Electronic Journal of Statistics
Vol. 16 (2022) 2406–2431
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS1995

Tail bounds for empirically standardized sums*

Guenther Walther

Department of Statistics,
Stanford University,
390 Jane Stanford Way, Stanford,
CA 94305
e-mail: gwalther@stanford.edu

Abstract: Exponential tail bounds for sums play an important role in statistics, but the example of the t-statistic shows that the exponential tail decay may be lost when population parameters need to be estimated from the data. However, it turns out that if Studentizing is accompanied by estimating the location parameter in a suitable way, then the t-statistic regains the exponential tail behavior. Motivated by this example, the paper analyzes other ways of empirically standardizing sums and establishes tail bounds that are sub-Gaussian or even closer to normal for the following settings: Standardization with Studentized contrasts for normal observations, standardization with the log likelihood ratio statistic for observations from an exponential family, and standardization via self-normalization for observations from a symmetric distribution with unknown center of symmetry. The latter standardization gives rise to a novel scan statistic for heteroscedastic data whose asymptotic power is analyzed in the case where the observations have a log-concave distribution.

MSC2020 subject classifications: Primary 62G32; secondary 60F10.
Keywords and phrases: Tail bounds, concentration inequality, t-statistic, Studentized contrast, likelihood ratio, self-normalization, scanning heteroscedastic data, moment bounds for log-concave distributions.

Received September 2021.

Contents

*Research supported by NSF grants DMS-1501767 and DMS-1916074

1. Introduction

Tail bounds and concentration inequalities for sums of independent random variables play a key role in statistics and machine learning, see e.g. van der Vaart and Wellner (1996), Boucheron et al. (2013), Vershynin (2018), or Wainwright (2019). Of particular importance are exponential tails bounds, which typically involve the expected value of the sum as well as a scale factor such as the variance. On the other hand, few results seem to be available when these parameters need to be estimated from the data, as may be required to make statistical methodology operational. The most prominent example is the t-statistic: If $X_1, \dots, X_m$ are i.i.d. $\mathrm{N}(\mu, \sigma^2)$, then

$$T \;:=\; \frac{\frac{1}{\sqrt{m}} \sum_{i=1}^{m} (X_i - \mu)}{\sqrt{\frac{1}{m-1} \sum_{i=1}^{m} \big(X_i - \overline{X}\big)^2}} \tag{1}$$

has the heavy algebraic tails of the t_{m-1}-distribution, so estimating σ^2 with the sample variance comes at the expense of losing the exponential tail decay. This paper explores the case where the expectation μ is also unknown and must be estimated. This is the typical setting for scan statistics, where observations in a scan window are assessed against an unknown baseline which is estimated with the sample mean of all observations, see e.g. Yao (1993). Corollary 1 below shows that, rather than exacerbating the situation, this additional estimation step actually restores the sub-Gaussian tail bound.

This result raises the question whether exponential tail bounds hold for other relevant ways of empirically (i.e. without using population parameters) standardizing sums. The answer turns out to be positive and this paper establishes tail bounds that are sub-Gaussian or even closer to normal for the following settings: Standardization by empirically centering and Studentizing sums of normal observations in Section 2, standardization with the log likelihood ratio statistic for observations from an exponential family in Section 3, and standardization via self-normalization for observations from a symmetric distribution with unknown center of symmetry in Section 4. The latter standardization give rise to a novel scan statistic for heteroscedastic data that is based on self-normalization, and its asymptotic power properties are also analyzed in Section 4. This analysis shows that the tail bounds are tight in the sense that they allow optimal detection in a certain scan problem; it is known that this optimality hinges on having the correct sub-Gaussian tail bound.

2. Normal tail bounds for Studentized constrasts and empirically centered sums

In order to derive a tail bound for empirically centered and Studentized sums it is convenient to establish a more general result about Studentized linear contrasts:

Theorem 1. *Let $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma^2)$ and $\mathbf{b} \in \mathbf{R}^n$ with $\sum_{i=1}^n b_i = 0$, $\sum_{i=1}^n b_i^2 = 1$. Then*

$$V := \frac{\sum_{i=1}^n b_i X_i}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2}}$$

is a pivot and satisfies a normal tail bound:

$$V \stackrel{d}{=} \frac{\sum_{i=1}^{n-1} Z_i}{\sqrt{\sum_{i=1}^{n-1} Z_i^2}} \qquad \text{for } Z_i \text{ i.i.d. } N(0,1),$$

$$\frac{V^2}{n-1} \sim \text{Beta}\left(\frac{1}{2}, \frac{n-2}{2}\right),$$

$$\mathbb{P}(V > t) \leq \mathbb{P}(N(0,1) > t) \quad \text{for } \begin{cases} t \geq 2.5 \text{ and } n \geq 10, & \text{or} \\ t \geq 2.75 \text{ and } n \geq 6, \end{cases}$$

and the analogous bound holds for the left tail of V.

In particular, Theorem 1 shows that the t-statistic regains the normal tail bound if the location parameter is estimated in a suitable way. This follows by setting $\mathbf{b} = \frac{\mathbf{c}}{\sqrt{\sum_i c_i^2}}$ with $c_i := 1 - \frac{m}{n}$ if $i \leq m$ and $c_i := -\frac{m}{n}$ otherwise, which implies $\sum_i c_i = 0$ and $\sum_i c_i^2 = m\left(1 - \frac{m}{n}\right)$:

Corollary 1. *Let $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma^2)$ and $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Then for $1 \leq m < n$:*

$$V := \frac{\frac{1}{\sqrt{m\left(1-\frac{m}{n}\right)}}\sum_{i=1}^m (X_i - \overline{X})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2}}$$

satisfies

$$\mathbb{P}(V > t) \;\leq\; \mathbb{P}(N(0,1) > t) \quad \text{for } \begin{cases} t \geq 2.5 \text{ and } n \geq 10, & \text{or} \\ t \geq 2.75 \text{ and } n \geq 6. \end{cases}$$

Studentization is a special case of self-normalization, see e.g. de la Peña et al. (2009) and Section 4. Self-normalization has certain advantages over standardizing with the population standard deviation because, roughly speaking, erratic fluctuations of the statistic are mirrored and therefore compensated by the random self-normalizing (Studentizing) term in the denominator, see Shao and Zhou (2016, 2017) for formal results. Corollary 1 shows that centering empirically rather than with the expected value can likewise be advantageous.

Remark: The algebraic tails of the t-distribution can be bounded by an exponential bound if the argument is small relative to the degrees of freedom, and this exponential tail bound may be useful for certain applications that do not require bounds far out in the tails. A referee pointed out the following more general example: If the X_i are symmetric about μ, then identity (1.1) in de la Peña et al. (2009) gives for T in (1):

$$\mathbb{P}(T > x) = \mathbb{P}\left(\frac{\sum_{i=1}^m (X_i - \mu)}{\sqrt{\sum_{i=1}^m (X_i - \mu)^2}} \geq \frac{\sqrt{m}x}{\sqrt{m - 1 + x^2}}\right) \leq \exp\left(-\frac{mx^2}{2(m - 1 + x^2)}\right)$$

where the inequality follows from (4). Hence T has a sub-Gaussian tail for $x = O(\sqrt{m})$. However, even for this restricted range of arguments this sub-Gaussian bound does not have the desired scale factor 1. For example, $x = \sqrt{m}$ yields the bound $\exp(-x^2/(2c))$ with $c = 2 - \frac{1}{m}$, so even for large m one obtains $c \approx 2$. The scale factor plays a key role in the theory and applications of sub-Gaussian tail bounds.

3. Sub-Gaussian tail bounds for the log likelihood ratio statistic

Let $X_1, \ldots, X_n$ be independent observations from a regular one-dimensional natural exponential family $\{f_\theta, \theta \in \Theta\}$, i.e. f_θ has a density with repect to some σ-finite measure ν which is of the form $f_\theta(x) = \exp(\theta x - A(\theta))\, h(x)$ and the natural parameter space $\Theta = \{\theta \in \mathbf{R} : \int \exp(\theta x) h(x) \nu(dx) < \infty\}$ is open.

In order to derive good finite sample tail bounds in this setting, it turns out that it is useful to standardize with the log likelihood ratio statistic rather than by centering and scaling. In more detail, let $1 \leq m < n$ and $\theta_0 \in \Theta$. Then the generalized log likelihood ratio statistic based on the observations $X_1, \ldots, X_m$ is

$$\begin{aligned} \log\mathrm{LR}_m(\theta_0) &= \log \frac{\sup_{\theta \in \Theta} \prod_{i=1}^m f_\theta(X_i)}{\prod_{i=1}^m f_{\theta_0}(X_i)} \\ &= \sup_{\theta \in \Theta} \left((\theta - \theta_0) \sum_{i=1}^m X_i - m\Big(A(\theta) - A(\theta_0)\Big) \right) \end{aligned} \tag{2}$$

The MLE $\hat{\theta}_m$ is defined as the argmax of (2) if the argmax exists. Note that $\log\mathrm{LR}_m(\theta_0)$ is always well defined whether $\hat{\theta}_m$ exists or not.

$\log\mathrm{LR}_m(\theta_0)$ represents a standardization of the sum $\sum_{i=1}^m X_i$ since by Wilk's theorem $2\log\mathrm{LR}_m(\theta_0)$ is asymptotically pivotal χ_1^2 if the population parameter is θ_0. The idea pursued in this section is that $\sqrt{2\log\mathrm{LR}_m(\theta_0)}$ is therefore approximately standard normal, and hence it might be possible to establish a *finite sample* sub-Gaussian tail bound. In the binomial case such a tail bound was indeed established by Rivera and Walther (2013), see also Harremoës (2016) for bounds when $m = 1$. This section first extends the binomial bound to the exponential family case and then addresses the case of empirical standardization where the typically unknown θ_0 is replaced by the MLE.

It should be pointed out that while the square root of the log likelihood ratio does not commonly appear in the current literature, it has a history as a statistic for inference in exponential families. Barndorff-Nielsen (1986) calls $\operatorname{sgn}(\hat{\theta}_m - \theta_0)\sqrt{2\log \mathrm{LR}_m(\theta_0)}$, as well as its empirically standardized counterpart below, the *signed likelihood ratio statistic*. Rivera and Walther (2013), Frick et al. (2014) and König et al. (2020) use this statistic for detection problems. An important advantage of working with this standardization is that it allows to make full use of the power of the Chernoff bound, as can be seen from the proof of Theorem 2(a). The resulting tail bound is therefore tighter than those obtained from various relaxations of the Chernoff bound such as the Hoeffding or Bennett bounds.

Usually θ_0 is not known. Then an empirical standardization is obtained with the MLE $\hat{\theta}_n$ substituted into the log likelihood ratio statistic for all the observations $X_1, \ldots, X_n$:

$$\begin{aligned} \log \mathrm{LR}_{m,n}(\hat{\theta}_n) &= \log \frac{\left(\sup_{\theta\in\Theta} \prod_{i=1}^m f_\theta(X_i)\right)\left(\sup_{\theta\in\Theta} \prod_{i=m+1}^n f_\theta(X_i)\right)}{\sup_{\theta\in\Theta} \prod_{i=1}^n f_\theta(X_i)} \\ &= \sup_{\theta\in\Theta}\left(\theta \sum_{i=1}^m X_i - mA(\theta)\right) + \sup_{\theta\in\Theta}\left(\theta \sum_{i=m+1}^n X_i - (n-m)A(\theta)\right) \\ &\quad - \sup_{\theta\in\Theta}\left(\theta \sum_{i=1}^n X_i - nA(\theta)\right). \end{aligned} \tag{3}$$

As an aside, this statistic can be interpreted as the generalized log likelihood ratio test statistic for testing a common θ against different θ for $X_1, \ldots, X_m$ and $X_{m+1}, \ldots, X_n$. The standardization V in Corollary 1 has the same interpretation. In fact, if f_θ is $\mathrm{N}(\theta, \sigma)$ with unknown mean θ and known σ, then one computes that $\sqrt{2\log \mathrm{LR}_{m,n}(\hat{\theta}_n)}$ equals V with the sample variance replaced by σ^2 in the definition of V.

As another example, if the X_i are Bernoulli with unknown parameter $p \in (0,1)$, then the natural parameter for the exponential family is $\theta = \log \frac{p}{1-p}$. One computes that $\log \mathrm{LR}_{m,n}(\hat{\theta}_n)$ equals

$$\begin{aligned} & m\left(\overline{X}_m \log \frac{\overline{X}_m}{\overline{X}} + (1-\overline{X}_m)\log \frac{1-\overline{X}_m}{1-\overline{X}}\right) \\ & \qquad + (n-m)\left(\overline{X}_{m^c} \log \frac{\overline{X}_{m^c}}{\overline{X}} + (1-\overline{X}_{m^c})\log \frac{1-\overline{X}_{m^c}}{1-\overline{X}}\right) \end{aligned}$$

where $\overline{X}_m := \frac{1}{m}\sum_{i=1}^m X_i$, $\overline{X}_{m^c} := \frac{1}{n-m}\sum_{i=m+1}^n X_i$ and $\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i$. This statistic was proposed as a scan statistic by Kulldorff (1997) and, despite its lengthy form, has been widely adopted for scanning problems in computer science and statistics, see e.g. Neill and Moore (2004a,b) and Walther (2010).

Theorem 2. *Let $X_1, \ldots, X_n$ be i.i.d. $f_{\theta_0} \in \{f_\theta, \theta \in \Theta\}$, a regular one-dimensional natural exponential family, and let $1 \leq m < n$. Then for $x > 0$:*

(a)

$$\mathbb{P}_{\theta_0}\left(\sqrt{2\,\mathrm{logLR}_m(\theta_0)} > x\right) \;\leq\; 2\exp\left(-\frac{x^2}{2}\right)$$

(b)

$$\mathbb{P}_{\theta_0}\left(\sqrt{2\,\mathrm{logLR}_{m,n}(\hat{\theta}_n)} > x\right) \;\leq\; \begin{cases} (4+2x^2)\exp\left(-\frac{x^2}{2}\right) \\ (4+2e)\exp\left(-\frac{x^2}{2}\right) \quad \text{if } x \leq (nC)^{1/6} \end{cases}$$

for a certain constant C.

The bounds can be divided by 2 if one considers the signed square-root for one-sided inference. The proof of (a) proceeds by inverting the Cramér-Chernoff tail bound as in Rivera and Walther (2013), where this technique is employed for the binomial case. The bounds in (b) do not quite match the bound in (a) and the author has not been able to establish the simple $2\exp(-x^2/2)$ bound for (b). Simulations suggest that in fact an even better bound holds which is closer to the standard normal bound, i.e. a bound that gains the factor $1/x$ on the sub-Gaussian bound as in (6). Establishing such a bound is a relevant open problem given its importance for scan statistics, see Walther and Perry (2019) and the references therein.

4. Tail bounds for self-normalized and empirically centered sums of symmetric random variables

The goal of this section is to extend the results for i.i.d. normal observations in Section 2 to a setting that allows heteroscedastic observations with not necessarily equal expected values. Clearly, some additional assumption is necessary. The methodology proposed below allows to treat the case of independent (not necessarily identically distributed) observations having symmetric distributions with unknown and possibly different centers of symmetry.

It is informative to recapitulate the short and well known argument for establishing a sub-Gaussian tail bound via self-normalization in the case where the center of symmetry is known to be zero, see e.g. de la Peña et al. (2009): If $X_1, \ldots, X_m$ are independent and symmetric about 0, then introduce i.i.d. Rademacher random variables $R_1, \ldots, R_m$, $\mathbb{P}(R_1 = 1) = \mathbb{P}(R_1 = -1) = \frac{1}{2}$, which are independent of the X_i. Then $X_i \overset{d}{=} R_i X_i$ and hence for $t > 0$:

$$\begin{aligned} \mathbb{P}\left(\frac{\sum_{i=1}^m X_i}{\sqrt{\sum_{i=1}^m X_i^2}} > t\right) &= \mathbb{P}\left(\frac{\sum_{i=1}^m R_i X_i}{\sqrt{\sum_{i=1}^m X_i^2}} > t\right) \\ &= \mathbb{E}\mathbb{P}\left(\frac{\sum_{i=1}^m R_i X_i}{\sqrt{\sum_{i=1}^m X_i^2}} > t \,\middle|\, X_1, \ldots, X_m\right) \leq \exp\left(-\frac{t^2}{2}\right) \end{aligned} \quad (4)$$

by Hoeffding's inequality. Hence the sub-Gaussian tail bound is inherited from the Rademacher sum. Sub-Gaussianity for self-normalized sums has been investigated in a number of papers. In the i.i.d. case, Giné et al. (1997) show that if the self-normalized sums are stochastically bounded (which always holds if the law of X_i is symmetric), then they are uniformly sub-Gaussian for some scale parameter. Also for the i.i.d. case, Shao (1999) established asymptotic Cramér-type large deviation results under the assumption of a finite third moment. For independent but not necessarily identically distributed X_i, Jing et al. (2003) establish a Cramér-type large deviation result under certain finite moment assumptions when $\mathbb{E}X_i = 0$. In the case where the distributions of the X_i are symmetric about 0, Efron (1969, pp. 1285–1288) suggested that it should be possible to lower the sub-Gaussian tail bound (4) to the normal tail $\mathbb{P}(\mathrm{N}(0,1) > t)$ in the usual hypothesis testing range $t > 1.65$, but Fig. 1 in Pinelis (2007) shows that the normal tail is too small by a factor of at least 1.2 for certain $t \in (2,3)$. However, recent remarkable results by Pinelis (2012) and Bentkus and Dzindzalieta (2015) show that the sub-Gaussian tail bound (4) for the Rademacher sum can be improved upon to a bound of the order $\frac{1}{t}\exp(-\frac{t^2}{2})$, namely to a multiple of $\mathbb{P}(\mathrm{N}(0,1) > t)$ where the multiple is at most 3.18 and is even close to 1 for large t. This tail bound will then translate to the sum $\sum_{i=1}^m X_i$ after self-normalization via the above argument. This makes the use of the self-normalization very attractive in this setting, cf. the remarks in Section 1.

The first aim of this section is to extend these results to the case where the center of symmetry is unknown and may vary between the X_i. At first glance, this would appear to be a hopeless undertaking since the above Rademacher argument depends crucially on the symmetry about zero. However, there are observations available outside the summation window $X_1, \ldots, X_m$ which can be used for an empirical standardization. The idea is to construct an empirical centering which eliminates the unknown center of symmetry from the symmetrization argument, or which at least results in certain bounds on the center of symmetry. The second step then is to show that these bounds still allow for nearly normal tails.

For simplicity of exposition it is assumed in the following that $n = mp$ for integers $m \geq 1$ and $p \geq 2$. If m is much smaller than n, as is typically the case for scan problems, then this can always be arranged by discarding a small fraction of the observations if necessary. The proposed empirical centering is given by a linear transformation $\widetilde{\mathbf{X}} = \mathbf{A}\mathbf{X}$, where the matrix $\mathbf{A}$ satisfies the conditions in Proposition 1. One example of such an empirical centering is

$$\widetilde{X}_i \;:=\; X_i - \frac{1}{p-1} \sum_{j=m+(i-1)(p-1)+1}^{m+i(p-1)} X_j, \qquad i = 1, \ldots, m \tag{5}$$

Corresponding to the linear tranformation $\mathbf{A}$ write $\widetilde{\mu}_i := \sum_{j=1}^n a_{ij}\mu_j$, where μ_j is the center of symmetry of X_j. Note that it is not assumed that the X_j have a finite expected value. In the following, the subscript I denotes averaging over the index set $I := \{1, \ldots, m\}$, so $\mu_I := \frac{1}{m}\sum_{i=1}^m \mu_i$ and $\mu_{I^c} := \frac{1}{n-m}\sum_{i=m+1}^n \mu_i$.

Proposition 1. *Let* $\mathbf{A}$ *be a* $m \times n$ *matrix that has* p *non-zero entries in each row and one non-zero entry in each column, and these entries are 1 in columns* $\{i : i \le m\}$ *and* $\frac{-1}{p-1}$ *in columns* $\{i : i > m\}$. [1]

Let $X_i,\ i = 1, \ldots, n$, *be independent and symmetric about* μ_i *(so the* $X_i - \mu_i$ *need not be identically distributed).*

(a) If $\widetilde{\mathbf{X}} = \mathbf{AX}$, *then the self-normalized sum of the* $\widetilde{X}_i$ *satisfies*

$$\frac{\sum_{i=1}^m \widetilde{X}_i}{\sqrt{\sum_{i=1}^m \widetilde{X}_i^2}} \;=\; \frac{n}{n-m} \frac{\sum_{i=1}^m (X_i - \overline{X})}{\sqrt{\mathbf{X}^T \mathbf{A}^T \mathbf{A} \mathbf{X}}} \;=:\; T_m$$

(b) If $\mu_I \le \mu_{I^c}$ *and* $\mu_i = \mu_I$ *for all* $i \le m$ *and* $\mu_i = \mu_{I^c}$ *for all* $i > m$, *then*

$$\mathbb{P}\left(T_m \ge t\right) \;\le\; \min\Big(3.18, g(t)\Big)\, \mathbb{P}\Big(\mathrm{N}(0,1) > t\Big) \tag{6}$$

for all $t > 0$, *where* $g(t) := 1 + \frac{14.11\phi(t)}{(9+t^2)(1-\Phi(t))} \to 1$ *as* $t \to \infty$.

(c) If $\mu_I \le \mu_{I^c}$ *and (7) or (8) hold, then the tail bound (6) holds for* $t \in (0, \sqrt{m}K)$ *for some* $K = K(v) > 0$.
Condition (7) requires that the $\widetilde{\mu}_i$ *don't vary much:*

$$\sum_{i=1}^m (\widetilde{\mu}_i - \widetilde{\mu}_I)^2 \le\; v \sum_{i=1}^m \widetilde{\mu}_i^2 \qquad \text{for some } v \in [0,1) \tag{7}$$

Condition (8) requires that the $\{\mu_i, i \le m\}$ *don't vary much and likewise for* $\{\mu_i, i > m\}$*:*

$$\left.\begin{array}{c} \dfrac{1}{m} \displaystyle\sum_{i=1}^m (\mu_i - \mu_I)^2 \\ \dfrac{1}{n-m} \displaystyle\sum_{i=m+1}^n (\mu_i - \mu_{I^c})^2 \end{array}\right\} \le v(\mu_I - \mu_{I^c})^2 \qquad \text{for some } v \ge 0. \tag{8}$$

(d) The analogous inequalities to (b) and (c) hold for the left tail of T_m *if* $\mu_I \ge \mu_{I^c}$.

The proof of Proposition 1 shows that the transformed $\widetilde{X}_i$ is symmetric about $\widetilde{\mu}_i$ which may not equal zero. Nevertheless, the self-normalized sum of the $\widetilde{X}_i$ satisfies the normal tail bound (6) if the μ_i satisfy the conditions given in (b) or (c). (b) is a standard assumption for testing against an elevated mean on I, see Yao (1993). Note that T_m is similar to the statistic V used in Corollary 1 for the homoscedastic case. Indeed, the proof of Theorem 1 shows that V is the self-normalized sum of $\mathbf{BX}$ for a certain $(n-1) \times n$ matrix $\mathbf{B}$.

[1] This uniquely determines A up to permutations of the columns $\{i : i \le m\}$ and permutations of the columns $\{i : i > m\}$.

4.1. Scanning heteroscedastic observations having symmetric log-concave distributions

As the statistic T_m appears to be new, it is incumbent to demonstrate its utility with an analysis of its power. To this end this section considers the scan problem where one observes independent X_i, $i = 1, \ldots, n$, where each X_i has a distribution that is symmetric about some μ_i and log-concave, i.e. X_i has a density of the form $f(x) = \exp \phi_i(x - \mu_i)$, where $\phi_i : \mathbf{R} \to [-\infty, \infty)$ is a concave function that is symmetric about 0. Special cases of log-concave distributions are the class of normal distributions, where ϕ_i is a quadratic, the class of Laplace distributions, where ϕ_i is piecewise linear, and more generally all gamma distributions with shape parameter ≥ 1, all Weibull distributions with exponent ≥ 1 and all beta distributions with both parameters ≥ 1. Log-concave distributions represent an attractive and useful nonparametric surrogate for the class of Gaussian distributions in a range of problems in inference and modeling, see e.g. the review papers of Walther (2009), Saumard and Wellner (2014) and Samworth (2018).

The goal of the scan problem under consideration here is to detect an elevated mean $\mu_I > \mu_{I^c}$ on some interval $I = (j, k]$. Both the starting point j and the length $|I| = k - j$ are unknown, likewise the μ_i and the distributions of the X_i, i.e. the functions ϕ_i, are unknown. Thus this is the general setting of Proposition 1 with the additional assumption of log-concavity. The log-concavity assumption allows to establish a result about the asymptotic detection power of the statistic T_m that is similar to the homoscedastic normal case.

T_m tests for an elevated mean on the interval $I = (0, m]$. It is straightforward to analyze a different interval $I = (j, k]$, e.g. by applying T_{k-j} to the rearranged data vector $(X_{j+1}, \ldots, X_n, X_1, \ldots, X_j)$. Denote this statistic by T_I. Analyzing all possible intervals $I \subset (0, n]$ gives rise to a multiple testing problem that is addressed by combining the corresponding T_I into a scan statistic. Walther and Perry (2019) analyze several ways for combining the T_I such that optimal inference is possible, e.g. the Bonferroni scan. The use of that scan requires the availability of a tail bound for the null distribution of T_I, such as (6). The Bonferroni scan and the normal tail bound (6) give T_I a critical value of the form $\sqrt{2 \log \frac{n}{|I|}} + \kappa_{n,I}(\alpha)$ with $\kappa_{n,I}(\alpha) = O(1)$, which follows as in the proof of Theorem 2 in Walther and Perry (2019). Thus (11) in the following theorem shows that the Bonferroni scan based on the T_I has asymptotic power 1 if the assumptions of the theorem are met. These assumptions are discussed following the statement of the theorem.

Theorem 3. *Let the X_i, $i = 1, \ldots, n$, be independent with a log-concave distribution that is symmetric about some μ_i. Set $\sigma_i^2 := \operatorname{Var} X_i$, $I := (0, m]$, let $\mathbf{A}$ be the linear transformation (5) and write $\widetilde{\sigma}_i^2 := \operatorname{Var} \widetilde{X}_i$. Assume the μ_i satisfy (7) or (8).*

If $\mu_I - \mu_{I^c} \geq \sqrt{\frac{(2+\epsilon_n)\sigma_I^2 R_I \log \frac{n}{|I|}}{|I|}}$ with $\epsilon_n \sqrt{\log \frac{n}{|I|}} \to \infty$, $|I| \geq (\log n)^2$ and

$R_I := \frac{\sum_{i\in I}\widetilde{\sigma}_i^2}{\sum_{i\in I}\sigma_i^2}$, *and if*

$$\frac{\sigma_j^2}{\sigma_I^2} \;\le\; S\sqrt{\max_{i\in I}(j-i)} \qquad \textit{for all } j\in\{1,\dots,n\} \textit{ and some } S>0, \tag{9}$$

then

$$R_I \;\le\; 1+2S\sqrt{\frac{|I|^2}{n}} \tag{10}$$

and

$$\mathbb{P}\left(T_I \;>\; \sqrt{2\log\frac{n}{|I|}+O(1)}\right) \;\to 1 \qquad (n\to\infty). \tag{11}$$

This result likewise holds for intervals $I=(j,j+m]$, $0\le j\le n-m$, by applying the theorem to $(X_{j+1},\dots,X_n,\ X_1,\dots,X_j)$.

In order to compare the power of this scan statistic to an optimal benchmark, we first consider the special case where $X_i\sim \mathrm{N}(\mu_i,\sigma^2)$. For this special case of homoscedastic normal observations it is known that there is a precise condition under which detection is possible with asymptotic power 1: $\mu_I-\mu_{I^c}\ge\sqrt{\frac{(2+\epsilon_n)\sigma^2\log\frac{n}{|I|}}{|I|}}$, provided that ϵ_n does not go to zero too quickly: $\epsilon_n\sqrt{\log\frac{n}{|I|}}\to\infty$. One the other hand, dedection is impossible if '$\sqrt{2+\epsilon_n}$' is replaced by '$\sqrt{2-\epsilon_n}$'. Hence $\sqrt{2}$ measures the difficulty of the detection problem, and the theory of that problem shows that it affects this difficulty as an *exponent*. This explains the efforts in the literature to approach $\sqrt{2}$ as fast as possible, and the rates $\sqrt{2\pm\epsilon_n}$ given above appear to be the currently best known rates. Attaining the factor $\sqrt{2}$ hinges on having the correct scale factor in the sub-Gaussian null distribution of the test statistic. References and summaries of these results are given in Walther and Perry (2019) and Walther (2022).

Theorem 3 shows that in the practically important range $|I|\le\sqrt{\frac{n}{\log n}}$, the Bonferroni scan based on the T_I does indeed have asymptotic power 1 if $\mu_I-\mu_{I^c}$ exceeds the above detection threshold for the homoscedastic normal case, since (10) gives $R_I=1+o(\epsilon_n)$ and $\sigma_I^2=\sigma^2$ by homoscedasticity. It is notable that this Bonferroni scan, which is designed to deal with heteroscedastic symmetric observations, allows optimal detection in the special case of homoscedastic normal data. In fact, Theorem 3 shows that it achieves the detection boundary for the homoscedastic normal case already provided only the σ_i, $i\in I$, are equal and the σ_i outside I don't grow too quickly, as required in (9).

If the data are heteroscedastic, then Theorem 3 requires that σ^2 needs to be replaced by $\sigma_I^2R_I=\frac{1}{|I|}\sum_{i\in I}\widetilde{\sigma}_i^2$ in the lower bound for $\mu_I-\mu_{I^c}$. It is beyond the scope of this paper to analyze whether this condition is optimal.

There appears to be not much literature about the scanning problem with heteroscedastic observations, presumably because it is difficult to derive appropriate methodology. For example, the recent work of Enikeeva et al. (2018)

considers the heteroscedastic Gaussian detection problem where σ is allowed to be different on I and I^c, but it is assumed that σ is constant and known on both I and on I^c. The finite-sample tail bound (6) holds without such a restriction and thus self-normalized statistics may prove to be quite useful for scanning problems.

The proof of Theorem 3 uses the following moment inequality for log-concave distributions, which may be of independent interest:

Proposition 2. *If X has a log-concave distribution that is symmetric about 0, then for all real numbers $r, s > 0$:*

$$\mathbb{E}|X|^s \;\le\; (\mathbb{E}|X|^r)^{\frac{s}{r}}\,\Gamma(s+1)(r+1)^{\frac{s}{r}}$$

If $0 < s < r$, then $\mathbb{E}|X|^s \le (\mathbb{E}|X|^r)^{\frac{s}{r}}$ by Jensen's inequality, without any assumptions on the law of X. The proposition shows that if the distribution is log-concave and symmetric, then it is possible to bound higher absolute moments in terms of lower absolute moments.

5. Proofs

5.1. Proof of Theorem 1

Write $\mathbf{X} = (X_1, \dots, X_n)^T$ and let $\mathbf{A}$ be an orthogonal $n \times n$ matrix with first row $(\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}})$. Then $\mathbf{Y} := \mathbf{A}\mathbf{X}$ is a vector of independent normal random variables with variance σ^2 and $\mathbb{E}Y_1 = \sqrt{n}\mu$, $\mathbb{E}Y_i = 0$, $i = 2, \dots, n$. Further $\sum_{i=1}^n (X_i - \overline{X})^2 = \sum_{i=1}^n X_i^2 - n\overline{X}^2 = \sum_{i=1}^n Y_i^2 - Y_1^2 = \sum_{i=2}^n Y_i^2$. Note that this is the same transformation that is commonly used in textbooks to derive the distribution of Student's t-statistic. In the latter case one is interested in $\sqrt{n}\,\overline{X} = Y_1$, which is independent of $\sum_{i=2}^n Y_i^2$. In contrast, the condition $\sum_{i=1}^n b_i = 0$ ensures that $\sum_{i=1}^n b_i X_i$ is a function of $(Y_2, \dots, Y_n)$ only:

$$V \;=\; \frac{\langle \mathbf{b}, \mathbf{X}\rangle}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n \big(X_i - \overline{X}\big)^2}} \;=\; \frac{\langle \mathbf{b}, \mathbf{A}^T\mathbf{Y}\rangle}{\sqrt{\frac{1}{n-1}\sum_{i=2}^n Y_i^2}} \;=\; \frac{\langle \mathbf{c}, \mathbf{Y}\rangle}{\sqrt{\frac{1}{n-1}\sum_{i=2}^n Y_i^2}}, \quad (12)$$

where $\mathbf{c} = \mathbf{A}\mathbf{b}$ has $c_1 = \sum_{i=1}^n \frac{1}{\sqrt{n}} b_i = 0$ and thus $\sum_{i=2}^n c_i^2 = \sum_{i=1}^n c_i^2 = \sum_{i=1}^n b_i^2 = 1$.

Set $U_i := \frac{Y_i}{\sqrt{\sum_{i=2}^n Y_i^2}}$, $i = 2, \dots, n$. Then $\mathbf{U} = (U_2, \dots, U_n)^T$ has the uniform distribution on the $(n-2)$-dimensional unit sphere in $\mathbf{R}^{n-1}$ since the Y_i are i.i.d. $\mathrm{N}(0, \sigma^2)$. Therefore $\sum_{i=2}^n w_i U_i$, the length of the projection of $\mathbf{U}$ onto a unit vector $\mathbf{w} = (w_2, \dots, w_n)^T$, has the same distribution for every unit vector $\mathbf{w}$.

Setting $\mathbf{w} = \left(\frac{1}{\sqrt{n-1}}, \ldots, \frac{1}{\sqrt{n-1}}\right)^T$ gives[2]

$$V \;=\; \sqrt{n-1}\sum_{i=2}^{n} c_i U_i \;\stackrel{d}{=}\; \sqrt{n-1}\sum_{i=2}^{n} w_i U_i \;=\; \frac{\sum_{i=2}^{n} Y_i}{\sqrt{\sum_{i=2}^{n} Y_i^2}} \;\stackrel{d}{=}\; \frac{\sum_{i=1}^{n-1} Z_i}{\sqrt{\sum_{i=1}^{n-1} Z_i^2}}$$

where the Z_i are i.i.d. $\mathrm{N}(0,1)$. Setting $\mathbf{w} = (1, 0, \ldots, 0)^T$ gives

$$V \;\stackrel{d}{=}\; \sqrt{n-1}\sum_{i=2}^{n} w_i U_i \;=\; \sqrt{n-1}\,\frac{Y_2}{\sqrt{\sum_{i=2}^{n} Y_i^2}} \;\stackrel{d}{=}\; \sqrt{n-1}\,\frac{Z_1}{\sqrt{\sum_{i=1}^{n-1} Z_i^2}},$$

so $\frac{V^2}{n-1} \sim \text{Beta}\left(\frac{1}{2}, \frac{n-2}{2}\right)$ follows from a well known fact about the beta distribution.

It is also known that the uniform distribution on the sphere in $\mathbf{R}^m$, $m := n-1$, gives U_2 the density $\frac{\Gamma(\frac{m}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{m-1}{2})}\left(1-u^2\right)^{\frac{m-3}{2}} 1(u \in (-1,1))$, hence $V \stackrel{d}{=} \sqrt{n-1}\,U_2$ has density

$$f_V(t) \;=\; \frac{1}{\sqrt{m}}\,\frac{\Gamma(\frac{m}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{m-1}{2})}\left(1-\frac{t^2}{m}\right)^{\frac{m-3}{2}} 1(-\sqrt{m} \le t \le \sqrt{m}).$$

The plan is to show that $f_V(t)$ is not larger than the standard normal density $\phi(t) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{t^2}{2})$ for t large enough. Clearly $f_V(t) \le \phi(t)$ for $t > \sqrt{m}$. For $t \in (0, \sqrt{m})$ one has $\Gamma(\frac{m}{2}) \le \Gamma(\frac{m-1}{2})\sqrt{\frac{m}{2}}$ for $m > 2$ by Gautschi's inequality, and $\log(1+x) \le x - \frac{x^2}{2}$ for $x \in (-1, 0)$:

$$\begin{aligned} f_V(t) &\le \frac{1}{\sqrt{2\pi}}\exp\left(\frac{m-3}{2}\log\left(1-\frac{t^2}{m}\right)\right) \\ &\le \frac{1}{\sqrt{2\pi}}\exp\left(\frac{m-3}{2}\left(-\frac{t^2}{m}-\frac{t^4}{2m^2}\right)\right) \\ &= \phi(t)\exp\left(\frac{3}{2m}t^2 - \frac{m-3}{4m^2}t^4\right) \\ &\le \phi(t) \qquad \text{for } \; t^2 \ge \frac{6m}{m-3} \end{aligned} \tag{13}$$

The condition is satisfied if e.g. $t \ge 3$ and $m = n-1 \ge 9$. Less conservative bounds obtain by employing higher order terms for bounding $\log(1+x)$. For example, $\log(1+x) \le x - \frac{x^2}{2} + \frac{x^3}{3}$ for $x = -\frac{t^2}{m} \in (-1, 0)$ yields

$$f_V(t) \;\le\; \phi(t)\exp\left(\frac{3}{2m}t^2 - \frac{m-3}{4m^2}t^4 - \frac{m-3}{6m^3}t^6\right).$$

[2] Alternatively, construct rows 2 to n of the orthogonal matrix $\mathbf{A}$ such that $\mathbf{A}\mathbf{b} = \mathbf{c} = \left(0, \sqrt{\frac{1}{n-1}}, \ldots, \sqrt{\frac{1}{n-1}}\right)^T$. Then (12) gives $V = \left(\sum_{i=2}^{n} Y_i\right)/\sqrt{\sum_{i=2}^{n} Y_i^2}$ without assuming that the X_i are normal. This also shows that V is a self-normalized sum. However, the Y_i may not be independent if the X_i are not normal.

Dividing the argument in the exponent by $\frac{m-3}{2m^2}t^2$ shows that the argument is non-positive if

$$\frac{3m}{m-3} - \frac{1}{2}t^2 - \frac{1}{3m}t^4 \;\leq\; 0$$

and this inequality holds for $t^2 \geq g(m) := \frac{3}{4}m\left(\sqrt{1+\frac{16}{m-3}}-1\right)$. One checks numerically that

$$\max_{m\in\{5,\ldots,8\}} g(m) \leq 2.75^2, \quad \max_{m\in\{9,\ldots,75\}} g(m) \leq 2.5^2. \tag{14}$$

Therefore $f_V(t) \leq \phi(t)$ follows for $t \geq 2.5$ and $m > 75$ from (13), for $t \geq 2.5$ and $9 \leq m \leq 75$ from (14), and for $t \geq 2.75$ and $m \geq 5$ from these results together with (14). The last claim of the theorem now obtains with $n = m-1$. □

5.2. Proof of Theorem 2

The proof of (a) proceeds by inverting the Cramér-Chernoff tail bound, as in Rivera and Walther (2013) for the binomial case. X_1 has moment generating function $\mathbb{E}_{\theta_0}\exp(tX_1) = \exp\left(A(\theta_0+t)-A(\theta_0)\right)$ for $\theta_0 + t \in \Theta$. Markov's inequality gives for $x > \mathbb{E}_{\theta_0}X_1$:

$$\begin{aligned}
\mathbb{P}_{\theta_0}\left(\frac{1}{m}\sum_{i=1}^m X_i > x\right) &\leq \inf_{t\geq 0}\frac{\mathbb{E}\exp(t\sum_{i=1}^m X_i)}{\exp(tmx)} \\
&\leq \exp\left\{-\sup_{t\geq 0, t+\theta_0\in\Theta}\Big(tmx - m\big(A(\theta_0+t)-A(\theta_0)\big)\Big)\right\} \\
&= \exp\left\{-\sup_{\theta\in\Theta:\,\theta\geq\theta_0} m\Big((\theta-\theta_0)x - \big(A(\theta)-A(\theta_0)\big)\Big)\right\} \\
&= \exp\left\{-\log\mathrm{LR}_m(x,\theta_0)\right\}
\end{aligned}$$

where $\log\mathrm{LR}_m(x,\theta_0) := \sup_{\theta\in\Theta} m\Big((\theta-\theta_0)x - (A(\theta)-A(\theta_0))\Big)$. This conclusion used the fact that the sup over $\{\theta\in\Theta:\,\theta\geq\theta_0\}$ equals the sup over $\{\theta\in\Theta\}$ since convexity of A yields

$$(\theta-\theta_0)x - (A(\theta)-A(\theta_0)) \;\leq\; (\theta-\theta_0)x - (\theta-\theta_0)A'(\theta_0) \tag{15}$$

and the RHS is negative if $\theta < \theta_0$ and $x > \mathbb{E}_{\theta_0}X_1 = A'(\theta_0)$. The following claim will be proved below:

The function $x \mapsto \log\mathrm{LR}_m(x,\theta_0)$ is continuous and strictly increasing on
$[\mathbb{E}_{\theta_0}X_1, \infty) \cap \mathcal{M}^0$ (16)

where $\mathcal{M}$ denotes the convex hull of the support of f_{θ_0}. Analogously one shows that for $x < \mathbb{E}_{\theta_0}X_1$:

$$\mathbb{P}_{\theta_0}\left(\frac{1}{m}\sum_{i=1}^m X_i < x\right) \;\leq\; \exp\big\{-\log\mathrm{LR}_m(x,\theta_0)\big\}$$

and $x \mapsto \log\mathrm{LR}_m(x, \theta_0)$ is continuous and strictly decreasing on $(-\infty, \mathbb{E}_{\theta_0} X_1] \cap \mathcal{M}^0$. Together with $\log\mathrm{LR}_m(\mathbb{E}_{\theta_0} X_1, \theta_0) = 0$, which follows from (15) and $\mathbb{E}_{\theta_0} X_1 = A'(\theta_0) \in \mathcal{M}^0$, one obtains

$$\mathbb{P}_{\theta_0}\left(\log\mathrm{LR}_m\Big(\frac{1}{m}\sum_{i=1}^m X_i, \theta_0\Big) > t\right) \;\le\; 2\exp(-t)$$

for $t > 0$ and claim (a) follows. It remains to prove (16). This follows from Lemma 6.7 in Brown (1986) or from a general result in convex analysis to the effect that the Legendre transform $\phi(x) := \sup_{\theta\in\Theta}(\theta x - A(\theta))$ satisfies $\phi'(x) = \arg\max_{\theta\in\Theta}(\theta x - A(\theta)) =: \theta(x)$ if $x \in \mathcal{M}^0$ (in which case the MLE $\theta(x)$ exists uniquely and is given by $\theta(x) = A'^{-1}(x)$ by exponential family theory) and $\phi''(x) = 1/A''(\theta(x)) = 1/\mathrm{Var}_{\theta(x)} X_1 > 0$ since the exponential family is minimal. Hence $\log\mathrm{LR}_m(x, \theta_0)$ is differentiable wrt $x \in \mathcal{M}^0$ and

$$\frac{d}{dx}\log\mathrm{LR}_m(x, \theta_0) \;=\; m\,\big(\theta(x) - \theta_0\big)$$

It was shown above that if $x > \mathbb{E}_{\theta_0} X_1$, then the maximizer $\theta(x)$ satisfies $\theta(x) \ge \theta_0$. Now (16) follows from $\frac{d}{dx}\theta(x) = \phi''(x) > 0$ for $x \in \mathcal{M}^0$. Part (a) of the theorem is proved.

As for part (b), by the definition (3)

$$\log\mathrm{LR}_{m,n}(\hat\theta_n) \;\le\; \log\mathrm{LR}_{m,n}(\theta_0) \;=\; \log\mathrm{LR}_I(\theta_0) + \log\mathrm{LR}_{I^c}(\theta_0) \qquad (17)$$

where $I := \{1, \ldots, m\}$ and $I^c := \{m+1, \ldots, n\}$ and for an index set J write

$$\log\mathrm{LR}_J(\theta_0) = \;\log\frac{\sup_{\theta\in\Theta}\prod_{i\in J} f_\theta(X_i)}{\prod_{i\in J} f_{\theta_0}(X_i)}$$

$$\overline{X}_J = \frac{1}{\#J}\sum_{i\in J} X_i \;\text{ and }\; \overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$$

So $\log\mathrm{LR}_I(\theta_0) = \log\mathrm{LR}_m(\theta_0)$. The proof of (a) established for $x > 0$:

$$\mathbb{P}_{\theta_0}\Big(\log\mathrm{LR}_I(\theta_0) > x,\, \overline{X}_I \ge \mathbb{E}_{\theta_0} X_1\Big) \;\le\; \exp(-x) \qquad (18)$$

and the same bound holds with $\overline{X}_I < \mathbb{E}_{\theta_0} X_1$ in place of $\overline{X}_I \ge \mathbb{E}_{\theta_0} X_1$ or with I^c in place of I. (18) shows that $\log\mathrm{LR}_I(\theta_0)\; 1(\overline{X}_I \ge \mathbb{E}_{\theta_0} X_1) \overset{d}{\le} E$, where $E \sim \mathrm{Exp}(1)$.

Since $\{X_i, i \in I\}$ and $\{X_i, i \in I^c\}$ are independent and stochastic order is preserved under convolution, one gets

$$\log\mathrm{LR}_I(\theta_0)\; 1(\overline{X}_I \ge \mathbb{E}_{\theta_0} X_1) + \log\mathrm{LR}_{I^c}(\theta_0)\; 1(\overline{X}_{I^c} \ge \mathbb{E}_{\theta_0} X_1) \;\overset{d}{\le}\; R \qquad (19)$$

where R has the Erlang distribution with density $te^{-t}1(t > 0)$. As (19) holds for all possible combinations of '$\ge$' and '$<$' in the indicator functions, the union

bound gives

$$\mathbb{P}_{\theta_0}\Big(\log\mathrm{LR}_I(\theta_0)+\log\mathrm{LR}_{I^c}(\theta_0) > x\Big) \le 4\,\mathbb{P}(R > x) = 4\int_x^\infty te^{-t}dt = 4(1+x)e^{-x}.$$

Now the first inequality in (b) follows with (17).

As for the second inequality, set $\alpha := \frac{\#I}{n} = \frac{m}{n}$. Then for $x > 0$:

$$\begin{aligned}\mathbb{P}_{\theta_0}\Big(\overline{X}_I - \overline{X}_{I^c} > x\Big) \;&=\; \mathbb{P}_{\theta_0}\left((1-\alpha)\sum_{i\in I} X_i - \alpha\sum_{i\notin I} X_i > \alpha(1-\alpha)nx\right)\\ &\le \inf_{t\ge 0}\frac{\mathbb{E}_{\theta_0}\exp\Big((1-\alpha)t\sum_{i\in I}X_i - \alpha t\sum_{i\notin I}X_i\Big)}{\exp\left(\alpha(1-\alpha)tnx\right)}\\ &= \exp\Big\{-\sup_{t\ge 0} n\Big(\alpha(1-\alpha)tx - \alpha\left[A(\theta_0+(1-\alpha)t) - A(\theta_0)\right]\\ &\qquad - (1-\alpha)\left[A(\theta_0-\alpha t) - A(\theta_0)\right]\Big)\Big\} \qquad (20)\end{aligned}$$

One way to proceed from here would be via a Taylor series approximation of A in order to derive an exponential tail bound for $\alpha(1-\alpha)\frac{(\overline{X}_I-\overline{X}_{I^c})^2}{2\sigma_0^2}$ and likewise approximate $\log\mathrm{LR}_{m,n}(\hat\theta_n)$ by this quantity. But these approximations will create notable slack in the tail bound, while the proof in (a) shows that tight bounds are possible by employing a statistic that conforms to the Cramér-Chernoff bound. To this end define for $x \ge 0$

$$\begin{aligned}&\widetilde{\log\mathrm{LR}}_n(x)\\ &:= \sup_{t\ge 0} n\Big(\alpha(1-\alpha)tx - \left[\alpha A(\theta_0+(1-\alpha)t) + (1-\alpha)A(\theta_0-\alpha t) - A(\theta_0)\right]\Big)\end{aligned}$$

Then (20) gives

$$\mathbb{P}_{\theta_0}\Big(\widetilde{\log\mathrm{LR}}_n\left(\overline{X}_I - \overline{X}_{I^c}\right) \ge x, \overline{X}_I - \overline{X}_{I^c} \ge 0\Big) \;\le\; \exp(-x) \qquad (21)$$

since $\widetilde{\log\mathrm{LR}}_n(\cdot)$ is strictly increasing with $\widetilde{\log\mathrm{LR}}_n(0) = 0$ by Jensen's inequality.

The goal now is to show that $\left|\widetilde{\log\mathrm{LR}}_n\left(\overline{X}_I - \overline{X}_{I^c}\right) - \log\mathrm{LR}_{m,n}(\hat\theta_n)\right|$ is small relative to $\widetilde{\log\mathrm{LR}}_n\left(\overline{X}_I - \overline{X}_{I^c}\right)$. This is done with the following Proposition 3, which gives a general result about the MLE in natural exponential families, and with Lemma 1. In order to motivate part (b) of the following proposition, recall that the exponential family $\{f_\theta(x), \theta \in \Theta\}$ can alternatively be parameterized by its mean value, and the mapping $\theta \mapsto \mathbb{E}_\theta X = A'(\theta)$ is a homeomorphism between Θ and $\mathcal{M}^0$, the interior of the convex hull of the support of f_{θ_0}, see e.g. Brown (1986). The MLE $\hat\theta$ is given by the solution of $A'(\hat\theta) = \overline{X}$ if it exists. It may fail to exist if $\overline{X}$ falls on the boundary of $\mathcal{M}$. For example, if a binomial(n,p) experiment results in n successes, then $\overline{X} = 1$, but in the

natural parametrization the supremum of the likelihood is approached as the natural parameter $\theta = \log \frac{p}{1-p} \to \infty$, so the MLE $\hat{\theta}$ does not exist. This issue usually becomes negligible in an asymptotic analysis of the MLE, but it has to be accounted for in a finite sample statement.

Proposition 3. *Let $X_1, \ldots, X_n$ be i.i.d. from a regular one-dimensional natural exponential family $\{f_\theta, \theta \in \Theta\}$. For $\theta_0 \in \Theta$ write $\mu_0 = \mathbb{E}_{\theta_0} X_1$, $\sigma_0^2 = \mathrm{Var}_{\theta_0} X_1$, and $\mathrm{logLR}_n(\theta_0)$ is defined in (2).*

(a) If the MLE $\hat{\theta}$ exists, then

$$\begin{aligned} n\Big(\frac{\overline{X} - \mu_0}{\sigma_0}\Big)^2 &\le 2\,\mathrm{logLR}_n(\theta_0)\ \frac{M}{\sigma_0^2} \\ n(\hat{\theta} - \theta_0)^2 \sigma_0^2 &\le 2\,\mathrm{logLR}_n(\theta_0)\ \frac{M\sigma_0^2}{m^2} \\ n(\overline{X} - \mu_0)(\hat{\theta} - \theta_0) &\le 2\,\mathrm{logLR}_n(\theta_0)\ \frac{M}{m} \end{aligned}$$

where $m = \min_{\theta \text{ between } \theta_0 \text{ and } \hat{\theta}} A''(\theta)$, $M = \max_{\theta \text{ between } \theta_0 \text{ and } \hat{\theta}} A''(\theta)$.

(b) Let $\delta > 0$ such that $[\theta_0 - \delta, \theta_0 + \delta] \in \Theta$ and set $d_\delta := \min_{\theta = \theta_0 \pm \delta} \big((\theta - \theta_0)A'(\theta) - (A(\theta) - A(\theta_0))\big)$. Then $d_\delta > 0$. If $\mathrm{logLR}_n(\theta_0) \le n d_\delta$, then the MLE $\hat{\theta}$ exists and satisfies $|\hat{\theta} - \theta_0| \le \delta$.

Proof of Proposition 3: As for part (a), Taylor's theorem gives for θ between θ_0 and $\hat{\theta}$:

$$A(\theta) - A(\theta_0) \ \le\ A'(\theta_0)(\theta - \theta_0) + \frac{M}{2}(\theta - \theta_0)^2$$

Therefore these θ satisfy

$$\frac{\mathrm{logLR}_n(\theta_0)}{n} \ \ge\ (\theta - \theta_0)\overline{X} - \big(A(\theta) - A(\theta_0)\big) \ \ge\ (\theta - \theta_0)(\overline{X} - \mu_0) - \frac{M}{2}(\theta - \theta_0)^2$$

as $A'(\theta_0) = \mu_0$. Setting $\theta := \theta_M := \theta_0 + \frac{\overline{X} - \mu_0}{M}$ one obtains

$$\frac{\mathrm{logLR}_n(\theta_0)}{n} \ \ge\ \frac{(\overline{X} - \mu_0)^2}{2M}$$

provided it can be shown that

$$\theta_M \text{ is between } \theta_0 \text{ and } \hat{\theta}. \tag{22}$$

To this end, define the functions

$$\begin{aligned} L(\theta) &:= (\theta - \theta_0)\overline{X} - \big(A(\theta) - A(\theta_0)\big) \\ g(\theta) &:= (\theta - \theta_0)\big(\overline{X} - \mu_0\big) - \frac{M}{2}(\theta - \theta_0)^2 \end{aligned}$$

Then $L'(\theta_0) = g'(\theta_0) = \overline{X} - \mu_0$ and $L''(\theta) \geq g''(\theta)$ for θ between θ_0 and $\hat{\theta}$. Hence one obtains $L'(\theta) \geq g'(\theta)$ for $\theta \in [\theta_0, \hat{\theta}]$ if $\hat{\theta} \geq \theta_0$, and $L'(\theta) \leq g'(\theta)$ for $\theta \in [\hat{\theta}, \theta_0]$ if $\hat{\theta} < \theta_0$.

Now consider the case $\hat{\theta} \geq \theta_0$. Since $A'(\hat{\theta}) = \overline{X}$ gives $L'(\hat{\theta}) = 0$, one gets $g'(\hat{\theta}) \leq 0$. Since θ_M is the maximizer of the quadratic function $g(\theta)$, $g'(\hat{\theta}) \leq 0$ implies $\theta_M \leq \hat{\theta}$.

$$\overline{X} - \mu_0 = A'(\hat{\theta}) - A'(\theta_0) = A''(\xi)(\hat{\theta} - \theta_0) \text{ for some } \xi \text{ between } \theta_0 \text{ and } \hat{\theta} \quad (23)$$

implies that $\overline{X} - \mu_0$ and $\hat{\theta} - \theta_0$ have the same sign, as $A'' > 0$. So $\overline{X} - \mu_0 \geq 0$, but then $g(\theta_0) = 0$, $g'(\theta_0) = \overline{X} - \mu_0 \geq 0$ and $g(\theta_M) = \frac{(\overline{X} - \mu_0)^2}{2M} \geq 0$ implies $\theta_0 \leq \theta_M$. (If $\overline{X} - \mu_0 = 0$, then the quadratic g has only one zero and $\theta_M = \theta_0$). This shows (22) in the case $\hat{\theta} \geq \theta_0$, the case $\hat{\theta} < \theta_0$ is analogous.

The second inequality in (a) follows from (23) which gives

$$(\hat{\theta} - \theta_0)^2 \sigma_0^2 \;\leq\; \frac{(\overline{X} - \mu_0)^2}{m^2} \sigma_0^2 \;\leq\; 2 \frac{\log \mathrm{LR}_n(\theta_0)}{n} \frac{M \sigma_0^2}{m^2}$$

as well as

$$(\overline{X} - \mu_0)(\hat{\theta} - \theta_0) \;\leq\; \frac{(\overline{X} - \mu_0)^2}{m} \;\leq\; 2 \frac{\log \mathrm{LR}_n(\theta_0)}{n} \frac{M}{m}.$$

As for part (b), the function $h(\theta) := (\theta - \theta_0)A'(\theta) - \big(A(\theta) - A(\theta_0)\big)$ is stricly decreasing for $\theta < \theta_0$ and strictly increasing for $\theta > \theta_0$ since $h'(\theta) = (\theta - \theta_0)A''(\theta)$. Further, $h(\theta_0) = 0$ and $\min_{\theta = \theta_0 \pm \delta} h(\theta) = d_\delta$. This shows that $d_\delta > 0$ and

$$\{\theta \in \Theta : \, h(\theta) \leq d_\delta\} \;=:\; [\theta_{low}, \theta_{up}] \;\subset\; [\theta_0 - \delta, \theta_0 + \delta]. \quad (24)$$

The motivation for defining h is that for each θ, $h(\theta)$ gives $\log \mathrm{LR}_n(\theta_0)/n$ when $\overline{X} = A'(\theta)$, with θ representing the argmax (i.e. the MLE). Indeed

$$h(\theta) \;=\; \sup_{t \in \Theta} \Big[(t - \theta_0)A'(\theta) - \big(A(t) - A(\theta_0)\big) \Big], \; \theta \in \Theta \quad (25)$$

as is readily seen by differentiating wrt t. To make clear the dependence of $\log \mathrm{LR}_n(\theta_0)$ on $\overline{X}$ we write similarly as before $\log \mathrm{LR}_n(\overline{X}, \theta_0) := \log \mathrm{LR}_n(\theta_0)$, i.e.

$$\frac{1}{n} \log \mathrm{LR}_n(x, \theta_0) \;=\; \sup_{t \in \Theta} \Big[(t - \theta_0)x - \big(A(t) - A(\theta_0)\big) \Big], \quad x \in \mathcal{M}. \quad (26)$$

This function is convex in x since it is the Legendre transform of the convex function $A(t)$ plus a linear function. Comparing (25) and (26) shows that

$$\frac{1}{n} \log \mathrm{LR}_n(x, \theta_0) \;=\; h(\theta) \qquad \text{with } x = A'(\theta),$$

so this identity holds for $\theta \in \Theta$ and $x \in \mathcal{M}^0$, with θ being the MLE when the mean is x. Therefore

$$\left\{ x \in \mathcal{M}^0 : \; \frac{1}{n} \log \mathrm{LR}_n(x, \theta_0) \leq d_\delta \right\} \;=\; \big[A'(\theta_{low}), A'(\theta_{up}) \big] \quad (27)$$

(recall that A' is strictly increasing and continuous). But this implies that a boundary point $x \in \mathrm{bd}\mathcal{M}$ cannot satisfy $\frac{1}{n}\log\mathrm{LR}_n(x,\theta_0) \le d_\delta$ because the function $x \mapsto \log\mathrm{LR}_n(x,\theta_0)$ is convex and hence $M_\delta := \{x \in \mathcal{M} : \frac{1}{n}\log\mathrm{LR}_n(x,\theta_0) \le d_\delta\}$ must be an interval. Together with (27) this shows that $M_\delta \in \mathcal{M}^0$ and so for every $x \in M_\delta$ the MLE exists and is given by $(A')^{-1}(x) \in [\theta_{low}, \theta_{up}] \subset [\theta_0 - \delta, \theta_0 + \delta]$. □

Lemma 1. *Let $T > 0$. If the MLEs $\hat\theta_I$ and $\hat\theta_{I^c}$ exist, then on the event $\{\overline{X}_I - \overline{X}_{I^c} \ge 0,\ \sqrt{\alpha}|\hat\theta_I - \theta_0| \le T,\ \sqrt{1-\alpha}|\hat\theta_{I^c} - \theta_0| \le T\}$:*

$$\log\mathrm{LR}_{m,n}(\hat\theta_n) \le \widetilde{\log\mathrm{LR}}_n\big(\overline{X}_I - \overline{X}_{I^c}\big) + \frac{5}{\sqrt{2\alpha(1-\alpha)}}\, nT^3 \max_{\theta:|\theta-\theta_0|\le \frac{3T}{\sqrt{2\alpha(1-\alpha)}}} |A'''(\theta)|$$

Proof of Lemma 1: In the case where the MLEs $\hat\theta_I$ and $\hat\theta_{I^c}$ exist, set $\widetilde\theta := \alpha\hat\theta_I + (1-\alpha)\hat\theta_{I^c}$. By definition (3):

$$\begin{aligned}
\log\mathrm{LR}_{m,n}(\hat\theta_n) &\le \log\mathrm{LR}_{m,n}(\widetilde\theta)\\
&= \alpha n\big(\hat\theta_I\overline{X}_I - A(\hat\theta_I)\big)\\
&\quad + (1-\alpha)n\big(\hat\theta_{I^c}\overline{X}_{I^c} - A(\hat\theta_{I^c})\big) - n\big(\widetilde\theta\overline{X} - A(\widetilde\theta)\big)\\
&= \alpha(1-\alpha)n\big(\hat\theta_I - \hat\theta_{I^c}\big)\big(\overline{X}_I - \overline{X}_{I^c}\big)\\
&\quad - n\big[\alpha A(\widetilde\theta + (1-\alpha)t) - (1-\alpha)A(\widetilde\theta - \alpha t) - A(\widetilde\theta)\big]\\
&\qquad \text{with } t := \hat\theta_I - \hat\theta_{I^c} \text{ and using } \overline{X} = \alpha\overline{X}_I + (1-\alpha)\overline{X}_{I^c}\\
&\le \widetilde{\log\mathrm{LR}}_n\big(\overline{X}_I - \overline{X}_{I^c}\big) + R\big(\hat\theta_I - \hat\theta_{I^c}, \widetilde\theta\big)
\end{aligned}$$

on $\{\overline{X}_I - \overline{X}_{I^c} \ge 0\}$, where

$$\begin{aligned}
R(t,\widetilde\theta) := n\Big[&\alpha A(\theta_0 + (1-\alpha)t) + (1-\alpha)A(\theta_0 - \alpha t) - A(\theta_0)\Big]\\
&- n\Big[\alpha A(\widetilde\theta + (1-\alpha)t) + (1-\alpha)A(\widetilde\theta - \alpha t) - A(\widetilde\theta)\Big].
\end{aligned}$$

The last inequality uses the fact that $\hat\theta_I - \hat\theta_{I^c}$ and $\overline{X}_I - \overline{X}_{I^c}$ have the same sign since $\overline{X}_I - \overline{X}_{I^c} = A'(\hat\theta_I) - A'(\hat\theta_{I^c})$ and $A'' > 0$.

Taylor's theorem gives for some ξ, τ between 0 and t:

$$\begin{aligned}
R(t,\widetilde\theta) &= \frac{1}{2}\alpha(1-\alpha)nt^2\Big[(1-\alpha)A''(\theta_0 + (1-\alpha)\xi) + \alpha A''(\theta_0 - \alpha\xi)\\
&\qquad - (1-\alpha)A''(\widetilde\theta + (1-\alpha)\tau) - \alpha A''(\widetilde\theta - \alpha\tau)\Big]\\
&\le \frac{1}{2}\alpha(1-\alpha)nt^2\Big[\max_{\theta:|\theta-\theta_0|\le|t|} A''(\theta) - \min_{\theta:|\theta-\widetilde\theta|\le|t|} A''(\theta)\Big]\\
&\le \frac{5}{\sqrt{2\alpha(1-\alpha)}} nT^3 \max_{\theta:|\theta-\theta_0|\le\frac{3T}{\sqrt{2\alpha(1-\alpha)}}} |A'''(\theta)|
\end{aligned}$$

since $|\widetilde{\theta}-\theta_0| \le \alpha|\hat\theta_I-\theta_0| + (1-\alpha)|\hat\theta_{I^c}-\theta_0| \le (\sqrt{\alpha}+\sqrt{1-\alpha})T \le \sqrt{2}T$ and $|t| = |\hat\theta_I - \hat\theta_{I^c}| \le \left(\sqrt{\frac{1}{\alpha}}+\sqrt{\frac{1}{1-\alpha}}\right)T \le \sqrt{\frac{2}{\alpha(1-\alpha)}}T$, so

$$\left\{\theta : \max\left(|\theta-\theta_0|, |\theta-\widetilde{\theta}|\right) \le |\hat\theta_I - \hat\theta_{I^c}|\right\} \subset \left\{\theta : |\theta-\theta_0| \le \frac{3T}{\sqrt{2\alpha(1-\alpha)}}\right\}$$

and

$$\max\left\{|\theta_1-\theta_2| :\ |\theta_1-\theta_0| \le |\hat\theta_I-\hat\theta_{I^c}|,\ |\theta_2-\widetilde{\theta}| \le |\hat\theta_I-\hat\theta_{I^c}|\right\} \le \frac{5T}{\sqrt{2\alpha(1-\alpha)}}. \quad \square$$

Now the proof of the theorem can be completed as follows: Let $\delta > 0$ such that $[\theta_0-\delta, \theta_0+\delta] \subset \Theta$. If $\mathrm{logLR}_I(\theta_0) \le x$ for some $x \in (0, \alpha(1-\alpha)nd_\delta)$, then part (b) of Proposition 3 implies that the MLE $\hat\theta_I$ exists and $|\hat\theta_I-\theta_0| \le \delta$. But then (a) of that Proposition implies that $\alpha n(\hat\theta_I-\theta_0)^2 \le 2x\frac{M}{m^2}$, where $M := \max_{\theta:|\theta-\theta_0|\le\delta} A''(\theta)$ and $m := \min_{\theta:|\theta-\theta_0|\le\delta} A''(\theta)$. Likewise, $\mathrm{logLR}_{I^c}(\theta_0) \le x$ implies $(1-\alpha)n(\hat\theta_{I^c}-\theta_0)^2 \le 2x\frac{M}{m^2}$, hence we can set $T := \sqrt{\frac{2xM}{nm^2}}$ in Lemma 1 to obtain on the event $\{\overline{X}_I - \overline{X}_{I^c} \ge 0,\ \mathrm{logLR}_I \le x,\ \mathrm{logLR}_{I^c}(\theta_0) \le x\}$:

$$\mathrm{logLR}_{m,n}(\hat\theta_n) \ \le\ \widetilde{\mathrm{logLR}}_n\big(\overline{X}_I - \overline{X}_{I^c}\big) + \sqrt{\frac{x^3}{n}}\,C$$

where $C := \frac{10}{m^3}\sqrt{\frac{M^3}{\alpha(1-\alpha)}}\max_{\theta:|\theta-\theta_0|\le\delta}|A'''(\theta)|$. So for $x \in \big(0, n\min\big(\alpha(1-\alpha)d_\delta, C^{-2}\big)\big)$:

$$\begin{aligned}
&\mathbb{P}_{\theta_0}\left(\mathrm{logLR}_{m,n}(\hat\theta_n) > x,\ \overline{X}_I - \overline{X}_{I^c} \ge 0\right)\\
&\quad\le \mathbb{P}_{\theta_0}\left(\widetilde{\mathrm{logLR}}_n\big(\overline{X}_I - \overline{X}_{I^c}\big) > x\left(1-\sqrt{\frac{x}{n}}C\right),\ \overline{X}_I - \overline{X}_{I^c} \ge 0\right)\\
&\qquad + \mathbb{P}_{\theta_0}\left(\mathrm{logLR}_I(\theta_0) > x\right) \ +\ \mathbb{P}_{\theta_0}\left(\mathrm{logLR}_{I^c}(\theta_0) > x\right)\\
&\quad\le \exp\left(-x\left(1-\sqrt{\frac{x}{n}}\,C\right)\right) + 2\exp(-x) \quad \text{by (21) and (a) of the theorem}\\
&\quad\le (2+e)\exp(-x) \qquad \text{if } x \le \left(nC^{-2}\right)^{1/3}.
\end{aligned}$$

The companion inequality with $\overline{X}_I - \overline{X}_{I^c} < 0$ obtains analogously. The claim for $\sqrt{2\,\mathrm{logLR}_{m,n}(\hat\theta_n)}$ follows for $\frac{1}{2}x^2 \le (nC^{-2})^{1/3}$, so one can use $8C^{-2}$ as the constant C in the statement of the theorem. $\square$

5.3. Proof of Proposition 1

The requirements for the matrix $\mathbf{A}$ imply that $\sum_{i=1}^m \widetilde{\mu}_i$ contains each μ_i, $i \le m$, exactly once with coefficient 1, and each μ_i, $i > m$, exactly once with coefficient

$\frac{-1}{p-1}$. Therefore

$$\widetilde{\mu}_I \;=\; \frac{1}{m}\sum_{i=1}^{m}\widetilde{\mu}_i \;=\; \frac{1}{m}\sum_{i=1}^{m}\mu_i - \frac{1}{m(p-1)}\sum_{i=m+1}^{n}\mu_i \;=\; \mu_I - \mu_{I^c} \tag{28}$$

Lemma 2. *If (7) or (8) hold, then*

$$\frac{m(\widetilde{\mu}_I)^2}{\sum_{i=1}^{m}\widetilde{\mu}_i^2} \;\geq\; \begin{cases} 1-v & \text{if (7) holds} \\ \frac{1}{4v+1} & \text{if (8) holds.} \end{cases}$$

Proof of Lemma 2: $\sum_{i=1}^{m}(\widetilde{\mu}_i-\widetilde{\mu}_I)^2 = \sum_{i=1}^{m}\widetilde{\mu}_i^2 - m(\widetilde{\mu}_I)^2$ since $\widetilde{\mu}_I = \frac{1}{m}\sum_{i=1}^{m}\widetilde{\mu}_i$. Hence (7) bounds the RHS by $v\sum_{i=1}^{m}\widetilde{\mu}_i^2$, while (8) will be shown to give the bound

$$\sum_{i=1}^{m}(\widetilde{\mu}_i - \widetilde{\mu}_I)^2 \;\leq\; 4mv(\widetilde{\mu}_I)^2 \tag{29}$$

so the claim follows in each case by collecting terms.

For simplicity of exposition (29) will be proved for the linear transformation $\mathbf{A}$ given by (5). The proof goes through in the same way for a general matrix A given in Proposition 1 by employing more cumbersome notation. Therefore $\widetilde{\mu}_i = \mu_i - \frac{1}{p-1}\sum_{j=m+(i-1)(p-1)+1}^{m+i(p-1)}\mu_j$ for $i = 1,\ldots,m$. Then it follows from (28) and Jensen's inequality that

$$\begin{aligned}
\sum_{i=1}^{m}(\widetilde{\mu}_i - \widetilde{\mu}_I)^2 &= 4\sum_{i=1}^{m}\left(\frac{\mu_i-\mu_I}{2} - \sum_{j=m+(i-1)(p-1)+1}^{m+i(p-1)}\frac{\mu_j-\mu_{I^c}}{2(p-1)}\right)^2 \\
&\qquad \text{as } \widetilde{\mu}_I = \mu_I - \mu_{I^c} \\
&\leq 4\sum_{i=1}^{m}\left(\frac{1}{2}(\mu_i-\mu_I)^2 + \frac{1}{2(p-1)}\sum_{j=m+(i-1)(p-1)+1}^{m+i(p-1)}(\mu_j-\mu_{I^c})^2\right) \\
&= 2\sum_{i=1}^{m}(\mu_i-\mu_I)^2 + \frac{2}{p-1}\sum_{j=m+1}^{n}(\mu_j-\mu_{I^c})^2 \\
&\leq 2mv(\widetilde{\mu}_I)^2 + \frac{2}{p-1}(n-m)v(\widetilde{\mu}_I)^2 \qquad \text{by (8)} \\
&= 4mv(\widetilde{\mu}_I)^2 \qquad \text{since } n-m = m(p-1). \qquad \square
\end{aligned}$$

As for proof of part (a) of the Proposition, by the construction of $\mathbf{A}$ the sum $\sum_{i=1}^{m}\widetilde{X}_i$ contains each X_i, $i \leq m$, exactly once with coefficient 1, and each X_i, $i > m$, exactly once with coefficient $\frac{-1}{p-1}$. Therefore

$$\sum_{i=1}^{m}\widetilde{X}_i = \sum_{i=1}^{m}X_i - \frac{1}{p-1}\sum_{i=m+1}^{n}X_i$$

$$= \frac{n}{n-m}\left(\sum_{i=1}^{m}\left(1-\frac{m}{n}\right)X_i - \frac{m}{n}\sum_{i=m+1}^{n}X_i\right) \qquad \text{since } n = mp$$
$$= \frac{n}{n-m}\sum_{i=1}^{m}(X_i - \overline{X}).$$

As for (b) and (c), since each column of $\mathbf{A}$ has only one non-zero entry, it follows that if $i_1 \neq i_2$, then $\widetilde{X}_{i_1}$ and $\widetilde{X}_{i_2}$ are functions of disjoint sets of X_j. Hence the $\widetilde{X}_1, \ldots, \widetilde{X}_m$ are independent. $X_j - \mu_j \stackrel{d}{=} \mu_j - X_j$ and the independence of the X_j yield

$$\mathbb{P}\left(\widetilde{X}_i - \widetilde{\mu}_i \leq t\right) = \mathbb{P}\left(\sum_{j=1}^{n} a_{ij}(X_j - \mu_j) \leq t\right) = \mathbb{P}\left(\sum_{j=1}^{n} a_{ij}(\mu_j - X_j) \leq t\right)$$
$$= \mathbb{P}\left(\sum_{j=1}^{n} a_{ij}(X_j - \mu_j) \geq -t\right) = \mathbb{P}\left(\widetilde{X}_i - \widetilde{\mu}_i \geq -t\right).$$

Hence $\widetilde{X}_i$ is symmetric about $\widetilde{\mu}_i$. Theorem 1.1 in Bentkus and Dzindzalieta (2015) gives the bound $\frac{\overline{\Phi}(t)}{4\overline{\Phi}(\sqrt{2})} \leq 3.18\overline{\Phi}(t)$ for the self-normalized Rademacher sum and Theorem 1.1 in Pinelis (2012) gives the bound $\overline{\Phi}(t) + \frac{14.11\phi(t)}{9+t^2}$. Hence the conditioning argument (4) yields

$$\mathbb{P}\left(\frac{\sum_{i=1}^{m}\left(\widetilde{X}_i - \widetilde{\mu}_i\right)}{\sqrt{\sum_{i=1}^{m}\left(\widetilde{X}_i - \widetilde{\mu}_i\right)^2}} > t\right) \leq \min\Big(3.18, g(t)\Big)\,\mathbb{P}\Big(\mathrm{N}(0,1) > t\Big) \tag{30}$$

for all $t > 0$, where $g(t) := 1 + \frac{14.11\phi(t)}{(9+t^2)(1-\Phi(t))} \to 1$ as $t \to \infty$.

Lemma 2 gives

$$\sqrt{\sum_{i=1}^{m}\widetilde{\mu}_i^2} \leq M\sqrt{m}\,|\widetilde{\mu}_I| \qquad \text{for some } M \geq 1. \tag{31}$$

Suppose $T_m = \frac{\sum_{i\leq m}\widetilde{X}_i}{\sqrt{\sum_{i\leq m}\widetilde{X}_i^2}} > t$ for some $t > 0$. Then $\sum_{i=1}^{m}(\widetilde{X}_i - \widetilde{\mu}_i) > 0$ since $\sum_{i=1}^{m}\widetilde{\mu}_i = m\widetilde{\mu}_I = m(\mu_I - \mu_{I^c}) \leq 0$ by (28). Hence Minkowski's inequality gives

$$\frac{\sum_{i=1}^{m}(\widetilde{X}_i - \widetilde{\mu}_i)}{\sqrt{\sum_{i=1}^{m}(\widetilde{X}_i - \widetilde{\mu}_i)^2}} \geq \frac{\sum_{i=1}^{m}\widetilde{X}_i - m\widetilde{\mu}_I}{\sqrt{\sum_{i=1}^{m}\widetilde{X}_i^2} + \sqrt{\sum_{i=1}^{m}\widetilde{\mu}_i^2}} \geq \frac{t\sqrt{\sum_{i=1}^{m}\widetilde{X}_i^2} + m\,|\widetilde{\mu}_I|}{\sqrt{\sum_{i=1}^{m}\widetilde{X}_i^2} + M\sqrt{m}|\widetilde{\mu}_I|} \geq t$$

if $\sqrt{m} \geq Mt$. Hence for $t \in (0, \sqrt{m}/M]$:

$$\mathbb{P}\left(T_m > t\right) \leq \mathbb{P}\left(\frac{\sum_{i=1}^{m}(\widetilde{X}_i - \widetilde{\mu}_i)}{\sqrt{\sum_{i=1}^{m}(\widetilde{X}_i - \widetilde{\mu}_i)^2}} > t\right)$$

and the last term satisfies (30), proving (c).

If the μ_i are constant for $i \le m$ and for $i > m$, then (8) holds with $v = 0$, so one can use $M = 1$ in (31). Then T_m satisfies the bound (6) for all positive t since $\mathbb{P}(T_m > t) = 0$ for $t > \sqrt{m}$ by Cauchy-Schwartz, proving (b). (d) is analogous. □

5.4. Proof of Theorem 3

On the event $E_n(I) := \left\{\sum_{i\le m} \widetilde{X}_i \ge 0\right\}$ Minkowski's inequality gives

$$T_I \;\ge\; \frac{\sum_{i\le m} \widetilde{\mu}_i \;+\; \sum_{i\le m}(\widetilde{X}_i - \widetilde{\mu}_i)}{\sqrt{\sum_{i\le m} \widetilde{\mu}_i^2} + \sqrt{\sum_{i\le m}(\widetilde{X}_i - \widetilde{\mu}_i)^2}}$$

By Lemma 2 there exists $M \ge 1$ such that

$$\sqrt{\sum_{i\le m} \widetilde{\mu}_i^2} \;\le\; M\sqrt{m}\,|\widetilde{\mu}_I| \;=\; M\sqrt{m}(\mu_I - \mu_{I^c}) \qquad \text{by (28)}$$

Set $Q_n(I) := \frac{\sum_{i\le m}(\widetilde{X}_i - \widetilde{\mu}_i)}{\sqrt{\sum_{i\le m} \widetilde{\mu}_i^2} + \sqrt{\sum_{i\le m}(\widetilde{X}_i - \widetilde{\mu}_i)^2}}$ and $F_n(I) := \left\{\sum_{i\le m}(\widetilde{X}_i - \widetilde{\mu}_i)^2 \le \sum_{i\le m} \widetilde{\sigma}_i^2 (1 + \frac{\epsilon_n}{4})\right\}$. Then on the event $E_n(I) \cap F_n(I)$:

$$\begin{aligned} T_I &\ge \frac{m(\mu_I - \mu_{I^c})}{M\sqrt{m}(\mu_I - \mu_{I^c}) + \sqrt{\sum_{i\le m} \widetilde{\sigma}_i^2 \left(1 + \frac{\epsilon_n}{4}\right)}} + Q_n(I) \qquad \text{by (28)} \\ &\ge \frac{m\,\mu_{\min}\left(\sum_{i\le m} \widetilde{\sigma}_i^2\right)^{-1/2}}{M\sqrt{m}\,\mu_{\min}\left(\sum_{i\le m} \widetilde{\sigma}_i^2\right)^{-1/2} + 1 + \frac{\epsilon_n}{8}} + Q_n(I) \end{aligned}$$

since $\mu_I - \mu_{I^c} \ge \mu_{\min} := \sqrt{\frac{(2+\epsilon_n)\sigma_I^2 R_I \log \frac{n}{m}}{m}}$ and the function $x \mapsto \frac{ax}{bx+c}$ with $a, b, c > 0$ is nondecreasing in $x > 0$. Now $m\,\mu_{\min}\left(\sum_{i\le m} \widetilde{\sigma}_i^2\right)^{-1/2} = \sqrt{(2+\epsilon_n)\log \frac{n}{m}}$ since $\sigma_I^2 R_I = m^{-1}\sum_{i\le m} \sigma_i^2 R_I = m^{-1}\sum_{i\le m} \widetilde{\sigma}_i^2$. Using $\frac{1}{1+y} \ge 1 - y$ for $y > 0$ one obtains on the event $E_n(I) \cap F_n(I)$:

$$\begin{aligned} T_I &\ge \sqrt{(2+\epsilon_n)\log \frac{n}{m}} \left(1 - \frac{\epsilon_n}{8} - M\sqrt{\frac{(2+\epsilon_n)\log \frac{n}{m}}{m}}\right) + Q_n(I) \\ &\ge \left(\sqrt{2} + \epsilon_n\left(\frac{1}{16} + o(1)\right)\right)\sqrt{\log \frac{n}{m}} + Q_n \end{aligned}$$

since $m \ge (\log n)^2$ and $(\log n)^{-1/2} = o(\epsilon_n)$.

Now

$$|Q_n(I)| \;\le\; \frac{\left|\sum_{i\le m}(\widetilde{X}_i - \mathbb{E}\widetilde{X}_i)\right|}{\sqrt{\sum_{i\le m}(\widetilde{X}_i - \mathbb{E}\widetilde{X}_i)^2}}$$

so both tails of $Q_n(I)$ satisfy the bound (6) by (30). Therefore $\mathbb{P}(T_I > \sqrt{2\log\frac{n}{m}} + O(1)) \to 1$ obtains (note that $\epsilon_n\sqrt{\log\frac{n}{m}} \to \infty$) once it is shown that $\mathbb{P}(E_n(I) \cap F_n(I)) \to 1$.

The proof of Proposition 1 shows that the $\widetilde{X}_i$ are are independent and symmetric about $\widetilde{\mu}_i$. Chebychev's inequality and $\sum_{i\le m}\widetilde{\mu}_i = m(\mu_I - \mu_{I^c})$ give

$$\begin{aligned}
\mathbb{P}(E_n(I)^c) &= \mathbb{P}\left(\sum_{i\le m}(\widetilde{X}_i - \widetilde{\mu}_i) < -\sum_{i\le m}\widetilde{\mu}_i\right) \;\le\; \frac{\sum_{i\le m}\widetilde{\sigma}_i^2}{m^2(\mu_I - \mu_{I^c})^2} \\
&\le \frac{\sum_{i\le m}\widetilde{\sigma}_i^2}{m(2+\epsilon_n)\sigma_I^2 R_m \log\frac{n}{m}} \;=\; \frac{1}{(2+\epsilon_n)\log\frac{n}{m}} \;\to 0 \\
\mathbb{P}(F_n(I)^c) &= \mathbb{P}\left(\sum_{i\le m}\left((\widetilde{X}_i - \widetilde{\mu}_i)^2 - \widetilde{\sigma}_i^2\right) > \sum_{i\le m}\widetilde{\sigma}_i^2\frac{\epsilon_n}{4}\right) \\
&\le 16\frac{\mathrm{Var}\left(\sum_{i\le m}(\widetilde{X}_i - \widetilde{\mu}_i)^2\right)}{\left(\sum_{i\le m}\widetilde{\sigma}_i^2\right)^2\epsilon_n^2} \;\le\; 16\frac{\sum_{i\le m}C(\widetilde{\sigma}_i^2)^2}{\left(\sum_{i\le m}\widetilde{\sigma}_i^2\right)^2\epsilon_n^2}
\end{aligned}$$

where $C := \Gamma(5)\,3^2 - 1$ obtains by setting $r = 2, s = 4$ in Proposition 2. This uses the fact that $\widetilde{X}_i - \widetilde{\mu}_i$ has a log-concave distribution since it is the sum of independent log-concave random variables, see e.g. Saumard and Wellner (2014). Now (9) implies for $i \le m$

$$\begin{aligned}
\widetilde{\sigma}_i^2 \;=\; \sigma_i^2 + \frac{1}{(p-1)^2}\sum_{j=m+(i-1)(p-1)+1}^{m+i(p-1)}\sigma_j^2 &\le \sigma_i^2 + \frac{1}{(p-1)^2}(p-1)S\sqrt{n}\sigma_I^2 \\
&\le \sigma_i^2 + 2S\sqrt{\frac{m}{p}}\sigma_I^2 \qquad (32)\\
&\le 3S\sqrt{m}\sigma_I^2 \qquad (33)
\end{aligned}$$

(33) yields

$$\mathbb{P}(F_n(I)^c) \le\; 16C\frac{3S\sqrt{m}\sigma_I^2}{\sum_{i\le m}\widetilde{\sigma}_i^2\epsilon_n^2} \;\le\; 48C\frac{S}{\sqrt{m}\epsilon_n^2} \to 0$$

since $m \ge (\log n)^2$ and $\epsilon_n\sqrt{\log n} \to \infty$.

Finally, (10) follows from (32). □

5.5. *Proof of Proposition 2*

The proof uses the following lemma repeatedly:

Lemma 3. *Let $w(x)$ be an integrable function on $(0,\infty)$ that does not change its sign from $+$ to $-$ as x increases from 0 to ∞. Then $\int_0^\infty w(x)\,dx = 0$ implies $\int_t^\infty w(x)\,dx \ge 0$ for all $t > 0$.*

The lemma obtains by observing that $\int_t^\infty w(x)\,dx < 0$ for some $t > 0$ implies $w(z) < 0$ for some $z > t$ as well as $\int_0^t w(x)\,dx = \int_0^\infty w(x)\,dx - \int_t^\infty w(x)\,dx > 0$, which implies $w(s) > 0$ for some $s \in (0, t)$, contradicting the assumption about the sign changes of w.

Since the density f of X is log-concave and symmetric about 0, it follows that f is non-increasing on $[0, \infty)$ and that $f_0 := f(0) > 0$. Hence

$$f(x) \begin{cases} \le u(x) := f_0\,\mathbf{1}\left(|x| \le \frac{1}{2f_0}\right) & \text{if } x \in \left(0, \frac{1}{2f_0}\right], \\ \ge u(x) & \text{if } x > \frac{1}{2f_0}. \end{cases}$$

Set $w(x) := f(x) - u(x)$. Then $\int_0^\infty w(x)\,dx = \frac{1}{2} - \frac{1}{2} = 0$ since both densities f and u are symmetric about 0. Hence Lemma 3 gives

$$\int_t^\infty f(x)dx \;\ge\; \int_t^\infty u(x)dx \qquad \text{for all } t > 0. \tag{34}$$

Let $U \sim \text{Unif}\left(-\frac{1}{2f_0}, \frac{1}{2f_0}\right)$. Then (34) yields for $s > 0$:

$$\begin{aligned} \mathbb{E}|X|^s = 2\mathbb{E}|X|^s\mathbf{1}(X > 0) \;&=\; 2\int_0^\infty \mathbb{P}(X > t^{\frac{1}{s}})dt \\ \ge 2\int_0^\infty \mathbb{P}(U > t^{\frac{1}{s}})dt \;&=\; \mathbb{E}|U|^s \;=\; 2f_0\int_0^{\frac{1}{2f_0}} u^s ds \;=\; \frac{(2f_0)^{-s}}{s+1} \end{aligned} \tag{35}$$

Let V have density $v(x) := f_0 \exp(-2f_0|x|)$. Since $\log f(x)$ is a concave function and $\log v(x)$ is linear on $[0, \infty)$, the function $g(x) := \log v(x) - \log f(x)$ is convex on $[0, \infty)$ and satsifies $g(0) = 0$. Hence $g(x)$ cannot change its sign from $+$ to $-$ as x increases from 0 to ∞, and this is therefore also true for $w(x) := v(x) - f(x)$. Again $\int_0^\infty w(x)\,dx = 0$ holds since both densities f and v are symmetric about 0, so Lemma 3 gives

$$\int_t^\infty f(x)dx \;\le\; \int_t^\infty v(x)dx \qquad \text{for all } t > 0.$$

Thus

$$\begin{aligned} \mathbb{E}|X|^s = \;& 2\int_0^\infty \mathbb{P}(X > t^{\frac{1}{s}})dt \le 2\int_0^\infty \mathbb{P}(V > t^{\frac{1}{s}})dt \\ &= \mathbb{E}|V|^s \;=\; 2f_0\int_0^\infty v^s \exp(-2f_0 v)dv \;=\; \Gamma(s+1)\,(2f_0)^{-s} \end{aligned}$$

Together with (35) this shows that

$$(s+1)^{-1} \;\le\; (2f_0)^s\,\mathbb{E}|X|^s \;\le\; \Gamma(s+1) \qquad \text{for all } s > 0.$$

Hence for any $r > 0$: $\mathbb{E}|X|^s \le \Gamma(s+1)\,((2f_0)^{-r})^{s/r} \le \Gamma(s+1)\,((r+1)\mathbb{E}|X|^r)^{s/r}$. □

Acknowledgments

The author would like to thank a referee for comments about an exponential tail bound for the t-statistic.

References

Barndorff-Nielsen, O.E. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika* **73**, 307–322. MR0855891

Bentkus, V.K. and Dzindzalieta, D. (2015). A tight Gaussian bound for weighted sums of Rademacher random variables. *Bernoulli* **21**, 1231–1237. MR3338662

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

Brown, L.D. (1986). *Fundamentals of Statistical Exponential Families*. Institute of Mathematical Statistics, Hayward, CA.

de la Peña, V.H., Lai, T.L. and Shao, Q.M. (2009). *Self-Normalized Processes: Theory and Statistical Applications*. Springer, Berlin. MR2488094

Enikeeva, F., Munk, A. and Werner, F. (2018). Bump detection in heterogeneous Gaussian regression. *Bernoulli* **24**, 1266–1306. MR3706794

Efron, B. (1969). Student's t-test under symmetry conditions. *J. Amer. Statist. Assoc.* **64**, 1278–1302. MR0251826

Frick, K., Munk, A. and Sieling, H. (2014). Multiscale change point inference. *J. R. Stat. Soc. Ser. B.* **76**, 495–580. MR3210728

Harremoës, P. (2016). Bounds on tail probabilities in exponential families. arXiv:1601.05179

Giné, E., Götze, F. and Mason, D. (1997). When is the Student t-statistic asymptotically normal? *Ann. Probab.* **25**, 1514-1531. MR1457629

Jing, B. Y., Shao, Q. M. and Wang, Q. Y. (2003). Self-normalized Cramér type large deviations for independent random variables. *Ann. Probab.* **31**, 2167–2215. MR2016616

König, C., Munk, A. and Werner, F. (2020). Multidimensional multiscale scanning in exponential families: Limit theory and statistical consequences. *Ann. Statist.* **48**, 655-678. MR4102671

Kulldorff, M. (1997). A spatial scan statistic. *Comm. Statist. Theory Methods* **26**, 1481–1496. MR1456844

Neill, D. and Moore, A. (2004a). A fast multi-resolution method for detection of significant spatial disease clusters. *Adv. Neural Inf. Process. Syst.* **10**, 651–658.

Neill, D. and Moore, A. (2004b). Rapid detection of significant spatial disease clusters. In *Proc. Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 256– 265. ACM, New York.

Pinelis, I. (2007). Toward the best constant factor for the Rademacher-Gaussian tail comparison. *ESAIM Probab. Stat.* **11**, 412–426. MR2339301

Pinelis, I. (2012). An asymptotically Gaussian bound on the Rademacher tails. *Electron. J. Probab.* **17**, 1–22. MR2924368

Rivera, C. and Walther, G. (2013). Optimal detection of a jump in the intensity of a Poisson process or in a density with likelihood ratio statistics. *Scand. J. Stat.* **40**, 752-769. MR3145116

Samworth, R.J. (2018). Recent progress in log-concave density estimation. *Statist. Sci.* **33**, 493-509. MR3881205

Saumard, A. and Wellner, J.A. (2014). Log-concavity and strong log-concavity: a review. *Statistics Surveys* **8**, 45-114. MR3290441

Shao, Q.-M. (1999). Cramér-type large deviation for Student's t statistic. *J. Theoret. Probab.* **12**, 387–398. MR1684750

Shao, Q. and Zhou, W. (2016). Cramér type moderate deviation theorems for self-normalized processes. *Bernoulli* **22**, 2029–2079. MR3498022

Shao, Q. and Zhou, W. (2017). Self-normalization: Taming a wild population in a heavy-tailed world. *Appl. Math. J. Chinese Univ.* **32**, 253–269. MR3694061

van der Vaart, A. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics*. Springer, New York. MR1385671

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge, UK. MR3837109

Wainwright, M.J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, Cambridge. MR3967104

Walther, G. (2009). Inference and modeling with log-concave distributions. *Statist. Sci.* **24**, 319–327. MR2757433

Walther, G. (2010). Optimal and fast detection of spatial clusters with scan statistics. *Ann. Statist.* **38**, 1010–1033. MR2604703

Walther, G. and Perry, A. (2019). Calibrating the scan statistic: finite sample performance vs. asymptotics. arXiv preprint arXiv:2008.06136.

Walther, G. (2022). Calbrating the scan statistic with size-dependent critical values: Heuristics, methodology and computation. In: Glaz, J, Koutras M.V. (eds) *Handbook of Scan Statistics*. Springer, New York, NY.

Yao, Q. (1993). Tests for change-points with epidemic alternatives. *Biometrika* **80**, 179–191. MR1225223