mathbb{P}\left(\exists t \in \mathbb{N} : S_t \geq a\sqrt{m} + \frac{b}{\sqrt{m}} \cdot t\sigma^2\right) = e^{-2ab}. \tag{5.4}$$

5.2. *Geometric relationship between Theorem 1 and Cramér-Chernoff*

Whenever a process (S_t) is sub-ψ with $V_t = t$, a fixed-time Cramér-Chernoff upper bound of the form (4.2) holds: for any fixed $t \in \mathbb{N}$, we have $\mathbb{P}(S_t \geq x) \leq \exp\{-t\psi^\star(x/t)\}$. Let $f_\alpha(t)$ denote the curve of such fixed-time bounds constructed for a fixed crossing probability α at each time t:

$$f_\alpha(t) := t\psi^{\star -1}\left(\frac{\log \alpha^{-1}}{t}\right), \tag{5.5}$$

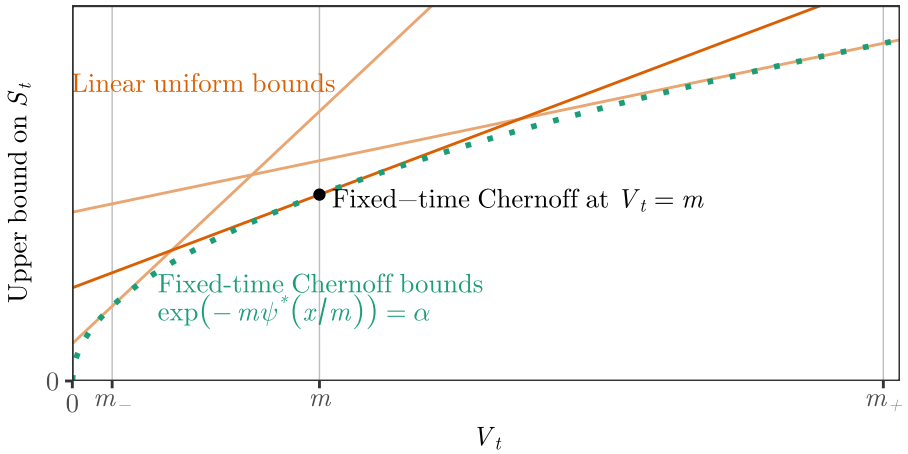

Fig 6: Geometric illustration of Theorem 1(b) and its relation to fixed-time Cramér-Chernoff bounds. Theorem 1(b) chooses the linear boundary which is optimal for $V_t = m$, but other linear boundaries with the same crossing probability are illustrated, each of which achieves the optimal fixed-time bound at some other time $V_t = m_\pm$. Each uniform Chernoff bound is tangent to the curve of fixed-time bounds, and indeed the curve of fixed-time bounds may be defined as the pointwise infimum of such linear uniform bounds.

where $\psi^{\star -1}(\lambda) = \inf\{u \geq 0 : \psi^\star(u) > \lambda\}$. For example, in the sub-Gaussian case $\psi(\lambda) = \psi_N(\lambda) = \lambda^2/2$, we have the standard formula $f_\alpha(t) = \sqrt{2t \log \alpha^{-1}}$.

Proposition 4. *Any line $a + bt$ which is tangent to $f_\alpha(t)$ satisfies $\mathbb{P}(\exists t \in \mathcal{T} : S_t \geq a + bt) \leq \alpha$.*

In words, the above proposition states that the set of linear boundaries from Theorem 1 is exactly the set of tangent lines to f_α, or conversely, f_α is defined as the pointwise infimum of this set of linear boundaries, as illustrated in Figure 6. We give the proof in Section 6.12. This observation provides some intuition for the appearance of the Legendre-Fenchel transform in the standard Cramér-Chernoff formula (4.2).

5.3. Future work

Characterizing families of sub-ψ processes Our Theorem 1 bounds the maximal line-crossing probability over each family $\mathbb{S}_\psi^{t_0}$, and Section 3 collects sufficient conditions for membership is many such families. It would be interesting to better delineate such families, for example by characterizing necessary conditions for inclusion. When ψ is CGF-like and (V_t) is predictable, it is necessary for the increments ΔS_t to have finite conditional CGFs a.s. When S_t is a cumulative sum of i.i.d., real-valued random variables and $V_t \propto t$, the exis-

tence of the CGF is sufficient as well (Fact 1(a)). When the increments are not i.i.d., however, existence of conditional CGFs is no longer sufficient. When (V_t) is not predictable, as with self-normalized bounds, it is no longer necessary for increments to have finite CGFs (e.g., Example 5).

Determining optimal l_0 values Smaller values of l_0 are preferable since they lead to tighter bounds in Theorem 1. Most of the results in this paper take either $l_0 = 1$ for scalar observations or $l_0 = d$ for $d \times d$ matrix observations. Taking $\lambda \downarrow 0$ in Definition 1 shows we cannot have $l_0 < 1$. Furthermore, asymptotic results about maxima of independent Gaussians show that $l_0 = d$ is an asymptotic lower bound as $d \uparrow \infty$ for operator-norm inequalities over any class that includes matrices with independent Gaussians on the diagonal (Galambos, 1978; Boucheron et al., 2013, Exercise 2.17). It would be useful to derive more results about optimal values of l_0 in various settings.

Generalizing assumptions Definition 1 can be further generalized, allowing it to subsume more known inequalities and yield sharper results for certain cases. However, the corresponding general theorem and specific results are less user-friendly. We have chosen our Definition 1 and Theorem 1 to balance generality and tractability, but in Appendix D we present one possible generalization of our assumption and a corresponding general theorem and specific bound.

Polynomial line-crossing inequalities We have focused on exponential inequalities, but polynomial concentration also plays an important role in the literature. A theory of polynomial line-crossing analogous to that presented here may begin with the Dubins-Savage inequality (see Appendix B) and its l_p extension in Khan (2009).

Banach spaces The Banach space bounds in Section 4.4 give dimension-free l_p bounds for $2 \leq p < \infty$, but do not give l_∞ bounds. In particular, this does not yield operator-norm bounds for infinite-dimensional Hilbert-Schmidt operators, as in Minsker (2017). Extending Minsker's "effective rank" approach to the uniform bounds of this paper would be an interesting future extension.

6. Proofs

6.1. Proof of Lemma 1

Define the stopping time $\tau \coloneqq \inf\{t \in \mathcal{T} : L_t \geq a\}$, where $\inf \emptyset = \infty$. For any fixed $m \in \mathcal{T}$, Markov's inequality implies

$$\mathbb{P}\left(\tau \leq m \,|\, \mathcal{F}_0\right) = \mathbb{P}\left(L_{\tau \wedge m} \geq a \,|\, \mathcal{F}_0\right) \leq \frac{\mathbb{E}\left(L_{\tau \wedge m} \,|\, \mathcal{F}_0\right)}{a} \leq \frac{L_0}{a}, \tag{6.1}$$

where we have used Doob's optional stopping theorem for bounded stopping times in the final step (e.g., Durrett, 2017, Exercise 4.4.2; or Protter, 2005, Theorem I.17). Taking $m \to \infty$ and using the bounded convergence theorem yields $\mathbb{P}\left(\tau < \infty \,|\, \mathcal{F}_0\right) \leq L_0/a$, which is the desired conclusion.

6.2. Proof of Proposition 1

Applying Taylor's theorem to ψ at the origin, we have $\psi(\lambda) = \left[\frac{\psi''(0_+)}{2} + h(\lambda)\right]\lambda^2$ where $h(\lambda) \to 0$ as $\lambda \downarrow 0$. Choose $\lambda_0 > 0$ small enough so that $\psi(\lambda) \leq \psi''(0_+)\lambda^2$ for all $0 \leq \lambda < \lambda_0$. Then, setting $a := 2\psi''(0_+)$, $c := 1/\lambda_0$ and using that fact that $\psi_{G,c} \geq \psi_N$ for $c \geq 0$, we have $\psi(\lambda) \leq a\psi_N(\lambda) \leq a\psi_{G,c}(\lambda)$ for all $0 \leq \lambda < 1/c$. The same argument holds with ψ_E in place of ψ_G. □

6.3. Proof of Proposition 2

In each case, we show an inequality between two ψ functions. The conclusion then follows from the fact that is $\psi_1 \leq \psi_2$, then $\exp\{\lambda S_t - \psi_2(\lambda)V_t\} \leq \exp\{\lambda S_t - \psi_1(\lambda)V_t\}$, showing that the key condition of Definition 1 continues to hold with ψ_2 in place of ψ_1.

Part (1): the proof of Theorem 1 in Hoeffding (1963) shows that, for all $\mu \in (0,1)$ and all $t \in [0, 1-\mu)$,

$$(\mu+t)\log\left(\frac{\mu+t}{\mu}\right) + (1-\mu-t)\log\left(\frac{1-\mu-t}{1-\mu}\right) \geq t^2 \begin{cases} \frac{1}{1-2\mu}\log\left(\frac{1-\mu}{\mu}\right), & 0 < \mu < \frac{1}{2}, \\ \frac{1}{2\mu(1-\mu)}, & \frac{1}{2} \leq \mu < 1, \end{cases} \tag{6.2}$$

with equality at $t = 1-2\mu$. Substituting $\mu = g/(g+h)$ and $t = u/(g+h)$ for $u \in [0,h)$, some algebra shows that the left-hand side is equal to $gh\psi_B^\star(u/gh)$ and the right-hand side is equal to $\psi_N^\star(u)/\varphi(g,h)$, so that, for all $g,h > 0$ and $u \in [0,h)$, $\psi_B^\star(u/gh) \geq \psi_N^\star(u)/[gh\varphi(g,h)]$, with equality at $u = h-g$. The order-reversing and scaling properties of the Legendre-Fenchel transform now imply $\psi_B^{\star\star}(\lambda) \leq \psi_N^{\star\star}(\varphi(g,h)\lambda)/[gh\varphi(g,h)]$ for all $\lambda \geq 0$. Finally, since ψ_B and ψ_N are convex and continuous, each is equal to its biconjugate $\psi^{\star\star}$ by the Fenchel-Moreau theorem, so that $\psi_B(\lambda) \leq \frac{\varphi(g,h)}{gh}\psi_N(\lambda)$.

Part (2): This follows directly from equation (4.15) in Hoeffding (1963) which, in our notation, says that $\psi_B(\lambda) \leq \frac{(g+h)^2}{4gh}\psi_N(\lambda)$ for all $\lambda \in \mathbb{R}$.

Part (3): In our notation, Lemma 2.32 of Bercu et al. (2015) shows that $(g\psi_{B,g,1})^\star(u) \geq (g\psi_{P,1-g})^\star(u)$ for all $u \in [0,1]$ and $g > 0$. The order-reversing and scaling properties of the Legendre-Fenchel transform imply $\psi_{B,g,1}^{\star\star}(\lambda) \leq \psi_{P,1-g}^{\star\star}(\lambda)$ for all $\lambda \geq 0$. Since $\psi_{B,g,1}$ and $\psi_{P,1-g}$ are convex and continuous, each is equal to its biconjugate $\psi^{\star\star}$ by the Fenchel-Moreau theorem, so that $\psi_{B,g,1}(\lambda) \leq \psi_{P,1-g}(\lambda)$. The result now follows from algebraic identities involving ψ_B and ψ_P: for any $g,h > 0$,

$$\psi_{B,g,h}(\lambda) = \frac{1}{h^2}\psi_{B,g/h,1}(h\lambda) \leq \frac{1}{h^2}\psi_{P,(h-g)/h}(h\lambda) = \psi_{P,h-g}(\lambda). \tag{6.3}$$

Part (4) is immediate from the definition $\psi_P = \psi_N$ when $c = 0$.

Part (5): since $\psi''_{P,c_P}(\lambda) = e^{c_P\lambda}$ and $\psi''_{G,c_G}(\lambda) = (1-c_G\lambda)^{-3}$,

$$\frac{\psi''_{P,c}(\lambda)}{\psi''_{G,c/3}(\lambda)} = (1-c\lambda/3)^3 e^{c\lambda} = f(1-c\lambda/3), \quad \text{where } f(y) = y^3 e^{3(1-y)}. \tag{6.4}$$

We have $f(1) = 1$ and $f'(y) = 3y^2 e^{3(1-y)}(1-y)$, so that $f'(y) \le 0$ for $y \ge 1$ and $f'(y) \ge 0$ for $y \le 1$. Hence $f(y) \le f(1) = 1$ for all y, i.e., $\psi''_{P,c}(\lambda) \le \psi''_{G,c/3}(\lambda)$ for all λ. Since $\psi_{P,c}(0) = \psi_{G,c/3}(0) = 0$ and $\psi'_{P,c}(0) = \psi'_{G,c/3}(0) = 0$, we conclude $\psi_{P,c}(\lambda) \le \psi_{G,c/3}(\lambda)$ for all λ.

Parts (6, 7, 8): some algebra shows that

$$\psi'_{G,c_G}(\lambda) - \psi'_{E,c_E}(\lambda) = \frac{\lambda^2[3c_G - 2c_E + c_G(c_E - 2c_G)\lambda]}{2(1-c_G\lambda)^2(1-c_E\lambda)}. \tag{6.5}$$

Since $\psi_{G,c_G}(0) = \psi_{E,c_E}(0) = 0$, we have $\psi_{G,c_G}(\lambda) \ge (\le)\, \psi_{E,c_E}(\lambda)$ for all λ if $\psi'_{G,c_G} \ge (\le)\, \psi'_{E,c_E}$ for all λ, and (6.5) shows the latter is true if and only if $f(\lambda) \coloneqq 3c_G - 2c_E + c_G(c_E - 2c_G)\lambda \ge (\le)\, 0$ for all λ. Note we need only check the domain $0 \le \lambda < c_E^{-1} \wedge (2c_G)^{-1}$ on which both functions are defined.

- For part (6), if $c_E = 3c_G/2$, then $f(\lambda) = -c_G^2\lambda/2 \le 0$, so that $\psi_{G,c} \le \psi_{E,3c/2}$ for $c \in \mathbb{R}$.
- For part (7), if $c_G = c_E \ge 0$ then we have $f(\lambda) = c(1-c\lambda) \ge 0$ for $0 \le \lambda < c^{-1}$, so that $\psi_{E,c} \le \psi_{G,c}$ for $c \ge 0$.
- For part (8), if $c_G = c_E/2 < 0$, then $f(\lambda) = -c_E/2 > 0$, so that $\psi_{E,c} \le \psi_{G,c/2}$ for $c < 0$.

Part (9): from $\psi'_{P,2c}(\lambda) = \frac{e^{2c\lambda}-1}{2c}$ and $\psi'_{G,c}(\lambda) = \frac{\lambda(2-c\lambda)}{2(1-c\lambda)^2}$, we have

$$\psi'_{P,2c}(\lambda) - \psi'_{G,c}(\lambda) = \frac{1 - f(1+|c|\lambda)}{2|c|(1-c\lambda)^2}, \quad \text{where } f(y) = y^2 e^{2(1-y)}. \tag{6.6}$$

We have $f(1) = 1$ and $f'(y) = 2ye^{2(1-y)}(1-y) \le 0$ for all $y \ge 1$, so that $f(y) \le 1$ for all $y \ge 1$. Hence $\psi'_{P,2c}(\lambda) \ge \psi'_{G,c}(\lambda)$ for all $\lambda \ge 0$. Together with $\psi_{P,2c}(0) = \psi_{G,c}(0) = 0$, we conclude $\psi_{P,2c}(\lambda) \ge \psi_{G,c}(\lambda)$ for all $\lambda \ge 0$.

Part (10) follows from the fact that $\psi_{P,c} \uparrow \psi_N$ as $c \uparrow 0$.

Part (11): for any $g, h > 0$, we have

$$\psi'_{B,g,h}(\lambda) = \frac{e^{h\lambda} - e^{-g\lambda}}{ge^{h\lambda} + he^{-g\lambda}}, \tag{6.7}$$

so $\lim_{h\downarrow 0} \psi'_{B,g,h}(\lambda) = (1-e^{-g\lambda})/g = \psi'_{P,-g}(\lambda)$. Since $\psi_{B,g,h}(0) = \psi_{P,c} = 0$ for all $g, h > 0$ and all $c \in \mathbb{R}$, we see that $\lim_{h\downarrow 0} \psi_{B,g,h}(\lambda) = \psi_{P,-g}(\lambda)$ for all $\lambda \ge 0$. Furthermore, differentiating (6.7) with respect to h reveals

$$\frac{d}{dh}\psi'_{B,g,h}(\lambda) = \frac{e^{(h-g)\lambda}(g+h)^2\psi_{P,-(g+h)}(\lambda)}{(ge^{h\lambda} + he^{-g\lambda})^2} \ge 0, \tag{6.8}$$

which implies $\psi_{B,g,h}(\lambda)$ is nondecreasing with h for all $\lambda \geq 0$. We conclude $\psi_{B,g,h}(\lambda) \downarrow \psi_{P,-g}(\lambda)$ as $h \downarrow 0$, hence $\psi_{P,c} \leq \psi_{B,-c,h}$ for all $h > 0$ whenever $c < 0$.

To see that no other implications are possible, observe that, as $\lambda \to \infty$, $\psi_B(\lambda) = \mathcal{O}(\lambda)$, $\psi_N(\lambda) = \mathcal{O}(\lambda^2)$, and when $c > 0$, $\psi_P(\lambda) = \mathcal{O}(e^{c\lambda})$, while $\psi_G(\lambda)$ and $\psi_E(\lambda)$ diverge at a finite value of λ. So we cannot use $a\psi_B$ to upper bound any of the other ψ functions for any constant a. Likewise, we cannot use $a\psi_N$ to upper bound ψ_P, ψ_G or ψ_E, and we cannot use $a\psi_P$ to upper bound ψ_G or ψ_E.

Now if S_t is a sum of i.i.d. $\mathcal{N}(0,1)$ random variables, then (S_t) is sub-Gaussian with variance process $V_t = t$, and the exponential process $\exp\left\{\lambda S_t - \lambda^2 t/2\right\}$ is a martingale. Under any scaling of V_t by a constant $a > 0$, (S_t) cannot be sub-Bernoulli, because $\mathbb{E}\exp\left\{\lambda \Delta S_t - a\psi_B(\lambda)\right\} = \exp\left\{\lambda^2/2 - a\psi_B(\lambda)\right\} > 1$ for sufficiently large λ, so the exponential process $\exp\left\{\lambda S_t - \psi_B(\lambda)t\right\}$ will be expectation-increasing. Analogous arguments shows that other reverse implications are not possible.

To see that the above constants are the best possible when we allow only scaling of V_t by a constant, consider the third-order expansions of each ψ function about $\lambda = 0$:

$$\psi_B(\lambda) = \left[\frac{\lambda^2}{2} + \frac{(h-g)\lambda^3}{6}\right] + o(\lambda^3) \tag{6.9}$$

$$\psi_N(\lambda) = \frac{\lambda^2}{2} \tag{6.10}$$

$$\psi_P(\lambda) = \frac{\lambda^2}{2} + \frac{c\lambda^3}{6} + o(\lambda^3) \tag{6.11}$$

$$\psi_E(\lambda) = \frac{\lambda^2}{2} + \frac{c\lambda^3}{3} + o(\lambda^3) \tag{6.12}$$

$$\psi_G(\lambda) = \frac{\lambda^2}{2} + \frac{c\lambda^3}{2} + o(\lambda^3). \tag{6.13}$$

It is clear from these expansions that parts (3), (4), (5), (6), and (11) have the best possible constants. Part (7) is unimprovable because ψ_E diverges at $\lambda = 1/c$, and using any scale parameter in ψ_G smaller than c would make ψ_G finite at $\lambda = 1/c$. For part (8), recall that when $c < 0$, $\bar{b} = |c|^{-1}$ for ψ_E, while $\bar{b} = |2c|^{-1}$ for ψ_G. Hence, if $c' < c/2 < 0$, then $\lim_{\lambda\to\infty} \psi'_{G,c'}(\lambda) = |2c'|^{-1} < |c|^{-1} = \lim_{\lambda\to\infty} \psi'_{E,c}(\lambda)$, so that $\psi_{G,c'}(\lambda)$ must be smaller than $\psi_{E,c}(\lambda)$ for sufficiently large λ. Part (9) is unimprovable by an analogous argument.

For part (1), when $g \geq h$, we know that the constant of one in front of $\psi_N(\lambda)$ is the best possible from the expansions above. When $g < h$, some algebra shows that the inequality $\psi_{B,g,h}(\lambda) \leq \frac{\varphi(g,h)}{gh}\psi_N(\lambda)$ holds with equality at $\lambda = (h-g)/\varphi(g,h)$, so the constant cannot be improved. For part (2), it is easy to see that $\varphi(g,h) = \left(\frac{g+h}{2}\right)^2 = g^2$ when $g = h$, so the constant $\frac{(g+h)}{4gh}$ is the best possible of the form k/gh where k is a function of $g + h$ alone. □

A brief remark on the rationale behind part (2). In the "Bernoulli I"

(Fact 1(b)) and "Bernoulli II" (Lemma 3(a)) conditions, $V_t = ght$, so applying Proposition 2, part (2) leads to $V_t = \left(\frac{g+h}{2}\right)^2 t$, a function of the total range $g + h$ alone. This is useful in the common case that observations are known to be bounded in a range $[a, b]$, and an inequality is desired which depends only on the range $b - a$ and not on the location of the means within $[a, b]$.

6.4. An intermediate condition for sub-ψ processes

In discrete time, the following result capture a useful general condition on a matrix-valued process (Y_t) that is sufficient to show that the maximum-eigenvalue process $S_t = \gamma_{\max}(Y_t)$ is sub-ψ.

Lemma 4. *Let ψ be a real-valued function with domain $[0, \lambda_{\max})$. Let $(Y_t)_{t\in\mathbb{N}}$ be an adapted, $\mathcal{H}^d$-valued process. Let $(W_t)_{t\in\mathbb{N}}$ be predictable, $\mathcal{H}^d$-valued, and nondecreasing in the semidefinite order, with $W_0 = 0$. Let $(U_t)_{t\in\mathbb{N}}$ be defined by $U_0 = 0$ and $\Delta U_t = u_t(\Delta Y_t)$ for some $u_t : \mathbb{R} \to \mathbb{R}_{\geq 0}$, for each t. If, for all $t \in \mathbb{N}$ and $\lambda \in [0, \lambda_{\max})$, we have*

$$\log \mathbb{E}_{t-1} \exp\{\lambda \Delta Y_t - \psi(\lambda)\Delta U_t\} \preceq \psi(\lambda)\Delta W_t, \tag{6.14}$$

then $S_t = \gamma_{\max}(Y_t)$ is d-sub-ψ with variance process $V_t = \gamma_{\max}(U_t + W_t)$.

For a familiar example, suppose $d = 1$ and (Y_t) has independent increments. Let $W_t = t$, $U_t \equiv 0$ and $\psi(\lambda) = \lambda^2/2$. Then (6.14) reduces to the usual definition of a 1-sub-Gaussian random variable (Boucheron et al., 2013). For a self-normalized example, let (ΔY_t) be i.i.d. from any distribution symmetric about zero. Then, again letting $\psi(\lambda) = \lambda^2/2$, an argument due to de la Peña (1999) shows that (6.14) holds with $W_t \equiv 0$ and $U_t = \sum_{i=1}^t \Delta Y_i^2$. See Lemma 3(d) for a general statement of this condition.

The value $l_0 = d$, the ambient dimension, leads to a pre-factor of d in all of our operator-norm matrix bounds. In cases when $\sup_{t\in\mathcal{T}} \operatorname{rank}(U_t + W_t) \leq r < d$ a.s., the pre-factor d in our bounds may be replaced by r via an argument originally due to Oliveira (2010b). See Appendix A for details.

Proof of Lemma 4. The key result here is Lieb's concavity theorem:

Fact 3 (Lieb, 1973; Tropp, 2012)**.** For any fixed $H \in \mathcal{H}^d$, the function $A \mapsto \operatorname{tr}\exp\{H + \log(A)\}$ is concave on the positive-definite cone.

Fixing $\lambda \in [0, \lambda_{\max})$, Lieb's theorem and Jensen's inequality together imply

$$\begin{aligned}&\mathbb{E}_{t-1} \operatorname{tr}\exp\{\lambda Y_t - \psi(\lambda)\cdot(U_t + W_t)\}\\ &\quad\leq \operatorname{tr}\exp\left\{\lambda Y_{t-1} - \psi(\lambda)\cdot(U_{t-1} + W_t) + \log \mathbb{E}_{t-1} e^{\lambda\Delta Y_t - \psi(\lambda)\cdot\Delta U_t}\right\}.\end{aligned} \tag{6.15}$$

Now we apply inequality (6.14) to the expectation and use the monotonicity of the trace exponential to obtain

$$\mathbb{E}_{t-1} \operatorname{tr}\exp\{\lambda Y_t - \psi(\lambda)\cdot(U_t + W_t)\} \leq \operatorname{tr}\exp\{\lambda Y_{t-1} - \psi(\lambda)\cdot(U_{t-1} + W_{t-1})\}. \tag{6.16}$$

This shows that the process $L_t \coloneqq \operatorname{tr}\exp\{\lambda Y_t - \psi(\lambda)\cdot(U_t+W_t)\}$ is a supermartingale, with $L_0 = d$. Next we show that the key condition of Definition 1 holds, $L_t \geq \exp\{\lambda\gamma_{\max}(Y_t) - \psi(\lambda)\gamma_{\max}(U_t+W_t)\}$ a.s. for all t. We repeat a short argument from Tropp (2012). First, by the monotonicity of the trace exponential,

$$\operatorname{tr}\exp\{\lambda Y_t - \psi(\lambda)\cdot(U_t+W_t)\} \tag{6.17}$$

$$\geq \operatorname{tr}\exp\{\lambda Y_t - \psi(\lambda)\gamma_{\max}(U_t+W_t)I_d\} \tag{6.18}$$

$$\geq \gamma_{\max}\left(\exp\{\lambda Y_t - \psi(\lambda)\gamma_{\max}(U_t+W_t)I_d\}\right) =: B. \tag{6.19}$$

using the fact that the trace of a positive semidefinite matrix is at least as large as its maximum eigenvalue. Then the spectral mapping property gives

$$B = \exp\left\{\gamma_{\max}\left(\lambda Y_t - \psi(\lambda)\gamma_{\max}(U_t+W_t)I_d\right)\right\}. \tag{6.20}$$

Finally, we use the fact that $\gamma_{\max}(A - cI_d) = \gamma_{\max}(A) - c$ for any $A \in \mathcal{H}^d$ and $c \in \mathbb{R}$ to see that $B = \exp\{\lambda\gamma_{\max}(Y_t) - \psi(\lambda)\gamma_{\max}(U_t+W_t)\}$, completing the argument. □

6.5. Proof of Lemma 3

We rely on the following transfer rule for the semidefinite ordering.

Fact 4 (Tropp, 2012, eq. 2.2)**.** If $f(a) \leq g(a)$ for all $a \in S$, then $f(A) \preceq g(A)$ when the eigenvalues of A lie in S.

We make frequent use of the martingale property $\mathbb{E}_{t-1}\Delta Y_t = 0$, and prove in most cases that

$$\mathbb{E}_{t-1}\exp\{\lambda\Delta Y_t - \psi(\lambda)\Delta U_t\} \preceq \exp\{\psi(\lambda)\Delta W_t\} \tag{6.21}$$

for some (U_t) and (W_t), then invoke Lemma 4. This a stronger condition than property (6.14); the latter is implied by taking logarithms on both sides, recalling the monotonicity of the matrix logarithm.

Part (a): we adapt the argument of Bennett (1962, p. 42). Fix $\lambda \geq 0$ and choose real numbers u, v, w so that $e^{\lambda x} \leq ux^2 + vx + w$ for all $x \leq h$, with equality at $x = h$ and $x = -g$. Using the assumption $\Delta Y_t \preceq hI_d$, the transfer rule implies

$$\mathbb{E}_{t-1}e^{\lambda\Delta Y_t} \preceq u\mathbb{E}_{t-1}\Delta Y_t^2 + v\mathbb{E}_{t-1}\Delta Y_t + wI_d \preceq (ugh + w)I_d, \tag{6.22}$$

where the second inequality uses the assumption $\mathbb{E}_{t-1}\Delta Y_t^2 \preceq ghI_d$ and the martingale property. Now consider the random matrix

$$Z = \begin{cases} -gI_d, & \text{with probability } \frac{h}{h+g}, \\ hI_d, & \text{with probability } \frac{g}{h+g}. \end{cases} \tag{6.23}$$

Evidently $\mathbb{E}Z = 0$ and $\mathbb{E}Z^2 = ghI_d$, so Z also satisfies the aforementioned assumptions. Note that for any function $f : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}f(Z) = \left[\frac{h}{h+g} \cdot f(-g) + \frac{g}{h+g} \cdot f(h)\right] I_d. \tag{6.24}$$

By our choice of u, v, w, we see that $\mathbb{E}e^{\lambda Z} = \mathbb{E}(uZ^2 + vZ + wI_d) = (ugh + w)I_d$, so by direct calculation,

$$(ugh + w)I_d = \mathbb{E}e^{\lambda Z} = \left[\frac{h}{h+g} \cdot e^{-\lambda g} + \frac{g}{h+g} \cdot e^{\lambda h}\right] I_d = e^{gh\psi_B(\lambda)} I_d. \tag{6.25}$$

Combining (6.25) with (6.22) shows that (6.21) holds with $U_t \equiv 0$ and $W_t = ghtI_d$, as desired.

Part (b): As in Lemma 1 of Hoeffding (1963), we use the fact that $e^{\lambda x} \leq \frac{g+x}{g+h}e^{h\lambda} + \frac{h-x}{g+h}e^{-g\lambda}$ for all $x \in [-g, h]$, along with the transfer rule, to conclude that, for each t,

$$\mathbb{E}_{t-1}e^{\lambda \Delta Y_t} \preceq \left(\frac{G_t}{G_t + G_t}e^{H_t\lambda} + \frac{H_t}{G_t + H_t}e^{-G_t\lambda}\right) I_d = \exp\left\{G_t H_t \psi_{B,G_t,H_t}(\lambda)\right\} I_d. \tag{6.26}$$

Now the proof of Proposition 2 part (1) shows that $\psi_{B,g,h}(\lambda) \leq \varphi(g,h)\psi_N(\lambda)/gh$, so we have

$$\mathbb{E}_{t-1}e^{\lambda \Delta Y_t} \preceq \exp\left\{\psi_N(\lambda)\varphi(G_t, H_t)I_d\right\}, \tag{6.27}$$

which shows that (6.21) holds with $U_t \equiv 0$ and $\Delta W_t = \varphi(G_t, H_t)I_d$, as desired.

Part (c): the argument is identical to that for part (a), except for the use of $\psi_{B,g,h}(\lambda) \leq \frac{(g+h)^2}{4gh}\psi_N(\lambda)$ from the proof of Proposition 2 part (2).

Part (d): From the standard inequality $\cosh x \leq e^{x^2/2}$ we see that $f(x) := e^{-x^2/2}\cosh x \leq 1$ for all x. Introducing an independent Rademacher random variable ε, we have for any t,

$$\mathbb{E}_{t-1}\exp\left\{\lambda \Delta Y_t - \frac{\lambda^2 \Delta Y_t^2}{2}\right\} = \mathbb{E}_{t-1}\exp\left\{\lambda\varepsilon \Delta Y_t - \frac{\lambda^2 \Delta Y_t^2}{2}\right\} \tag{6.28}$$

$$= \mathbb{E}_{t-1}\mathbb{E}\left[\exp\left\{\lambda\varepsilon \Delta Y_t - \frac{\lambda^2 \Delta Y_t^2}{2}\right\} \,\middle|\, \mathcal{F}_{t-1}, \Delta Y_t\right] \tag{6.29}$$

$$= \mathbb{E}_{t-1}f(\lambda \Delta Y_t) \tag{6.30}$$

$$\preceq I_d, \tag{6.31}$$

applying the transfer rule in the last step. This shows that (6.21) holds with $U_t = [Y]_t$ and $W_t \equiv 0$.

Part (e): Lemma 4.1 of Fan et al. (2015) shows that

$$\exp\left\{\lambda x - [\log(1-\lambda)^{-1} - \lambda]x^2\right\} \leq 1 + \lambda x, \text{ for all } x \geq -1 \text{ and } 0 \leq \lambda < 1. \tag{6.32}$$

Applying the transfer rule and taking expectations, we have for any t,

$$\mathbb{E}_{t-1} \exp\left\{\lambda \cdot \frac{\Delta Y_t}{c} - [\log(1-\lambda)^{-1} - \lambda] \cdot \frac{\Delta Y_t^2}{c^2}\right\} \preceq I_d. \tag{6.33}$$

Replace λ with $c\lambda$ and identify ψ_E to complete the argument that (6.21) holds with $U_t = [Y]_t$ and $W_t \equiv 0$.

Part (f): Proposition 12 of Delyon (2009) shows that $e^{x-x^2/6} \leq 1 + x + x^2/3$ for all $x \in \mathbb{R}$. This implies, by the transfer rule,

$$\mathbb{E}_{t-1} \exp\left\{\lambda \Delta Y_t - \frac{\lambda^2}{6} \Delta Y_t^2\right\} \preceq I_d + \frac{\lambda^2}{3} \mathbb{E}_{t-1} \Delta Y_t^2 \tag{6.34}$$

$$\preceq \exp\left\{\frac{\lambda^2}{3} \mathbb{E}_{t-1} \Delta Y_t^2\right\}. \tag{6.35}$$

This shows that (6.21) holds with $U_t = [Y]_t/3$ and $W_t = 2\langle Y \rangle_t /3$.

Part (g): Proposition 12 of Delyon (2009), together with the fact that $e^{-x} + x - 1 \leq x^2/2$ for $x \geq 0$, shows that $e^{x - x_+^2/2} \leq 1 + x + x_-^2/2$. Again the transfer rule implies

$$\mathbb{E}_{t-1} \exp\left\{\lambda \Delta Y_t - \frac{\lambda^2}{2} (\Delta Y_t)_+^2\right\} \preceq I_d + \frac{\lambda^2}{2} \mathbb{E}_{t-1} (\Delta Y_t)_-^2 \tag{6.36}$$

$$\preceq \exp\left\{\frac{\lambda^2}{2} \mathbb{E}_{t-1} (\Delta Y_t)_-^2\right\}. \tag{6.37}$$

This shows that (6.21) holds with $U_t = [Y_+]_t/2$ and $W_t = \langle Y_- \rangle_t /2$.

Part (h): we appeal to part (d) to see that S_t is d-sub-Gaussian with variance process $V_t = \gamma_{\max}(\frac{1}{3}[Y]_t + \frac{2}{3}\langle Y \rangle_t)$. Now the condition $\Delta Y_t^2 \preceq A_t^2$ ensures that $\frac{1}{3}[Y]_t + \frac{2}{3}\langle Y \rangle_t \preceq \sum_{i=1}^t A_i^2$, hence $V_t \leq \gamma_{\max}(\sum_{i=1}^t A_i^2)$. Substituting this larger variance process only makes the exponential process in Definition 1 smaller, so the assumption remains satisfied.

Part (i): the proof of Corollary 2.2 in Fan et al. (2015) is based on the inequality $e^{x-x^2/2} \leq 1 + x + x_-^3/3$ for all $x \in \mathbb{R}$. The transfer rule implies

$$\mathbb{E}_{t-1} \exp\left\{\lambda \Delta Y_t - \frac{\lambda^2}{2} \Delta Y_t^2\right\} \preceq I_d + \frac{\lambda^3}{3} \mathbb{E}_{t-1} (\Delta Y_t)_-^3 \preceq \exp\left\{\frac{\lambda^3}{3} \mathbb{E}_{t-1} (\Delta Y_t)_-^3\right\}. \tag{6.38}$$

Setting $c = 1/6$ in ψ_G, we have for all $x \in [0, 6)$ the obvious inequality $x^2/2 \leq \psi_G(x)$ and we claim $x^3/3 \leq \psi_G(x)$ as well; indeed,

$$\frac{x^3/3}{x^2/2(1 - x/6)} = \frac{x(6-x)}{9}, \tag{6.39}$$

which reaches a maximum value of one at $x = 3$. The transfer rule now implies

$$\mathbb{E}_{t-1} \exp\left\{\lambda \Delta Y_t - \psi_G(\lambda) \Delta Y_t^2\right\} \preceq \mathbb{E}_{t-1} \exp\left\{\lambda \Delta Y_t - \frac{\lambda^2}{2} \Delta Y_t^2\right\} \tag{6.40}$$

$$\preceq \exp\left\{\frac{\lambda^3}{3}\mathbb{E}_{t-1}(\Delta Y_t)_-^3\right\} \tag{6.41}$$

$$\preceq \exp\left\{\psi_G(\lambda)\mathbb{E}_{t-1}(\Delta Y_t)_-^3\right\}, \tag{6.42}$$

which shows that (6.21) holds with $U_t = [Y]_t$ and $V_t = \sum_{i=1}^t \mathbb{E}_{i-1}|\Delta Y_i|^3$.

6.6. *Proof of Corollary 2*

Define the $\mathcal{H}^{d_1+d_2}$-valued process (Y_t) using the dilation of B_t:

$$\Delta Y_t := \epsilon_t \begin{pmatrix} 0 & B_t \\ B_t^\star & 0 \end{pmatrix}. \tag{6.43}$$

Since the dilation operation is linear and preserves spectral information, we have $S_t = \gamma_{\max}(Y_t) = \|\sum_{i=1}^t \epsilon_i B_i\|_{\mathrm{op}}$ (Tropp, 2012, Eq. 2.12). Furthermore, since each B_i is fixed and ϵ_i is 1-sub-Gaussian (in the usual sense for scalar random variables), (Y_t) satisfies the conditions of Lemma 4 with $\psi = \psi_N$, $U_t \equiv 0$, and

$$W_t = \sum_{i=1}^t \begin{pmatrix} B_i B_i^\star & 0 \\ 0 & B_i^\star B_i \end{pmatrix}, \tag{6.44}$$

by Tropp (2012, Lemma 4.3). Hence (S_t) is (d_1+d_2)-sub-Gaussian with variance process

$$V_t = \|W_t\|_{\mathrm{op}} = \max\left\{\left\|\sum_{i=1}^t B_i B_i^\star\right\|_{\mathrm{op}}, \left\|\sum_{i=1}^t B_i^\star B_i\right\|_{\mathrm{op}}\right\}. \tag{6.45}$$

The result now follows from Theorem 1(b).

6.7. *Proof of Corollary 3*

First, observe $\mathfrak{s}^{-1}(u) = \psi'(D(u))$ for any $u \in (0, \bar{b})$. Indeed, from the definition of $\mathfrak{s}(\cdot)$ and Lemma 2(v) we see that if $u = \mathfrak{s}(v)$ then $D(u) = \psi^{\star\prime}(v) = \psi'^{-1}(v)$, so that $v = \psi'(D(u))$. This identity will be used below.

Now let $h(b) := m\psi^\star(b) + (x - bm)\psi^{\star\prime}(b)$. We will show the following:

(I) If $m = 0$ or $b \le \mathfrak{s}\left(\frac{x}{m}\right)$, then $h(b) \le (x - (b \wedge \bar{b})m)D(b)$.
(II) If $m > 0$ then $h(b) \le m\psi^\star\left(\frac{x}{m}\right) = h\left(\frac{x}{m}\right)$.

Together with Theorem 1(d) these prove that (4.10) holds, and (4.11) follows upon setting $x = bm$.

First suppose $m = 0$, so it suffices to show $\psi^{\star\prime}(b) \le D(b)$ to prove (I) in this case. But Lemma 2(vi) implies $u \le \mathfrak{s}^{-1}(u)$ for any $u \in [0, \bar{b})$, and together with the convexity of $\psi^\star$, we have $\psi^{\star\prime}(b) \le \psi^{\star\prime}(\mathfrak{s}^{-1}(b))$. Then the identities $\mathfrak{s}^{-1}(u) = \psi'(D(u))$ and $\psi^{\star\prime} = \psi'^{-1}$ imply $\psi^{\star\prime}(\mathfrak{s}^{-1}(b)) = D(b)$.

Now suppose $m > 0$. It is easy to see that $h'(b) = (x - bm)\psi^{\star\prime\prime}(b)$. The convexity of $\psi^\star$ now implies h is nondecreasing for $b \le x/m$ and nonincreasing for $b \ge x/m$. Hence $h(b)$ is maximized at $b = x/m$, which proves (II). To prove (I) in this case, we claim that $h(\mathfrak{s}^{-1}(b)) = (x - bm)D(b)$. Then the condition $b \le \mathfrak{s}(x/m)$ and Lemma 2(vi) imply $b < \mathfrak{s}^{-1}(b) \le x/m$, so that $h(b) \le h(\mathfrak{s}^{-1}(b))$ since h is nondecreasing on this region, which is (I).

To prove the claim, substitute the identity $\mathfrak{s}^{-1}(u) = \psi'(D(u))$ into the definition of $h(\cdot)$, yielding

$$h(\mathfrak{s}^{-1}(b)) = h(\psi'(D(b))) = m\psi^\star(\psi'(D(b))) + [x - m\psi'(D(b))]D(b). \tag{6.46}$$

Now use the identity $\psi^\star(u) = u\psi^{\star\prime}(u) - \psi(\psi^{\star\prime}(u))$ to obtain

$$h(\mathfrak{s}^{-1}(b)) = xD(b) - m\psi(D(b)) \tag{6.47}$$

$$= xD(b) - mbD(b), \tag{6.48}$$

where the final equality uses Lemma 2(v), proving the claim.

The second statement (4.11) follows directly from Theorem 1(d) with $x = mb$. When $b \le \bar{b}$, Lemma 2(vi) implies $s(x/m) \le x/m = b$, so the second case in (2.10) applies. When $b > \bar{b}$, we have $x > m\bar{b}$, so the first case in (2.10) applies. Noting that $D(b) = \infty = \psi^\star(b)$ in this case using Lemma 2(i), we see that (4.11) remains valid. □

6.8. Proof of Corollary 10

We invoke arguments from Pinelis (1994) and Pinelis (1992) to show that Definition 1 is satisfied.

For **part (a)**, the proofs of Theorem 3 in Pinelis (1994) and Theorem 3 in Pinelis (1992) show that, for each $t \in \mathbb{N}$,

$$\mathbb{E}_{t-1} \cosh\big(\lambda\Psi(Y_t)\big) \le e^{\lambda^2 D_\star^2 c_t^2/2} \cosh\big(\lambda\Psi(Y_{t-1})\big). \tag{6.49}$$

Hence $L_t := \cosh\big(\lambda\Psi(Y_t)\big)e^{-\lambda^2 D_\star^2 \sum_{i=1}^t c_i^2/2}$ is a supermartingale, and the inequality $\cosh x > e^x/2$ implies that Definition 1 is satisfied for $S_t = \Psi(Y_t)$, $V_t = D_\star^2 \sum_{i=1}^t c_i^2$ and $\psi = \psi_N$ with $\lambda_{\max} = \infty$ and $l_0 = 2$. The conclusion (4.28) follows from a slight reparametrization of V_t to make $D_\star^2$ explicit in the bound.

For **part (b)**, the proof of Theorem 3 in Pinelis (1994) shows that

$$\mathbb{E}_{t-1} \cosh\big(\lambda\Psi(Y_t)\big) \le \exp\Big\{D_\star^2 \mathbb{E}_{t-1}\Big[e^{\lambda\|\Delta Y_t\|} - \lambda\|\Delta Y_t\| - 1\Big]\Big\} \cosh\big(\lambda\Psi(Y_{t-1})\big) \tag{6.50}$$

$$\le \exp\left\{D_\star^2 \left(\frac{e^{c\lambda} - c\lambda - 1}{c^2}\right) \mathbb{E}_{t-1}\|\Delta Y_t\|^2\right\} \cosh\big(\lambda\Psi(Y_{t-1})\big), \tag{6.51}$$

using the fact that $(e^{c\lambda} - c\lambda - 1)/c^2$ is nondecreasing. Hence the process $L_t := \cosh\big(\lambda\Psi(Y_t)\big)e^{-\psi_P(\lambda)D_\star^2 \sum_{i=1}^t \mathbb{E}_{i-1}\|X_i\|^2}$ is a supermartingale, and we see that Definition 1 is satisfied for $S_t = \Psi(Y_t)$, $V_t = D_\star^2 \sum_{i=1}^t \mathbb{E}_{i-1}\|X_i\|^2$ and $\psi = \psi_P$ with $\lambda_{\max} = \infty$ and $l_0 = 2$. The conclusion (4.29) follows as in part (a). □

6.9. Proof of Corollary 13

We invoke Theorem 2 of Robbins and Siegmund (1970) for the sum S_n/σ with $g(t) = a/\sigma + b\sigma t$, noting that

$$\lim_{m\to\infty} \mathbb{P}\left(\exists t \in \mathbb{N} : \frac{S_n}{\sqrt{m}} \geq a + \frac{bt\sigma^2}{m}\right) = \lim_{m\to\infty} \mathbb{P}\left(\exists t \in \mathbb{N} : \frac{S_n}{\sigma} \geq \sqrt{m} g\left(\frac{t}{m}\right)\right). \tag{6.52}$$

It is easy to verify the conditions of parts (i) and (ii) of Robbins and Siegmund's theorem, yielding the conclusion

$$\lim_{m\to\infty} \mathbb{P}\left(\exists t \in \mathbb{N} : \frac{S_n}{\sigma} \geq \sqrt{m} g\left(\frac{t}{m}\right)\right) = \mathbb{P}\left(\exists t \in (0,\infty) : B_t \geq g(t)\right), \tag{6.53}$$

where (B_t) is standard Brownian motion. The latter probability is equal to e^{-2ab} by the standard line-crossing formula for Brownian motion (e.g., Durrett, 2017, Exercise 7.5.2). □

6.10. Proof of Proposition 3

From the definition of $D(\cdot)$, we see that $M_t = \exp\{D(b) \cdot (S_t - bV_t)\}$. Since τ is a stopping time, $(M_{t\wedge\tau})$ is a martingale, so $1 = \mathbb{E}M_{t\wedge\tau}$ for each $t \in \mathbb{N}$. The third condition of the proposition ensures that $M_{t\wedge\tau} \leq e^{D(b)\cdot(a+\epsilon)}$ for all t a.s., so by dominated convergence we have $\mathbb{E}M_{t\wedge\tau} \to \mathbb{E}M_\tau = 1$, where M_τ is defined as the a.s. limit of $(M_{t\wedge\tau})$, whose existence is guaranteed since the stopped process is a nonnegative martingale. The second condition of the proposition implies $M_t \overset{\text{a.s.}}{\to} 0$, hence

$$1 = \mathbb{E}M_\tau = \mathbb{E}M_\tau 1_{(\tau<\infty)} + \mathbb{E}M_\infty 1_{(\tau=\infty)} \tag{6.54}$$

$$\leq \exp\{D(b)\cdot(a+\epsilon)\}\,\mathbb{P}(\tau < \infty), \tag{6.55}$$

which gives the desired lower bound on $\mathbb{P}(\tau < \infty)$.

6.11. Proof of Corollary 14

The conclusion follows immediately from Proposition 3 with $\epsilon = 0$ once we show that the conditions of the proposition are satisfied for (S_t) with $V_t = [S]_t$ and $\psi = \psi_N$.

In this case, since (S_t) has continuous paths a.s, (M_t) is the stochastic exponential of the process $(D(b)S_t)$ (Protter, 2005, Ch. II, Theorem 37). Kazamaki's criterion is sufficient to ensure (M_t) is a martingale (Protter, 2005, Ch. III, Theorem 44) and $M_0 = 1$ since $S_0 = 0$. This shows that condition (1) of Proposition 3 holds. Condition (3) follows directly from the continuity of paths of (S_t).

It remains to show that condition (2) holds. For this we express (S_t) as a time change of Brownian motion (Protter, 2005, Ch. II, Theorem 42): $S_t = B_{[S]_t}$ where (B_t) is a standard Brownian motion (with respect to a different filtration). From the law of the iterated logarithm we know that $B_t/t \overset{\text{a.s.}}{\to} 0$ as $t \to \infty$, hence $S_t - b[S]_t = [S]_t(B_{[S_t]}/[S]_t - b) \to -\infty$ since $[S]_t \uparrow \infty$. □

6.12. Proof of Proposition 4

Lemma 2.4 of Boucheron et al. (2013) shows that

$$f_\alpha(t) = \inf_\lambda \left[\frac{\log \alpha^{-1}}{\lambda} + \frac{\psi(\lambda)}{\lambda} \cdot t \right], \tag{6.56}$$

so that $f_\alpha(t)$ is a pointwise infimum of lines indexed by λ with intercepts $a_\lambda = (\log \alpha^{-1})/\lambda$ and slopes $b_\lambda = \psi(\lambda)/\lambda$. Hence $D(b_\lambda) = \lambda$, and by Theorem 1 the crossing probability of each such line is $e^{-a_\lambda D(b_\lambda)} = \alpha$. Note we have also shown that f_α is concave. The optimizer $\lambda_\star(t)$ in (6.56) is the solution in λ of $\lambda\psi'(\lambda) - \psi(\lambda) = (\log \alpha^{-1})/t$. The left-hand side of this equation has positive derivative in λ by the convexity of ψ, so the map $t \mapsto \lambda_\star(t)$ is injective. Hence the optimum line $a_{\lambda_\star(m)} + b_{\lambda_\star(m)} t$ is tangent to the curve $f_\alpha(t)$ at $t = m$.

Appendix A: Sharpened pre-factors based on rank

This argument is adapted from Wainwright (2017), though the idea originates in Oliveira (2010b). Suppose the conditions of Lemma 4 hold and

$$\sup_{t \in \mathcal{T}} \operatorname{rank}(U_t + W_t) \leq r < d, \text{ a.s.} \tag{A.1}$$

Since $\Delta U_t \succeq 0$ and $\Delta W_t \succeq 0$ for all t, we know that $\operatorname{range}(U_t + W_t) \subseteq S$ for all t a.s., where S is an r-dimensional subspace. Inequality (6.14) implies that $\operatorname{range}(Y_t) \subseteq S$ for all t a.s. as well. Let M be a $d \times r$ matrix whose columns form an orthonormal basis for this subspace. Then the r-dimensional process $\widetilde{Y}_t \coloneqq M^\star Y_t M$ has the same spectrum as Y_t for all t a.s., so we may apply our bounds to $(\widetilde{Y}_t)$, with $(\widetilde{U}_t)$ and $(\widetilde{W}_t)$ defined analogously, to obtain bounds with $l_0 = r$. □

Appendix B: Relation to the Dubins-Savage inequality

The Dubins-Savage inequality (Dubins and Savage, 1965) says that for any martingale S_t in discrete time with $S_0 = 0$, setting $V_t = \sum_{i=1}^t \operatorname{Var}_{i-1}(S_t - S_{t-1})$, we have

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t \geq a + bV_t\right) \leq \frac{1}{1+ab}. \tag{B.1}$$

The Dubins-Savage inequality may be proved by means similar to ours, invoking Ville's inequality for a suitable supermartingale. The relationship of our bounds to the Dubins-Savage inequality is analogous to that between fixed-time Cramér-Chernoff bounds and Chebyshev's inequality. More precisely, the Dubins-Savage inequality is analogous to Uspensky's one-sided version of Chebyshev's inequality (Uspensky, 1937; Bennett, 1962):

$$\mathbb{P}(X - \mathbb{E}X \geq x) \leq \frac{\operatorname{Var} X}{\operatorname{Var} X + x^2}. \tag{B.2}$$

Similar to our Theorem 1(b), we may optimize the RHS of (B.1) over all lines passing through a point (m, x) to obtain the equivalent bound

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t \geq x + \frac{x}{2m}(V_t - m)\right) \leq \frac{m}{m + x^2/4}, \tag{B.3}$$

recovering Uspensky's inequality (B.2) with $x/2$ in place of x. The Dubins-Savage inequality does not recover Uspensky's inequality at the fixed time m—something is necessarily lost in going from a fixed time to a uniform bound. Compare our Theorem 1(b), which exactly recovers the fixed-time Cramér-Chernoff bound (4.2). For these exponential bounds, we lose nothing in going from a fixed time to a uniform bound.

Appendix C: Graphical comparison of ψ functions

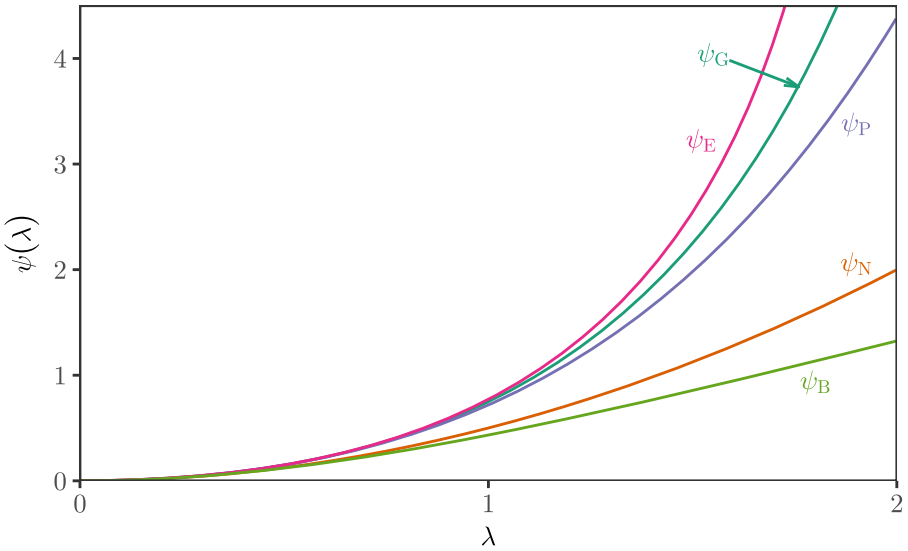

Fig 7: Comparison of ψ functions given in Table 2. We have set $g = h = 1$ in ψ_B, $c = 1$ in ψ_P, $c = 1/3$ in ψ_G, and $c = 1/2$ in ψ_E. These are all values that might be used in bounding a process with $[-1, 1]$-valued increments using the same variance process; see Figure 3 and Proposition 2. In general, bounds based on different ψ functions may have different assumptions and variance processes, so may not be comparable based on ψ functions alone. However, with identical variance processes, a smaller ψ function yields a tighter bound. Note all functions behave like $\psi_N(\lambda) = \lambda^2/2$ near the origin.

Figure 7 illustrates together the five standard ψ functions discussed in Section 3, to help the reader gain intuition. With the given parameter settings, the inequalities apparent in the figure do hold for all $\lambda \geq 0$: $\psi_B(\lambda) \leq \psi_N(\lambda) \leq \psi_P(\lambda) \leq \psi_G(\lambda) \leq \psi_E(\lambda)$. See the proof of Proposition 2 in Section 6.3.

Appendix D: A more general boundary-crossing result

The following assumption weakens Definition 1, replacing the product $\psi(\lambda)\Delta V_t$ with a function $f(\lambda, \Delta V_t)$.

Assumption 1. *Let $(S_t)_{t\in\mathbb{N}\cup\{0\}}$ and $(V_t)_{t\in\mathbb{N}\cup\{0\}}$ be two real-valued processes adapted to an underlying filtration $(\mathcal{F}_t)_{t\in\mathbb{N}\cup\{0\}}$ with $S_0 = V_0 = 0$ a.s. and $V_t \geq 0$ a.s. for all $t \in \mathbb{N}$. Let $f : [0, \lambda_{\max}) \times (0, \infty) \to \mathbb{R}$ be concave in its second argument for each value of the first, and let $l_0 \in [1, \infty)$. We assume, for each $\lambda \in [0, \lambda_{\max})$, there exists a supermartingale $(L_t(\lambda))_{t\in\mathbb{N}\cup\{0\}}$ with respect to $(\mathcal{F}_t)$ such that $L_0 \leq l_0$ a.s. and $\exp\left\{\lambda S_t - \sum_{i=1}^t f(\lambda, \Delta V_i)\right\} \leq L_t(\lambda)$ a.s. for all $t \in \mathbb{N}$.*

Clearly, when $f(\lambda, v) \equiv \psi(\lambda) \cdot v$ for some ψ, Definition 1 holds and Theorem 1 applies. Under the weaker Assumption 1 we have the following results:

Theorem 2. *If Assumption 1 holds for some real-valued processes (S_t) and (V_t), then for any $\lambda \in [0, \lambda_{\max})$ and $a > 0$, we have*

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t \geq a + \frac{t f(\lambda, V_t/t)}{\lambda}\right) \leq l_0 e^{-a\lambda}. \tag{D.1}$$

Furthermore, if $f_v(\cdot) \coloneqq f(\cdot, v)$ is CGF-like for each $v > 0$, then for any $n \in \mathbb{N}$, $m > 0$ and $0 \leq x < n \sup_\lambda f'_{m/n}(\lambda)$, we have

$$\mathbb{P}\left(\exists t \leq n : S_t \geq x + \frac{n}{\lambda_\star}\left[f\left(\lambda_\star, \frac{V_t}{n}\right) - f\left(\lambda_\star, \frac{m}{n}\right)\right]\right) \leq l_0 \exp\left\{-n f^\star_{m/n}\left(\frac{x}{n}\right)\right\} \tag{D.2}$$

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t \geq x + \frac{t f(\lambda_\star, m/t) - n f(\lambda_\star, m/n)}{\lambda_\star}\right) \leq l_0 \exp\left\{-n f^\star_{m/n}\left(\frac{x}{n}\right)\right\} \tag{D.3}$$

where $\lambda_\star \coloneqq (f^\star_{m/n})'(x/n)$.

The proof follows the same principles as that of Theorem 1 and is omitted for brevity. One application of this result is to martingales with bounded increments, making use of ψ_B:

Corollary 15. *Let $(Y_t)_{t\in\mathbb{N}}$ be an $\mathcal{H}^d$-valued martingale and let $S_t \coloneqq \gamma_{\max}(Y_t)$. Suppose $\gamma_{\max}(\Delta Y_t) \leq c$ for all t for some $c > 0$, and let $V_t \coloneqq \gamma_{\max}(\langle Y\rangle_t)$. Then for any $x, m > 0, n \in \mathbb{N}$ we have*

$$\mathbb{P}\left(\exists t \leq n : S_t \geq x + n\left[g\left(\frac{V_t}{n}\right) - g\left(\frac{m}{n}\right)\right]\right) \leq \left[\left(\frac{m}{x+m}\right)^{x+m}\left(\frac{n}{n-x}\right)^{n-x}\right]^{n/(n+m)}, \tag{D.4}$$

and

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t \geq x + tg\left(\frac{m}{t}\right) - ng\left(\frac{m}{n}\right)\right) \leq \left[\left(\frac{m}{x+m}\right)^{x+m}\left(\frac{n}{n-x}\right)^{n-x}\right]^{n/(n+m)}, \quad \text{(D.5)}$$

where

$$g(v) \coloneqq \frac{m+cn}{n(v+c)\log\xi}\left[v\xi^{\frac{cn}{m+cn}} + c\xi^{-\frac{vn}{m+cn}}\right] \quad \textit{and} \quad \xi \coloneqq \frac{1+x/m}{1-x/cn}. \quad \text{(D.6)}$$

This generalizes Theorem 2.1 of Fan et al. (2012) [B, D].

One can further generalize Assumption 1 by replacing $\sum_i f(\lambda, \Delta V_i)$ with $\sum_i f_i(\lambda, \Delta V_i)$, permitting f_i to vary with time, but the added generality further weakens the conclusions we can draw.

Appendix E: Equivalent sub-exponential conditions

Here we show that our sub-exponential condition (E.1) is equivalent to another common definition (E.2) (Wainwright, 2017). We rephrase both conditions for the right tail of a mean-zero random variable X.

Proposition 5. *For a zero-mean random variable X, the following are equivalent:*

1. *There exist $\sigma^2 > 0$ and $c > 0$ such that*
$$\log \mathbb{E}e^{\lambda X} \leq \frac{[-\log(1-c\lambda) - c\lambda]\,\sigma^2}{c^2} \qquad \textit{for all } \lambda \in [0, 1/c). \quad \text{(E.1)}$$
2. *There exist $\nu > 0$ and $\alpha > 0$ such that*
$$\log \mathbb{E}e^{\lambda X} \leq \frac{\lambda^2 \nu}{2} \qquad \textit{for all } \lambda \in [0, 1/\alpha). \quad \text{(E.2)}$$

Proof. Suppose the first condition holds. A Taylor expansion of $[-\log(1-c\lambda) - c\lambda]/c^2$ about $\lambda = 0$ yields

$$\frac{[-\log(1-c\lambda) - c\lambda]\,\sigma^2}{c^2} = \frac{\lambda^2\sigma^2}{2} + \lambda^2\sigma^2 \sum_{k=1}^{\infty} \frac{(c\lambda)^k}{2+k} = \frac{\lambda^2\sigma^2}{2} + o(\lambda^2). \quad \text{(E.3)}$$

So choosing $\nu > \sigma^2$, we can find α sufficiently large to ensure that

$$\frac{[-\log(1-c\lambda) - c\lambda]\,\sigma^2}{c^2} \leq \frac{\lambda^2\nu}{2} \qquad \text{for all } \lambda \in [0, 1/\alpha), \quad \text{(E.4)}$$

implying the second condition holds.

Now suppose the second condition holds. Then since $\lambda \geq 0$, the above series expansion shows that the first condition holds with $\sigma^2 = \nu$ and $c = \alpha$. □

Note that if the first condition (E.1) applies to both X and $-X$, then X satisfies the usual, two-tailed sub-exponential condition, $\log \mathbb{E}e^{\lambda X} \leq \lambda^2\nu/2$ for all $|\lambda| < 1/\alpha$.

Acknowledgments

We thank Victor de la Peña and Joel Tropp for their detailed perspectives on our work. We also thank Eli Ben-Michael, Lester Mackey, Xinjia Chen, and referees for helpful comments. Howard thanks Office of Naval Research (ONR) Grant N00014-15-1-2367. Sekhon thanks ONR grants N00014-17-1-2176 and N00014-15-1-2367.

References

Ahlswede, R. and Winter, A. (2002), 'Strong converse for identification via quantum channels', *IEEE Trans. Inf. Theor.* **48**(3), 569–579. MR1889969

Azuma, K. (1967), 'Weighted sums of certain dependent random variables.', *Tohoku Mathematical Journal* **19**(3), 357–367. MR0221571

Bacry, E., Gaïffas, S. and Muzy, J.-F. (2018), 'Concentration inequalities for matrix martingales in continuous time', *Probability Theory and Related Fields* **170**(1-2), 525–553. MR3748330

Ball, K., Carlen, E. A. and Lieb, E. H. (1994), 'Sharp uniform convexity and smoothness inequalities for trace norms', *Inventiones Mathematicae* **115**(1), 463–482. MR1262940

Barlow, M. T., Jacka, S. D. and Yor, M. (1986), 'Inequalities for a pair of processes stopped at a random time', *Proceedings of the London Mathematical Society* **s3-52**(1), 142–172. MR0812449

Bennett, G. (1962), 'Probability inequalities for the sum of independent random variables', *Journal of the American Statistical Association* **57**(297), 33–45. MR0163335

Bercu, B., Delyon, B. and Rio, E. (2015), *Concentration inequalities for sums and martingales*, Springer International Publishing, Cham. MR3363542

Bercu, B. and Touati, A. (2008), 'Exponential inequalities for self-normalized martingales with applications', *The Annals of Applied Probability* **18**(5), 1848–1869. MR2462551

Bernstein, S. (1927), *Theory of probability*, Gastehizdat Publishing House, Moscow.

Blackwell, D. (1997), Large deviations for martingales, *in* 'Festschrift for Lucien Le Cam', Springer, New York, NY, pp. 89–91. MR1462940

Blackwell, D. and Freedman, D. A. (1973), 'On the amount of variance needed to escape from a strip', *The Annals of Probability* **1**(5), 772–787. MR0356214

Boucheron, S., Lugosi, G. and Massart, P. (2013), *Concentration inequalities: a nonasymptotic theory of independence*, 1st edn, Oxford University Press, Oxford. MR3185193

Chen, X. (2012a), A statistical approach for performance analysis of uncertain systems, *in* R. E. Karlsen, D. W. Gage, C. M. Shoemaker and G. R. Gerhart, eds, 'Unmanned Systems Technology XIV', Vol. 8387, International Society for Optics and Photonics, SPIE, pp. 583–594.

Chen, X. (2012b), 'New optional stopping theorems and maximal inequalities on stochastic processes', 1207.3733 [math, stat].

Chernoff, H. (1952), 'A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations', *The Annals of Mathematical Statistics* **23**(4), 493–507. MR0057518

Christofides, D. and Markström, K. (2007), 'Expansion properties of random Cayley graphs and vertex transitive graphs via matrix martingales', *Random Structures & Algorithms* **32**(1), 88–100. MR2371053

Chung, F. and Lu, L. (2006), 'Concentration inequalities and martingale inequalities: a survey', *Internet Mathematics* **3**(1), 79–127. MR2283885

Craig, C. C. (1933), 'On the Tchebychef inequality of Bernstein', *The Annals of Mathematical Statistics* **4**(2), 94–102.

Cramér, H. (1938), 'Sur un nouveau théorème-limite de la théorie des probabilités', *Actualités Scientifiques* **736**.

Darling, D. A. and Robbins, H. (1967), 'Iterated logarithm inequalities', *Proceedings of the National Academy of Sciences* **57**(5), 1188–1192. MR0211441

de la Peña, V. H. (1999), 'A general class of exponential inequalities for martingales and ratios', *The Annals of Probability* **27**(1), 537–564. MR1681153

de la Peña, V. H. and Giné, E. (1999), *Decoupling*, Probability and its applications, Springer New York, New York, NY. MR1666908

de la Peña, V. H., Klass, M. J. and Lai, T. L. (2000), Moment bounds for self-normalized martingales, *in* 'High Dimensional Probability II', Birkhäuser, Boston, MA, pp. 3–11. MR1857311

de la Peña, V. H., Klass, M. J. and Lai, T. L. (2001), Self-normalized processes: exponential inequalities, moments, and limit theorems. Stanford University Technical Report No. 2001-6.

de la Peña, V. H., Klass, M. J. and Lai, T. L. (2004), 'Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws', *The Annals of Probability* **32**(3), 1902–1933. MR2073181

de la Peña, V. H., Klass, M. J. and Lai, T. L. (2007), 'Pseudo-maximization and self-normalized processes', *Probability Surveys* **4**, 172–192. MR2368950

de la Peña, V. H., Klass, M. J. and Lai, T. L. (2009), 'Theory and applications of multivariate self-normalized processes', *Stochastic Processes and their Applications* **119**(12), 4210–4227. MR2565565

de la Peña, V. H., Lai, T. L. and Shao, Q.-M. (2009), *Self-normalized processes: limit theory and statistical applications*, Springer, Berlin. MR2488094

Delyon, B. (2009), 'Exponential inequalities for sums of weakly dependent variables', *Electronic Journal of Probability* **14**, 752–779. MR2495559

Delyon, B. (2015), Exponential inequalities for dependent processes, Technical report.

Dembo, A. and Zeitouni, O. (2010), *Large deviations techniques and applications*, Springer, Berlin, Heidelberg. MR2571413

Doob, J. L. (1940), 'Regularity properties of certain families of chance variables', **47**(3), 455–486. MR0002052

Dubins, L. E. and Savage, L. J. (1965), 'A Tchebycheff-like inequality for stochastic processes', *Proceedings of the National Academy of Sciences* **53**(2), 274–275. MR0182042

Durrett, R. (2017), *Probability: theory and examples*, 5th edn. MR3930614

Efron, B. (1969), 'Student's t-test under symmetry conditions', *Journal of the American Statistical Association* **64**(328), 1278–1302. MR0251826

Fan, X., Grama, I. and Liu, Q. (2012), 'Hoeffding's inequality for supermartingales', *Stochastic Processes and their Applications* **122**(10), 3545–3559. MR2956116

Fan, X., Grama, I. and Liu, Q. (2015), 'Exponential inequalities for martingales with applications', *Electronic Journal of Probability* **20**(1), 1–22. MR3311214

Freedman, D. A. (1975), 'On tail probabilities for martingales', *The Annals of Probability* **3**(1), 100–118. MR0380971

Galambos, J. (1978), *The asymptotic theory of extreme order statistics*, Wiley. Google-Books-ID: O06aAAAAIAAJ. MR0489334

Gittens, A. and Tropp, J. A. (2011), 'Tail bounds for all eigenvalues of a sum of random matrices', *ACM Report 2014-02, Caltech* .

Godwin, H. J. (1955), 'On generalizations of Tchebychef's inequality', *Journal of the American Statistical Association* **50**(271), 923–945. MR0071657

Hoeffding, W. (1963), 'Probability inequalities for sums of bounded random variables', *Journal of the American Statistical Association* **58**(301), 13–30. MR0144363

Jorgensen, B. (1997), *The theory of dispersion models*, CRC Press. MR1462891

Kearns, M. and Saul, L. (1998), Large deviation methods for approximate probabilistic inference, *in* 'Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence', UAI'98, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 311–319.

Khan, R. A. (2009), 'L_p-version of the Dubins–Savage inequality and some exponential inequalities', *Journal of Theoretical Probability* **22**(2), 348. MR2501324

Koltchinskii, V. and Lounici, K. (2017), 'Concentration inequalities and moment bounds for sample covariance operators', *Bernoulli* **23**(1), 110–133. MR3556768

Lepingle, D. (1978), Sur le comportement asymptotique des martingales locales, *in* A. Dold, B. Eckmann, C. Dellacherie, P. A. Meyer and M. Weil, eds, 'Séminaire de Probabilités XII', Vol. 649, Springer, Berlin, Heidelberg, pp. 148–161. MR0520004

Lieb, E. H. (1973), 'Convex trace functions and the Wigner–Yanase–Dyson conjecture', *Advances in Mathematics* **11**, 267–288. MR0332080

Logan, B. F., Mallows, C. L., Rice, S. O. and Shepp, L. A. (1973), 'Limit distributions of self-normalized sums', *The Annals of Probability* **1**(5), 788–809. MR0362449

Mackey, L., Jordan, M. I., Chen, R. Y., Farrell, B. and Tropp, J. A. (2014), 'Matrix concentration inequalities via the method of exchangeable pairs', *The Annals of Probability* **42**(3), 906–945. MR3189061

Mazliak, L. and Shafer, G., eds (2009), *The splendors and miseries of martingales*, *Electronic Journal for History of Probability and Statistics* **5**(1). [Special issue]. `http://www.jehps.net/juin2009.html`. MR2520660

McDiarmid, C. (1998), Concentration, *in* M. Habib, C. McDiarmid, J. Ramirez-Alfonsin and B. Reed, eds, 'Probabilistic Methods for Algorithmic Discrete

Mathematics', Springer, New York, pp. 195–248. MR1678578
Minsker, S. (2017), 'On some extensions of Bernstein's inequality for self-adjoint operators', *Statistics and Probability Letters* **127**, 111–119. MR3648301
Oliveira, R. (2010a), 'The spectrum of random k-lifts of large graphs (with possibly large k)', *Journal of Combinatorics* **1**(3), 285–306. MR2799213
Oliveira, R. (2010b), 'Sums of random Hermitian matrices and an inequality by Rudelson', *Electronic Communications in Probability* **15**, 203–212. MR2653725
Papapantoleon, A. (2008), 'An introduction to Lévy processes with applications in finance', arXiv:0804.0482 .
Pinelis, I. (1992), An approach to inequalities for the distributions of infinite-dimensional martingales, *in* 'Probability in Banach Spaces, 8: Proceedings of the Eighth International Conference', Birkhäuser, Boston, MA, pp. 128–134. MR1227615
Pinelis, I. (1994), 'Optimum bounds for the distributions of martingales in Banach spaces', *The Annals of Probability* **22**(4), 1679–1706. MR1331198
Prokhorov, A. V. (1995), Bernstein inequality, *in* M. Hazewinkel, ed., 'Encyclopaedia of Mathematics', Vol. 1 of *Encyclopaedia of Mathematics*, Springer, Boston, MA, p. 365. MR1492593
Prokhorov, Y. V. (1959), 'An extremal problem in probability theory', *Theory of Probability & Its Applications* **4**(2), 201–203. MR0121857
Protter, P. E. (2005), *Stochastic integration and differential equations*, Springer Science & Business Media. MR2020294
Raginsky, M. and Sason, I. (2012), 'Concentration of measure inequalities in information theory, communications and coding (second edition)', arXiv:1212.4663 [cs, math].
Robbins, H. (1970), 'Statistical methods related to the law of the iterated logarithm', *The Annals of Mathematical Statistics* **41**(5), 1397–1409. MR0277063
Robbins, H. and Siegmund, D. (1970), 'Boundary crossing probabilities for the Wiener process and sample sums', *The Annals of Mathematical Statistics* **41**(5), 1410–1429. MR0277059
Rockafellar, R. T. (1970), *Convex analysis*, Princeton mathematical series, Princeton University Press, Princeton, N.J. MR0274683
Rudelson, M. (1999), 'Random vectors in the isotropic position', *Journal of Functional Analysis* **164**(1), 60–72. MR1694526
Shao, Q.-M. (1997), 'Self-normalized large deviations', *The Annals of Probability* **25**(1), 285–328. MR1428510
Shorack, G. R. and Wellner, J. A. (1986), *Empirical processes with applications to statistics*, Wiley, New York. MR0838963
Siegmund, D. (1985), *Sequential analysis*, Springer New York, New York, NY. MR0799155
Tropp, J. A. (2011), 'Freedman's inequality for matrix martingales', *Electronic Communications in Probability* **16**, 262–270. MR2802042
Tropp, J. A. (2012), 'User-friendly tail bounds for sums of random matrices', *Foundations of Computational Mathematics* **12**(4), 389–434. MR2946459
Tropp, J. A. (2015), 'An introduction to matrix concentration inequalities',

Foundations and Trends in Machine Learning **8**(1-2), 1–230. MR3311439

Uspensky, J. V. (1937), *Introduction to mathematical probability*, 1st edn., McGraw-Hill Book Company, Inc, New York, London.

van de Geer, S. (1995), 'Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes', *The Annals of Statistics* **23**(5), 1779–1801. MR1370307

Vershynin, R. (2012), Introduction to the non-asymptotic analysis of random matrices, *in* Y. C. Eldar and G. Kutyniok, eds, 'Compressed Sensing: Theory and Applications', Cambridge University Press, pp. 210–268. MR2963170

Ville, J. (1939), *Étude Critique de la Notion de Collectif*, Gauthier-Villars, Paris.

Wainwright, M. J. (2017), *High-dimensional statistics: a non-asymptotic viewpoint*, Cambridge University Press. MR3967104

Wald, A. (1945), 'Sequential tests of statistical hypotheses', *Annals of Mathematical Statistics* **16**(2), 117–186. MR0013275