High-dimensional Location Estimation via Norm Concentration for Subgamma Vectors

Shivam Gupta [1] **Jasper C.H. Lee** [2] **Eric Price** [1]

Abstract

In location estimation, we are given n samples from a known distribution f shifted by an unknown translation λ, and want to estimate λ as precisely as possible. Asymptotically, the maximum likelihood estimate achieves the Cramér-Rao bound of error $N(0, \frac{1}{n\mathcal{I}})$, where $\mathcal{I}$ is the Fisher information of f. However, the n required for convergence depends on f, and may be arbitrarily large. We build on the theory using *smoothed* estimators to bound the error for finite n in terms of $\mathcal{I}_r$, the Fisher information of the r-smoothed distribution. As $n \to \infty$, $r \to 0$ at an explicit rate and this converges to the Cramér-Rao bound. We (1) improve the prior work for 1-dimensional f to converge for constant failure probability in addition to high probability, and (2) extend the theory to high-dimensional distributions. In the process, we prove a new bound on the norm of a high-dimensional random variable whose 1-dimensional projections are subgamma, which may be of independent interest.

1. Introduction

Location estimation—a variant of mean estimation—is a fundamental problem in parametric statistics. Suppose there is a translation-invariant model $f^\lambda(x) = f(x - \lambda)$ for some known distribution f over $\mathbb{R}^d$. The statistician receives n i.i.d. samples from f^λ for some arbitrarily chosen *true parameter* $\lambda \in \mathbb{R}^d$, and the goal is to estimate λ with high accuracy, succeeding with high probability over the samples.

In contrast to general mean estimation, which aims to estimate the mean under minimal assumptions on the distribution, here we know the exact shape of the distribution up to translation. Such additional information allows us to estimate λ to higher accuracy.

The classic "textbook" theory for location estimation, and indeed for parametric estimation in general, recommends using the *Maximum Likelihood Estimate* (MLE). The MLE enjoys asymptotic normality: if we fix a distribution f and take the number of samples n to infinity, the distribution of the MLE converges to the multivariate Gaussian $\mathcal{N}(\lambda, \frac{1}{n}\mathcal{I}^{-1})$, where $\mathcal{I}$ is the *Fisher information* matrix, defined by

$$\mathcal{I} = \mathop{\mathbb{E}}_{x\sim f}\left[(\nabla \log f(x))\,(\nabla \log f(x))^\top\right]$$

As a basic property, if we denote the covariance matrix of f by Σ, then we always have $\mathcal{I}^{-1} \preceq \Sigma$, implying that the asymptotic performance of the MLE is always at least as good as the sample mean, whose performance is controlled by the covariance Σ. Furthermore, the Crámer-Rao bound states that no unbiased location estimator can have covariance smaller than $\frac{1}{n}\mathcal{I}^{-1}$, and so the MLE has the best asymptotic performance of any unbiased estimator.

Even though the textbook theory is satisfying in that the Fisher information essentially captures the information-theoretic limits of location estimation, its predictions may be misleading in practice. Specifically, this is due to the asymptotic nature of the MLE performance guarantee: we need to take the number of samples n to infinity in order to achieve subgaussian estimation error. The asymptotic result may have arbitrarily bad dependence on n in terms of the model f. While bounds exist in terms of regularity properties of f (Miao, 2010; Spokoiny, 2011; Pinelis, 2017), these bounds are infinite for simple examples like the Laplace distribution. The research goal, therefore, is to establish a *finite-sample* theory of location estimation, which bounds the estimation error explicitly as a function of n, applies to every f, and ideally attains even optimal constants in the estimation error.

Recent work by Gupta et al. (2022) addressed this question in the special case of 1 dimension. They showed that, while the MLE can have bad finite-sample performance, it is possible to improve the behavior by a simple adaptation: add Gaussian noise of some appropriately chosen radius r, where r decreases with the number of samples, to both the

[1]The University of Texas at Austin [2]Department of Computer Sciences and Institute for Foundations of Data Science, University of Wisconsin-Madison. Correspondence to: Shivam Gupta <shivamgupta@utexas.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

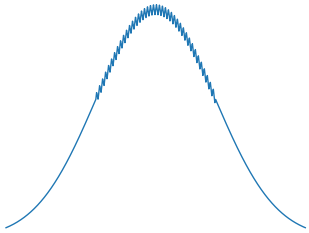


Figure 1. Gaussian+Sawtooth Distribution

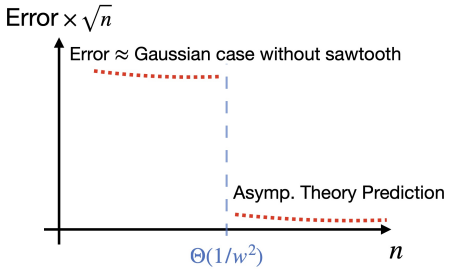


Figure 2. Constant probability error lower bound for Gaussian+Sawtooth

samples and model before performing MLE. Accordingly, the theoretical guarantees for the *smoothed* MLE replaces the Fisher information of f with the Fisher information of the smoothed distribution $f_r = f * \mathcal{N}(0, r^2)$, also called the smoothed Fisher information $\mathcal{I}_r$. Smoothed MLE achieves finite-sample subgaussian error bounds analogous to a Gaussian with variance $(1 + o(1))\mathcal{I}_r^{-1}$, where the $o(1)$ term can be explicitly calculated and is *independent* of f.

Characterization by smoothed Fisher information. Our results will follow the approach of Gupta et al. (2022) and show finite sample bounds in terms of the smoothed Fisher information. Here, focusing on the 1-dimensional case, we briefly discuss why Fisher information is inadequate and why smoothed Fisher information is a suitable substitute.

Consider the "Gaussian+Sawtooth" distribution shown in Figure 1, which is a sawtooth of tooth width w and slope $\pm\Delta$ added to the central section of the standard Gaussian density. As $w \to 0$, the density converges to the standard Gaussian, yet the Fisher information grows to $\Theta(\Delta^2)$ as $\Delta \to \infty$. The asymptotic theory thus predicts an error of $O(1/(\Delta\sqrt{n}))$ with constant probability.

However, Gupta et al. (2022) showed that for $n \ll 1/w^2$, the constant probability error for *every* algorithm is in fact at least $\Omega(1/\sqrt{n})$, as if the distribution were just a standard Gaussian. Intuitively, we need to align the model to within a single sawtooth width of w in order to leverage the sawtooth structure for high accuracy estimation. For a standard Gaussian, $\Omega(1/w^2)$ samples are needed for error less than w. Figure 2 shows a plot of the constant probability error lower bound for the Gaussian+Sawtooth model, with the error scaled by $\sqrt{n}$ for normalization.

Since the sample threshold depends on w, this example shows that there is no algorithm that converges to the asymptotic error in a distribution-independent way. Concretely, no algorithm can be within a $1 + o(1)$ factor of the $\mathcal{N}(0, 1/(n\mathcal{I}))$ error for a distribution-independent $o(1)$ term. We therefore need an alternative quantity to replace $\mathcal{I}$ for finite-sample error bounds, which can capture the phase transition in Figure 2.

Smoothed Fisher information exhibits this phase transition behavior. Smoothing by radius $r \gg w$ blurs out the sawtooth structure—$\mathcal{I}_r$ is small and close to the standard Gaussian Fisher information of 1. On the other hand, smoothing by radius $r \ll w$ preserves the sawtooth and keeps $\mathcal{I}_r$ close to $\mathcal{I} = \Theta(\Delta^2)$. Both Gupta et al. (2022) and we leverage this behavior to show finite sample bounds analogous to $(1 + o(1))\mathcal{N}(0, 1/(n\mathcal{I}_r))$, with a $o(1)$ term that is distribution-independent.

We need to choose the smoothing parameter *carefully*, as the smoothed Fisher information can depend delicately on r. Intuitively, we expect $r \to 0$ as $n \to \infty$; however, this is not true of Gupta et al.'s results. Their choice of smoothing vanishes only in the high-probability regime, i.e. when both $n \to \infty$ and $\delta \to 0$ for failure probability δ. Thus, for small constant δ, their results can be very sub-optimal. One of our new results removes the spurious dependence of r on δ.

Our results. In this paper, we improve and extend the result of Gupta et al. (2022) in two ways. First, we show that a variant of the algorithm has a simpler and better analysis in one dimension. This better analysis supports smaller smoothing radius r, and hence higher Fisher information $\mathcal{I}_r$:

Theorem 1.1 (1-d Smoothed MLE)**.** *Given a model f, let the r-smoothed Fisher information of a distribution f be $\mathcal{I}_r$, and let* IQR *be the interquartile range of f. Fix the failure probability be $\delta \le 0.5$, and assume that $n \ge c \cdot \log \frac{2}{\delta}$ for some sufficiently large constant c.*

*Choose $r^* = \Omega((\frac{\log \frac{2}{\delta}}{n})^{1/8})$*IQR*. Then, with probability at least $1 - \delta$, the output $\hat{\lambda}$ of Algorithm 2 satisfies*

$$|\hat{\lambda} - \lambda| \le \left(1 + O\left(\frac{\log \frac{2}{\delta}}{n}\right)^{\frac{1}{10}}\right)\sqrt{\frac{2\log \frac{2}{\delta}}{n\mathcal{I}_{r^*}}}$$

The main difference between this result and (Gupta et al., 2022) is the dependence on δ: the previous result needed $\delta \to 0$ for r to decay to 0 and for the leading constant to decay to 1. In ours, both decay polynomially in n for constant δ.

Consider how this result behaves on the Gaussian+Sawtooth example above (Figure 1), for constant δ. For small n, we will choose $r^* = \frac{1}{\text{poly}(n)} > w$ and get error within $1 + \frac{1}{\text{poly}(n)}$ of the regular Gaussian tail; for large n, $r^* \ll w$ and the error is within $1 + \frac{1}{\text{poly}(n)}$ of the asymptotically optimal $\mathcal{N}(0, 1/(n\mathcal{I}_r))$. Thus we get the same qualitative transition behavior as Figure 2, albeit at a different transition point ($\frac{1}{w^8}$ rather than $\frac{1}{w^2}$). The prior work (Gupta et al., 2022) additionally required vanishing δ, roughly $\delta < 2^{-\text{poly}(n)}$, to observe this behavior.

Second, our simpler approach lets us generalize the result to high dimensions. We show an analogous result to the one-dimensional result. In an ideal world, since the (unsmoothed) MLE satisfies $(\widehat{\lambda} - \lambda) \to \mathcal{N}(0, \frac{1}{n}\mathcal{I}^{-1})$ asymptotically, we would aim for the Gaussian tail error (Boucheron et al. (2013), Example 5.7)

$$\|\widehat{\lambda} - \lambda\|_2 \le \sqrt{\frac{\text{Tr}(\mathcal{I}^{-1})}{n}} + \sqrt{2\|\mathcal{I}^{-1}\|\frac{\log\frac{1}{\delta}}{n}} \qquad (1)$$

with probability $1 - \delta$. We show that this *almost* holds. Let $d_{\text{eff}}(A) = \frac{\text{Tr}(A)}{\|A\|}$ denote the effective dimension of a positive semidefinite matrix A. If we smooth by a spherical Gaussian $R = r^2 I_d$ for some $r^2 \le \|\Sigma\|$, then for a sufficiently large n as a function of $\|\Sigma\|/r^2, \log\frac{1}{\delta}$, $d_{\text{eff}}(\Sigma)$, and $d_{\text{eff}}(\mathcal{I}_R^{-1})$, our error is close to (1) replacing $\mathcal{I}$ with the smoothed Fisher information $\mathcal{I}_R$.

Theorem 1.2 (High-dimensional MLE, Informal; see Theorem B.16)**.** *Let f have covariance matrix Σ. For any $r^2 \le \|\Sigma\|$, let $R = r^2 I_d$ and $\mathcal{I}_R$ be the R-smoothed Fisher information of the distribution. For any constant $0 < \eta < 1$,*

$$\|\widehat{\lambda} - \lambda\|_2 \le (1+\eta)\sqrt{\frac{\text{Tr}(\mathcal{I}_R^{-1})}{n}} + 5\sqrt{\frac{\|\mathcal{I}_R^{-1}\|\log\frac{4}{\delta}}{n}}$$

with probability $1 - \delta$, for

$$n > O_\eta\left(\left(\frac{\|\Sigma\|}{r^2}\right)^2\left(\log\frac{2}{\delta} + d_{\mathit{eff}}(\mathcal{I}_R^{-1}) + \frac{d_{\mathit{eff}}(\Sigma)^2}{d_{\mathit{eff}}(\mathcal{I}_R^{-1})}\right)\right)$$

When $d_{\text{eff}}(\mathcal{I}_R^{-1}) \gg \log\frac{1}{\delta}$, the bound is $(1 + \eta + o(1))\sqrt{\text{Tr}(\mathcal{I}_R^{-1})}$. This is very close to the Cramer-Rao bound for the expected error of $\sqrt{\text{Tr}(\mathcal{I}^{-1})}$ for unbiased estimators (Bickel & Doksum (2015), Theorem 3.4.3).

The formal version of this theorem, Theorem B.16, also gives bounds for general distances $\|\widehat{\lambda} - \lambda\|_M$ induced by symmetric PSD matrices M; the exact bound, and the n required for convergence, depend on M.

One key piece of our proof, which may be of independent interest, is a concentration bound for the norm of a high-dimensional vector x with subgamma marginals in every direction. If a vector is Gaussian in every direction, it is a high-dimensional Gaussian and satisfies the tail bound (1) (replacing $\mathcal{I}^{-1}$ by the covariance matrix Σ). It was shown in (Hsu et al., 2012) that the same bound applies even if the marginals are merely *subgaussian* with parameter Σ. We extend this to get a bound for *subgamma* marginals:

Theorem 1.3 (Norm concentration for subgamma random vectors; see Theorem 5.1)**.** *Let x be a mean-zero random vector in $\mathbb{R}^d$ that is (Σ, C)-subgamma, i.e., it satisfies that for any vector $v \in \mathbb{R}^d$,*

$$\mathbb{E}[e^{\lambda\langle x, v\rangle}] \le e^{\lambda^2 v^T \Sigma v/2}$$

for $|\lambda| \le \frac{1}{\|Cv\|}$. Then with probability $1 - \delta$,

$$\|x\| \le \sqrt{\text{Tr}(\Sigma)} + 4\sqrt{\|\Sigma\|\log\frac{2}{\delta}} + 16\|C\|\log\frac{2}{\delta} + \min\left(4\|C\|_F\sqrt{\log\frac{2}{\delta}}, 8\frac{\|C\|_F^2}{\sqrt{\text{Tr}(\Sigma)}}\log\frac{1}{\delta}\right)$$

The first, trace term is the expected norm and the next two terms are (up to constants) the tight bound from 1-dimensional subgamma concentration. When x is an average of n samples, both Σ and C drop by a factor n; thus, the terms involving C decay at a rate of $1/n$, versus the terms involving only Σ, which decay at a rate of $1/\sqrt{n}$. As $n \to \infty$, the terms involving C disappear compared with the Gaussian terms involving Σ.

To better understand the last term, consider x to be the average of n samples X_i drawn from the spherical case ($\Sigma = \sigma^2 I, C = cI$). We also focus on the high-dimensional regime where $d \ge (2/\eta^2)\log(1/\delta)$ for some small η, where the target error bound of (1) becomes $(1+\eta)\sqrt{\text{tr}(\Sigma)/n}$, that is, within a $(1+\eta)$ factor of the expected ℓ_2 norm error. In the subgamma setting, the bound of Theorem 1.3 implies an error of $(1 + O(\eta))\sqrt{\text{Tr}(\Sigma)/n}$ whenever $n \gtrsim (c/\sigma)^2 d$, where the threshold for n is due to comparing the last "min" term in the bound with the $\sqrt{\|\Sigma\|\log\frac{2}{\delta}}$ term.

Under the stronger assumption that the random vectors have distance at most c from their expectation, one can compare our tail bound with Talagrand's/Bousquet's suprema concentration inequality (Boucheron et al. (2013), Theorem 12.5). Focusing again on the high-dimensional, spherical regime where $d \ge (2/\eta^2)\log(1/\delta)$ and $\Sigma = \sigma^2 I, C = cI$, Bousquet's inequality implies an almost-identical ℓ_2 error of $(1 + O(\eta))\sqrt{\text{Tr}(\Sigma)/n}$ whenever $n \gtrsim (c/\sigma)^2 d$, albeit with smaller hidden constant. Given that the n threshold for our bound is due to our last "min" term, it is likely that such a term is qualitatively necessary, and that our last term is not too large at least in the relevant regimes we consider in this paper.

Results by (Adamczak, 2008) and (van de Geer & Lederer, 2011) can also be used to bound the norm of random vectors.

Adamczak (2008) shows a high probability bound similar to (1) on $\sum_{i=1}^n X_i$ when $\|\max_i X_i\|$ is subexponential with a bounded Orlicz norm $\|\max_i X_i\|_{\psi_1}$. By contrast, our Theorem 1.3 assumes only that X_i is direction-by-direction subgamma instead of the *norm* being subexponential, and is independent of n. Moreover, even in the simplest case when the X_i are i.i.d. Gaussian, the Orlicz norm $\|\max_i X_i\|_{\psi_1}$ is as large as $\sqrt{\operatorname{Tr}(\Sigma)} + \sqrt{\|\Sigma\| \log n}$, and so (Adamczak, 2008) needs $n \gg \log^2 \frac{1}{\delta}$ to get close to the standard norm bound in (1). Theorem 1.3 on the other hand yields error $\sqrt{n \operatorname{Tr}(\Sigma)} + 4\sqrt{n\|\Sigma\| \log \frac{2}{\delta}}$ for any value of n, since a Gaussian is subgamma with $C = 0$. This almost completely recovers the standard Gaussian bound of (1), except for replacing the constant $\sqrt{2}$ by 4 in the second term.

van de Geer & Lederer (2011) implies a variant of Theorem 1.3 with three main differences: (a) it is designed for the spherical setting, (b) it loses a constant factor on the $\sqrt{\operatorname{Tr}(\Sigma)}$ term (and more if Σ is not spherical), and (c) it replaces our possibly-lossy terms involving $\|C\|_F$ with $\|C\|d$, which is incomparable in general, and worse if $d > \log \frac{1}{\delta}$. In the context of this work, the main issue is (b) since we aim for a $(1 + o(1))$-approximation to (1).

1.1. Notation

We denote the known distribution by f. In 1 dimension, f_r is the r-smoothed distribution $f * \mathcal{N}(0, r^2)$, with smoothed Fisher information $\mathcal{I}_r$. In high dimensions, f_R is the R-smoothed distribution $f * \mathcal{N}(0, R)$ with smoothed Fisher information $\mathcal{I}_R$—note the quadratic difference between r and R, analogous to the usual conventions for the (co)variance of 1-dimensional vs high-dimensional Gaussians.

The true parameter is denoted by λ. Both our 1-dimensional and high-dimensional algorithms first gets an initial estimate λ_1, before refining it into the final estimate $\hat{\lambda}$.

Unless otherwise specified, for a given vector x, $\|x\|$ denotes the ℓ_2 norm, and similarly $\|A\|$ is the operator norm of a square matrix A. Given a square positive semidefinite matrix A, we define its *effective dimension* to be $d_{\text{eff}}(A) = \operatorname{tr}(A)/\|A\|$. The effective dimension of a matrix A is d when it is spherical, but decays if one or more of its eigenvalues deviate from the maximum eigenvalue.

2. Related work

For an in-depth textbook treatment of the asymptotic theory of location estimation and parametric estimation in general, see (van der Vaart, 2000). There have also been finite-sample analysis of the MLE ((Spokoiny, 2011) in high dimensions, (Pinelis, 2017; Miao, 2010) in 1 dimension), but they require strong regularity conditions in addition to losing (at least) multiplicative constants in the estimation error bounds. Most related to this paper is the prior work of Gupta et al. (2022), which introduced smoothed MLE in the context of location estimation in 1 dimension, as well as formally analyzed its finite sample performance in terms of the smoothed Fisher information for large n and small δ.

There has been a flurry of work in recent years on the closely related problem of mean estimation, under the minimal assumption of finite (co)variance. The bounds then depend on this variance, rather than the Fisher information. In 1 dimension, the seminal paper of Catoni (2012) initiated the search for a subgaussian mean estimator with estimation error tight to within a $1 + o(1)$ factor; improvements by Devroye et al. (2016) and Lee and Valiant (2022a) have given a 1-dimensional mean estimator that works for all distributions with finite (but unknown) variance, with accuracy that is optimal to within a $1 + o(1)$ factor. Crucially, the $o(1)$ term is independent of the underlying distribution.

It remains an open problem to find a subgaussian mean estimator with tight constants under bounded covariance in high dimensions. A line of work (Lugosi & Mendelson, 2017; Hopkins, 2018; Cherapanamjeri et al., 2019) has shown how to achieve the subgaussian rate, ignoring constants, in polynomial time. More recently, Lee and Valiant (2022b) has achieved linear time and a sharp constant, but requires the effective dimension of the distribution to be much larger than $\log^2 \frac{1}{\delta}$.

Our other contribution is our novel norm concentration bound for subgamma random vectors. The norm concentration for Gaussian vectors has long been understood, see for example the textbook (Boucheron et al. (2013), Example 5.7). Hsu et al. (2012) generalized this bound to the case of direction-by-direction subgaussian vectors. Norm concentration can also be viewed as the supremum of an empirical process. Bousquet's version (2002; 2003) of Talagrand's suprema concentration inequality implies a norm concentration bound for random vectors bounded within an ℓ_2 ball of their expectation. Our bound generalizes this case of Bousquet's inequality from bounded vectors to all subgamma vectors. As discussed after Theorem 1.3, the results are quite similar for spherical Σ and C. Other results that can be used to bound the norm of random vectors are (Adamczak, 2008; van de Geer & Lederer, 2011). However, as discussed after Theorem 1.3, neither is adequate in our setting.

3. 1-dimensional location estimation

We discuss our 1-dimensional location estimation algorithm and its analysis at a high level in this section. See Appendix A for the complete analysis.

Algorithm 1 below is a *local* algorithm in the sense that it assumes we have an initial estimate λ_1 that is within some distance ϵ of λ, with the goal of refining the estimate to high accuracy.

Algorithm 1 Local smoothed MLE for one dimension

Input Parameters:

- Description of f, smoothing parameter r, samples $x_1, \ldots, x_n \overset{i.i.d.}{\sim} f^\lambda$ and initial estimate λ_1 of λ

1. Let $s(\hat{\lambda})$ be the score function of f_r, the r-smoothed version of f.
2. For each sample x_i, compute a perturbed sample $x_i' = x_i + \mathcal{N}(0, r^2)$ where all the Gaussian noise are drawn independently across all the samples.
3. Compute the empirical score at λ_1, namely $\hat{s}(\lambda_1) = \frac{1}{n}\sum_{i=1}^n s(x_i' - \lambda_1)$.
4. Return $\hat{\lambda} = \lambda_1 - (\hat{s}(\lambda_1)/\mathcal{I}_r)$.

Let $\mathcal{I}_r$ be the Fisher information of f_r, the r-smoothed version of f. Basic facts about the score $s(x)$ are:

$$0 = \mathop{\mathbb{E}}_{x\sim f_r}[s(x)]$$
$$\mathcal{I}_r = \mathop{\mathbb{E}}_{x\sim f_r}[-s'(x)] = \mathop{\mathbb{E}}_{x\sim f_r}[s(x)^2].$$

First, Algorithm 1 adds $N(0, r^2)$ perturbation independently to each x_i to get x_i', which are drawn as $(y_1 + \lambda, y_2 + \lambda, \ldots, y_n + \lambda)$ for $y_i \sim f_r$. It then computes

$$\widehat{s}(\lambda_1) := \frac{1}{n}\sum_{i=1}^n s(x_i' - \lambda_1) = \frac{1}{n}\sum_{i=1}^n s(y_i - \epsilon)$$

which is, in expectation,

$$\mathop{\mathbb{E}}_{x\sim f_r}[s(x-\epsilon)] \approx \mathop{\mathbb{E}}_{x\sim f_r}[s(x) - \epsilon s'(x)] = \epsilon\mathcal{I}_r.$$

Thus we expect $\widehat{\lambda} = \lambda_1 - \widehat{s}(\lambda_1)/\mathcal{I}_r \approx \lambda$.

There are two sources of error in this calculation: (I) the Taylor approximation to $s(x-\epsilon)$, and (II) the difference between the empirical and true expectations of $s(x-\epsilon)$. When $\epsilon = 0$, the Taylor error is 0 and the empirical estimator has variance

$$\frac{\mathrm{Var}(s(x))}{n} = \frac{\mathcal{I}_r}{n}.$$

Thus, when $\lambda_1 = \lambda$, $\widehat{\lambda}$ would be an unbiased estimator of λ with variance $\frac{1}{n\mathcal{I}_r}$: exactly the Cramér-Rao bound. Moreover, one can show that $s(x)$ is subgamma with variance proxy $\mathcal{I}_r$ and tail parameter $1/r$, giving tails on $\widehat{\lambda} - \lambda$ matching the $\frac{1}{n\mathcal{I}_r}$-variance Gaussian (up to some point depending on r). All we need to show, then, is that shifting by ϵ introduces little excess error in (I) and (II); intuitively, this happens for $|\epsilon| \ll r$ because f_r has been smoothed by radius r.

In fact, (Gupta et al., 2022) *already* bounded both errors: for (I), their Lemma C.2 shows that

$$\mathop{\mathbb{E}}_{x\sim f_r}[s(x-\epsilon)] = \mathcal{I}_r\epsilon \pm O(\sqrt{\mathcal{I}_r}\frac{\epsilon^2}{r^2}) \tag{2}$$

for all $|\epsilon| \le r/2$, and for (II), their Corollary 3.3 and Lemma C.3 together imply that a subgamma concentration of

$$|\hat{s}(\lambda_1) - \mathop{\mathbb{E}}_{x\sim f_r}[s(x-\epsilon)]| \lesssim$$
$$(1+o(1))\sqrt{\frac{\mathcal{I}_r \log\frac{2}{\delta}}{n}} + \frac{\log\frac{2}{\delta}}{nr} \tag{3}$$

when $r \gg |\epsilon|$.

Therefore, for sufficiently large r, the total error in $\widehat{s}(\lambda_1)$ is dominated by the leading $\sqrt{\frac{\mathcal{I}_r \log\frac{2}{\delta}}{n}}$ term, giving a result within $1 + o(1)$ of optimal.

Getting an initial estimate. We estimate λ by the empirical α-quantile of a small κ fraction of the samples, for some α; one can show that this has error at most $O(\mathrm{IQR}\cdot\sqrt{\frac{\log\frac{1}{\delta}}{\kappa n}})$ with $1-\delta$ probability, where IQR denotes the interquartile range. This strategy is essentially identical to (Gupta et al., 2022), except we use fresh samples for the two stages while they reuse samples.

Algorithm 2 Global smoothed MLE for one dimension

Input Parameters:

- Failure probability δ, description of f, n i.i.d. samples drawn from f^λ for some unknown λ

1. Let q be $\sqrt{2}(\log\frac{2}{\delta}/n)^{2/5}$.
2. Compute an $\alpha \in [q, 1-q]$ to minimize the width of interval defined by the $\alpha \pm q$ quantiles of f.
3. Take the sample α-quantile of the first $(\log\frac{1}{\delta}/n)^{1/10}$ fraction of the n samples.
4. Let $r^* = \Omega((\frac{\log\frac{1}{\delta}}{n})^{1/8})\mathrm{IQR}$.
5. Run Algorithm 1 on the rest of the samples, using initial estimate $\lambda_1 = x_\alpha$ and r^*-smoothing, and return the final estimate $\hat{\lambda}$.

Combining the above strategies and balancing parameters gives our final Algorithm 2. We prove in Appendix A that the algorithm gives our 1-dimensional result, Theorem 1.1.

Comparison to prior work. All the properties of the score function we need for this 1-dimensional result were shown in (Gupta et al., 2022), but that paper uses a different algorithm for which they could only prove a worse result. The (Gupta et al., 2022) algorithm looks for a root of $\widehat{s}$, while we essentially perform one step of Newton's method to approximate the root. General root finding requires *uniform* convergence of $\widehat{s}$, which (Gupta et al., 2022) could not prove without additional loss factors. By using one step, and (a small number of) fresh samples for the initial estimate, our algorithm only needs pointwise convergence.

4. High-dimensional location estimation

The high-dimensional case is conceptually analogous to the 1-d case. The complete analysis can be found in Appendix B. The main differences are: 1) The initial estimate comes from a heavy-tailed subgaussian estimator, and 2) We bound the difference between our estimate and the true mean using our concentration inequality for the norm of a subgamma vector (Theorem 5.1).

Let λ be the true location, and $\widehat{\lambda}$ our final estimate. We first state our main theorem, which gives a bound on $\|\widehat{\lambda} - \lambda\|_M$, induced by symmetric PSD matrices M.

Theorem 4.1 (High-dimensional MLE, Informal; see Theorem B.16)**.** *Let f have covariance matrix Σ. For any $r^2 \le \|\Sigma\|$, let $R = r^2 I_d$ and $\mathcal{I}_R$ be the R-smoothed Fisher information of the distribution. Let M be any symmetric PSD matrix, and let $T = M^{1/2}\mathcal{I}_R^{-1}M^{1/2}$. For any constant $0 < \eta < 1$,*

$$\|\widehat{\lambda} - \lambda\|_M \le (1+\eta)\sqrt{\frac{\mathrm{Tr}(T)}{n}} + 5\sqrt{\frac{\|T\|\log\frac{4}{\delta}}{n}}$$

with probability $1-\delta$, for

$$n > O_\eta\left(\left(\frac{\|\Sigma\|}{r^2}\right)^2\left(\log\frac{2}{\delta} + d_{\mathit{eff}}(T) + \frac{d_{\mathit{eff}}(\Sigma)^2}{d_{\mathit{eff}}(T)}\right)\right)$$

As a Corollary, we obtain Theorem 1.2 which bounds $\|\widehat{\lambda} - \lambda\|_2$, as well as the following, which bounds the Mahalanobis distance $\|\widehat{\lambda} - \lambda\|_{\mathcal{I}_R}$.

Corollary 4.2. *Let f have covariance matrix Σ. For any $r^2 \le \|\Sigma\|$, let $R = r^2 I_d$ and $\mathcal{I}_R$ be the R-smoothed Fisher information of the distribution. For any constant $0 < \eta < 1$,*

$$\|\widehat{\lambda} - \lambda\|_{\mathcal{I}_R} \le (1+\eta)\sqrt{\frac{d}{n}} + 5\sqrt{\frac{\log\frac{4}{\delta}}{n}}$$

with probability $1-\delta$, for

$$n > O_\eta\left(\left(\frac{\|\Sigma\|}{r^2}\right)^2\left(\log\frac{2}{\delta} + d + \frac{d_{\mathit{eff}}(\Sigma)^2}{d}\right)\right)$$

We now sketch our analysis. Algorithm 3 below takes an initial estimate λ_1 of the mean, and refines it to a precise estimate $\widehat{\lambda}$, analogously to Algorithm 1 for the 1-d case.

Algorithm 3 High-dimensional Local MLE

Input Parameters:

- Description of distribution f on $\mathbb{R}^d$, smoothing R, samples $x_1, \ldots, x_n \overset{i.i.d.}{\sim} f^\lambda$, and initial estimate λ_1

1. Let $\mathcal{I}_R$ be the Fisher information matrix of f_R, the R-smoothed version of f. Let s_R be the score function of f_R.
2. For each sample x_i, compute a perturbed sample $x_i' = x_i + \mathcal{N}(0, R)$ where all the Gaussian noise are drawn independently across all the samples.
3. Let $\hat{\epsilon} = \frac{1}{n}\sum_{i=1}^n \mathcal{I}_R^{-1} s_R(x_i' - \lambda_1)$ and return $\hat{\lambda} = \lambda_1 - \hat{\epsilon}$.

We discuss the runtime of Algorithm 3 in Appendix D.

Let f be a distribution on $\mathbb{R}^d$, and let $\mathcal{I}_R$ be the Fisher information matrix of f_R, the R-smoothed version of f. Then, for score s_R, if $\mathbf{J}_{s_R}$ is the Jacobian of s_R,

$$\mathcal{I}_R = \mathop{\mathbb{E}}_{x\sim f_R}\left[s_R(x)s_R(x)^T\right] = \mathop{\mathbb{E}}_{x\sim f_R}\left[-\mathbf{J}_{s_R}(x)\right]$$

Analogously to the 1-d case, Algorithm 3 takes an initial estimate $\lambda_1 = \lambda + \epsilon$ with $\epsilon^T R^{-1}\epsilon \le 1/4$. The algorithm first adds $N(0, R)$ independently to each sample x_i, to get x_i' which are drawn as $y_i + \lambda$ for $y_i \sim f_R$. Then, it computes

$$\hat{\epsilon} = \frac{1}{n}\sum_{i=1}^n \mathcal{I}_R^{-1} s_R(x_i' - \lambda_1) = \frac{1}{n}\sum_{i=1}^n \mathcal{I}_R^{-1} s_R(y_i - \epsilon)$$

which is in expectation

$$\mathop{\mathbb{E}}_{x\sim f_R}\left[\mathcal{I}_R^{-1} s_R(x-\epsilon)\right] \approx \mathop{\mathbb{E}}_{x\sim f_R}\left[-\mathcal{I}_R^{-1}\mathbf{J}_{s_R}(x)\epsilon\right] = \epsilon$$

So, again, we expect $\hat{\lambda} = \lambda_1 - \hat{\epsilon} \approx \lambda$ up to error from (I) the Taylor approximation to $s_R(x-\epsilon)$, and (II) the difference between the empirical and true expectations of $s_R(x-\epsilon)$.

For (I), Lemma B.3 shows that

$$\|\epsilon - \mathop{\mathbb{E}}_{x\sim f_R}\left[\mathcal{I}_R^{-1} s_R(x-\epsilon)\right]\|^2 \lesssim \|\mathcal{I}_R^{-1}\|(\epsilon^T R^{-1}\epsilon)$$

for $\epsilon^T R^{-1}\epsilon \le 1/4$. For (II), Corollary B.12 shows that for any unit direction v, $v^T\mathcal{I}_R^{-1}s_R(x-\epsilon)$ is subgamma:

$$v^T\mathcal{I}_R^{-1}s_R(x-\epsilon) \in \Gamma(\mathcal{I}_R^{-1}(1+o(1)), \mathcal{I}_R^{-1}R^{-1/2})$$

when $\epsilon^T R^{-1}\epsilon \le 1/4$ and $\sqrt{(\epsilon^T R^{-1}\epsilon)\log\left(\|I_R^{-1}\|\|R^{-1}\|\right)} \ll 1$, so that together

with our norm concentration inequality for subgamma vectors (Theorem 5.1), Lemma B.13 shows

$$\|\hat{\epsilon} - \mathop{\mathbb{E}}_{x\sim f_R}\left[\mathcal{I}_R^{-1} s_R(x-\epsilon)\right]\| \leq (1+o(1))\left(\sqrt{\frac{\mathrm{Tr}(\mathcal{I}_R^{-1})}{n}} + 4\sqrt{\frac{\|\mathcal{I}_R^{-1}\|\log\frac{2}{\delta}}{n}} + 16\frac{\|\mathcal{I}_R^{-1}R^{-1/2}\|\log\frac{2}{\delta}}{n} + 8\frac{\|\mathcal{I}_R^{-1}R^{-1/2}\|_F^2\log\frac{2}{\delta}}{n^{3/2}\sqrt{\mathrm{Tr}(\mathcal{I}_R^{-1})}}\right)$$

For $R = r^2 I_d$, when r is large, the total error is dominated by the first two terms in the above bound, which correspond to subgaussian concentration with covariance $\mathcal{I}_R^{-1}$.

Getting an initial estimate. For our initial estimate λ_1, we make use of a heavy-tailed estimator (Hopkins, 2018; Cherapanamjeri et al., 2019; Diakonikolas et al., 2020), which guarantee subgaussian error dependent on the covariance Σ of f, up to constants.

As in the 1-d case, combining our initial estimate with Algorithm 3 gives our final theorem, Theorem B.16. Below, Algorithm 4 shows how to compute our initial estimate and combine it with the local MLE Algorithm 3 to obtain our final estimate.

Algorithm 4 High-dimensional Global MLE

Input Parameters:

- Failure probability δ, description of distribution f, n samples from f^λ, Smoothing R, Approximation parameter η

1. Let Σ be the covariance matrix of f. Compute an initial estimate λ_1 using the first η/C fraction of of the n samples for large constant C, using an estimator from Theorem B.15.
2. Run Algorithm 3 using the remaining $1-\eta/C$ fraction of samples using R-smoothing and our initial estimate λ_1, returning the final estimate $\hat{\lambda}$.

5. Norm concentration for subgamma vectors

Theorem 5.1 (Norm concentration for subgamma vectors)**.** *Let x be a mean-zero random vector in $\mathbb{R}^d$ that is (Σ, C)-subgamma, i.e., for all $v \in \mathbb{R}^d$, $v^T x \in \Gamma(v^T\Sigma v, \|Cv\|)$. In other words, it satisfies that for any vector $v \in \mathbb{R}^d$,*

$$\mathbb{E}[e^{\lambda\langle x,v\rangle}] \leq e^{\lambda^2 v^T\Sigma v/2}$$

for $|\lambda| \leq \frac{1}{\|Cv\|}$. Let $\gamma > 0$. Then,

$$\mathbb{P}\left[\|x\| \geq \sqrt{\mathrm{Tr}(\Sigma)} + t\right] \leq 2e^{-\frac{1}{16}\min(\frac{t^2}{\|\Sigma\|}, \frac{t}{\|C\|}, \frac{2t\sqrt{\mathrm{Tr}(\Sigma)}+t^2}{\|C\|_F^2})}.$$

Thus, with probability $1-\delta$,

$$\|x\| \leq \sqrt{\mathrm{Tr}(\Sigma)} + 4\sqrt{\|\Sigma\|\log\frac{2}{\delta}} + 16\|C\|\log\frac{2}{\delta} + \min\left(4\|C\|_F\sqrt{\log\frac{2}{\delta}}, 8\frac{\|C\|_F^2}{\sqrt{\mathrm{Tr}(\Sigma)}}\log\frac{2}{\delta}\right)$$

The proof idea, similar to (Hsu et al., 2012) for the subgaussian case, is as follows. Define $v \sim N(0, I)$. We relate $\mathbb{P}[\|x\| > t]$ to the MGF $\mathbb{E}_x[e^{\lambda^2\|x\|^2}]$, which equals $\mathbb{E}_{x,v}[e^{\lambda\langle x,v\rangle}]$. If we interchange the order of expectation, as long as $\|Cv\| \leq 1/|\lambda|$, this is at most $\mathbb{E}_v[e^{\lambda^2 v^T\Sigma v}]$. Since v is Gaussian, we can compute the last MGF precisely.

To handle the subgamma setting, we need a way to control $\mathbb{E}_{x,v}[e^{\lambda\langle x,v\rangle}]$ over those v with $\|Cv\| > 1/|\lambda|$. We do so by showing that (I) WLOG $\|x\|$ is never strictly larger than the bound we want to show, and (II) then the contribution to the expectation from such cases is small.

Proof. Define $\gamma = \frac{t}{\sqrt{\mathrm{Tr}(\Sigma)}}$, so we want to bound $\mathbb{P}[\|x\| \geq (1+\gamma)\sqrt{\mathrm{Tr}(\Sigma)}]$. We start by showing that WLOG $\|x\|$ never *strictly* exceeds this threshold.

Introducing a bounded norm assumption. We first show that, without loss of generality, we can assume $\|x\| \leq (1+\gamma)\sqrt{\mathrm{Tr}(\Sigma)}$ always. Let $s \in \{\pm 1\}$ be distributed uniformly independent of x, and define

$$y = s \cdot x \cdot \min\left(1, \frac{(1+\gamma)\sqrt{\mathrm{Tr}(\Sigma)}}{\|x\|}\right).$$

to clip x's norm and symmetrize. For any v and x,

$$\mathop{\mathbb{E}}_s[e^{\lambda\langle y,v\rangle}] = \cosh\left(\lambda\langle x,v\rangle \cdot \min\left(1, \frac{(1+\gamma)\sqrt{\mathrm{Tr}(\Sigma)}}{\|x\|}\right)\right) \leq \cosh(\lambda\langle x,v\rangle)$$

Now, since x is (Σ, C)-subgamma,

$$\begin{aligned}\mathop{\mathbb{E}}_x[\cosh(\lambda\langle x,v\rangle)] &= \frac{1}{2}\left(\mathop{\mathbb{E}}_x[e^{\lambda\langle x,v\rangle}] + \mathop{\mathbb{E}}_x[e^{\lambda\langle x,-v\rangle}]\right) \\ &\leq \frac{1}{2}\left(e^{\lambda^2 v^T\Sigma v/2} + e^{\lambda^2(-v)^T\Sigma(-v)/2}\right) \\ &= e^{\lambda^2 v^T\Sigma v/2}\end{aligned}$$

and so

$$\mathop{\mathbb{E}}_y[e^{\lambda\langle y,v\rangle}] \leq e^{\lambda^2 v^T\Sigma v/2}.$$

Thus y is also (Σ, C)-subgamma. The target quantity in our theorem is the same for y as for x: $\mathbb{P}[\|x\| \geq (1+\gamma)\sqrt{\mathrm{Tr}(\Sigma)}] = \mathbb{P}[\|y\| \geq (1+\gamma)\sqrt{\mathrm{Tr}(\Sigma)}]$. Since $\|y\| \leq (1+\gamma)\sqrt{\mathrm{Tr}(\Sigma)}$ always, by considering y instead of x, we can WLOG assume that $\|x\| \leq (1+\gamma)\sqrt{\mathrm{Tr}(\Sigma)}$ in our theorem proof.

Relating probability to $\mathbb{E}_{x,v}[e^{\lambda\langle x,v\rangle}]$. Define

$$\alpha := \mathbb{P}\left[\|x\| \geq (1+\gamma)\sqrt{\mathrm{Tr}(\Sigma)}\right]$$

so that by Markov's inequality applied to $e^{\lambda^2\|x\|^2/2}$,

$$\alpha \leq \frac{\mathbb{E}[e^{\lambda^2\|x\|^2/2}]}{e^{\lambda^2(1+\gamma)^2\,\mathrm{Tr}(\Sigma)/2}}$$

for any λ. Now, let $v \sim N(0, I_d)$. For any x,

$$\mathop{\mathbb{E}}_{v}[e^{\lambda\langle x,v\rangle}] = e^{\lambda^2\|x\|^2/2}$$

so

$$\alpha \leq \mathop{\mathbb{E}}_{x,v}[e^{\lambda\langle x,v\rangle}]e^{-\lambda^2(1+\gamma)^2\,\mathrm{Tr}(\Sigma)/2}. \tag{4}$$

Upper bounding $\mathbb{E}_{x,v}[e^{\lambda\langle x,v\rangle}]$. We will bound the RHS above by making the inner expectation over x. Since x is (Σ, C)-subgamma, for every v,

$$\mathop{\mathbb{E}}_{x}[e^{\lambda\langle x,v\rangle}] \leq e^{\lambda^2 v^T\Sigma v/2} \qquad \forall|\lambda| \leq \frac{1}{\|Cv\|},$$

Therefore

$$\begin{aligned}
\mathop{\mathbb{E}}_{x,v}[e^{\lambda\langle x,v\rangle}] &= \mathop{\mathbb{E}}_{x,v}[e^{\lambda\langle x,v\rangle}1_{\|Cv\|\leq 1/|\lambda|} + e^{\lambda\langle x,v\rangle}1_{\|Cv\|>1/|\lambda|}] \\
&\leq \mathop{\mathbb{E}}_{v}[e^{\lambda^2 v^T\Sigma v/2}1_{\|Cv\|\leq 1/|\lambda|}] + \mathop{\mathbb{E}}_{x,v}[e^{\lambda\langle x,v\rangle}1_{\|Cv\|>1/|\lambda|}] \\
&\leq \mathop{\mathbb{E}}_{v}[e^{\lambda^2 v^T\Sigma v/2}] + \mathop{\mathbb{E}}_{x}[\mathop{\mathbb{E}}_{v}[e^{\lambda\langle x,v\rangle}1_{\|Cv\|>1/|\lambda|}]] \qquad (5)
\end{aligned}$$

We start with the first term. Let the eigenvalues of Σ be $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_d^2$. Then, $v^T\Sigma v/2$ is a generalized chi-squared distribution, distributed as $\sum_i u_i^2$ for independent Gaussian variables $u_i \sim N(0, \sigma_i^2/2)$. It is easy to check that u^2 for $u \sim N(0,1)$ is $(4,4)$-subgamma, i.e.,

$$\mathbb{E}[e^{\lambda(u^2-\mathbb{E}[u^2])}] = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{2\lambda^2} \qquad \forall|\lambda| \leq \frac{1}{4}.$$

Therefore $\sum u_i^2$ is $(\sum_i \sigma_i^4, 2\max\sigma_i^2) = (\|\Sigma\|_F^2, 2\|\Sigma\|)$-subgamma. Since $\|\Sigma\|_F^2 \leq \|\Sigma\|\,\mathrm{Tr}(\Sigma)$, $v^T\Sigma v$ is also $(\|\Sigma\|\,\mathrm{Tr}(\Sigma), 2\|\Sigma\|)$-subgamma.

Including the mean term as well ($\mathbb{E}[v^T\Sigma v/2] = \mathrm{Tr}(\Sigma)/2$), we have

$$\mathop{\mathbb{E}}_{v}[e^{\lambda^2 v^T\Sigma v/2}] \leq e^{\lambda^2\,\mathrm{Tr}(\Sigma)/2}\cdot e^{\lambda^4\,\mathrm{Tr}(\Sigma)\|\Sigma\|/2} \quad \forall\lambda^2 \leq \frac{1}{2\|\Sigma\|}. \tag{6}$$

We now bound the second term in (5) for each x. Since v is i.i.d. gaussian, $\|Cv\| \leq \|C\|_F + \|C\|\sqrt{2\log\frac{1}{\delta}}$ with probability $1-\delta$ (see Equation 1). Therefore, for all $|\lambda| < \frac{1}{2\|C\|_F}$,

$$\mathbb{P}[\|Cv\| > 1/|\lambda|] \leq e^{-\frac{(1/|\lambda|-\|C\|_F)^2}{2\|C\|^2}} \leq e^{-\frac{1}{8\lambda^2\|C\|^2}}$$

and so by Cauchy-Schwarz, and our bound on $\|x\|$,

$$\begin{aligned}
\mathop{\mathbb{E}}_{v}[e^{\lambda\langle x,v\rangle}1_{\|Cv\|>1/|\lambda|}] &\leq \sqrt{\mathop{\mathbb{E}}_{v}[e^{2\lambda\langle x,v\rangle}]\,\mathbb{P}[\|Cv\| > 1/|\lambda|]} \\
&\leq \sqrt{e^{2\lambda^2\|x\|^2}e^{-\frac{1}{8\lambda^2\|C\|^2}}} \\
&= e^{\lambda^2(1+\gamma)^2\,\mathrm{Tr}(\Sigma) - \frac{1}{16\lambda^2\|C\|^2}}.
\end{aligned}$$

Therefore, as long as $\lambda^2 \leq \min(\frac{1}{4(1+\gamma)\sqrt{\mathrm{Tr}(\Sigma)}\|C\|}, \frac{1}{4\|C\|_F^2})$,

$$\mathop{\mathbb{E}}_{v}[e^{\lambda\|x\|v_1}1_{\|Cv\|>1/|\lambda|}] \leq 1.$$

Combining with (6) (which is a bound always larger than 1) and (5),

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x,v}[e^{\lambda\langle x,v\rangle}] \leq 2e^{\lambda^2\,\mathrm{Tr}(\Sigma)/2}\cdot e^{\lambda^4\,\mathrm{Tr}(\Sigma)\|\Sigma\|/2} \\
&\qquad \forall\lambda^2 \leq \min(\frac{1}{2\|\Sigma\|}, \frac{1}{4(1+\gamma)\sqrt{\mathrm{Tr}(\Sigma)}\|C\|}, \frac{1}{4\|C\|_F^2})
\end{aligned}$$

and with (4),

$$\begin{aligned}
&\alpha \leq 2e^{\frac{1}{2}\lambda^2\,\mathrm{Tr}(\Sigma)(\lambda^2\|\Sigma\|-2\gamma-\gamma^2)} \\
&\forall\lambda^2 \leq \min(\frac{1}{2\|\Sigma\|}, \frac{1}{4(1+\gamma)\sqrt{\mathrm{Tr}(\Sigma)}\|C\|}, \frac{1}{4\|C\|_F^2})
\end{aligned}$$

Final bound. By also restricting λ^2 to be at most $\frac{2\gamma+\gamma^2}{2\|\Sigma\|}$, and setting λ^2 to the maximum of this range, we get

$$\alpha \leq 2e^{-\frac{(2\gamma+\gamma^2)\,\mathrm{Tr}(\Sigma)}{4}\min(\frac{1}{2\|\Sigma\|}, \frac{2\gamma+\gamma^2}{2\|\Sigma\|}, \frac{1}{4(1+\gamma)\sqrt{\mathrm{Tr}(\Sigma)}\|C\|}, \frac{1}{4\|C\|_F^2})}$$

The first two cases can be merged: $\min(\frac{2\gamma+\gamma^2}{2}, \frac{(2\gamma+\gamma^2)^2}{2}) \geq \frac{\gamma^2}{2}$. Thus:

$$\alpha \leq 2e^{-\frac{1}{16}\min(\frac{\gamma^2\,\mathrm{Tr}(\Sigma)}{\|\Sigma\|}, \frac{\gamma\sqrt{\mathrm{Tr}(\Sigma)}}{\|C\|}, \frac{(2\gamma+\gamma^2)\,\mathrm{Tr}(\Sigma)}{\|C\|_F^2})}.$$

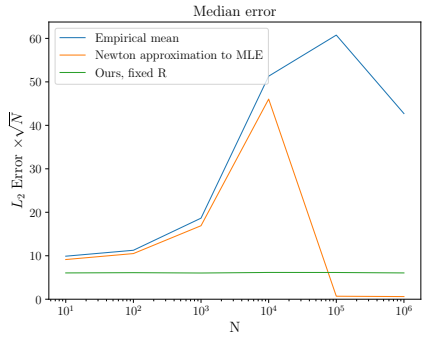
Plugging in $\gamma = \frac{t}{\sqrt{\mathrm{Tr}(\Sigma)}}$ gives the first result, and setting t such that the exponent is $\log\frac{2}{\delta}$ gives the second. □

6. Experiments

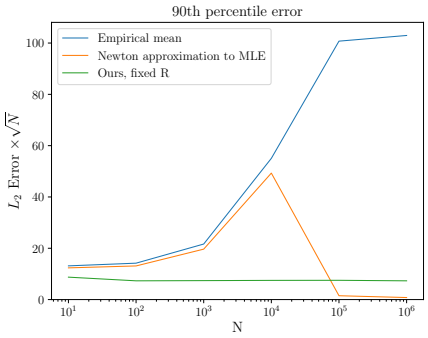
We perform experimental validation[1] on a synthetic high-dimensional example. We consider a mixture of three gaussians, two at similar scales and one very narrow and rare: $d = 20$, and $x \sim N(-e_1, I) + N(e_1, 9I) + 10^{-4}N(10^4 e_2, 10^{-6}I)$.

We consider three algorithms: our algorithm with smoothing radius 0.1; the empirical mean; and an approximation to the MLE given by Newton's method (i.e., our algorithm except

[1] Our implementation is available here: https://github.com/shivamgupta2/High-dimensional-location



(a) Median error



(b) 90th percentile error

Figure 3. Error scaled by $\sqrt{N}$ for different algorithms, for a synthetic Gaussian mixture.

with $R = 0$ and multiple steps; we use 10 steps). We use this approximation because we do not know how to compute the actual MLE efficiently in high dimensions.

For each algorithm, and for a variety of sample sizes N, we compute $\sqrt{N}$ times the estimation ℓ_2 error. Our theorem suggests that this should be about $\sqrt{\mathrm{Tr}(\mathcal{I}_R^{-1})} \approx 6.0$ for our algorithm with $R = 0.01I$, which is significantly better than the empirical mean's typical error of $\sqrt{\mathrm{Tr}(\Sigma)} \approx 70$, but significantly worse than the Cramer-Rao bound $\sqrt{\mathrm{Tr}(\mathcal{I}^{-1})} \approx 0.6$.

We find that:

- The observed constant for our algorithm's median error lies within $[6.0, 6.2]$, which is within 4% of the constant in the main term in our theorem ($\sqrt{\mathrm{Tr}(\mathcal{I}_R^{-1})} \approx 6.0$);
- The empirical mean does not benefit from the Fisher information being much better than the covariance, so it performs much worse than our algorithm.
- Asymptotically the approximate MLE is optimal, and it does work very well for large enough N, but struggles for small N.

In this experiment we fixed R as n varies; for very large n, one should run our algorithm with smaller R to make $\mathcal{I}_R$ converge down to 0.6.

7. Conclusion and Future Work

In this paper we gave an algorithm for location estimation in high dimensions, getting non-asymptotic error bounds approaching those of $\mathcal{N}(0, \frac{\mathcal{I}_R^{-1}}{n})$, where $\mathcal{I}_R$ is the Fisher information matrix of our distribution when smoothed using $\mathcal{N}(0, R)$ for small R that decays with n. In the process of proving this result, we obtained a new concentration inequality for the norm of high-dimensional random variables whose 1-dimensional projections are subgamma, which may be of independent interest. Even in 1 dimension, our results give improvement for constant failure probability. For function classes such as a mixture of Laplacians, no previous work gives a rate for the asymptotic convergence to the Cramér-Rao bound as $n \to \infty$ for fixed δ.

This paper is one step in the finite-sample theory of parameter estimation. Our quantitative bounds could be improved: our bound on the rate of convergence to Cramér-Rao is $1 + \frac{1}{\mathrm{poly}(n)}$, but one could hope for faster convergence ($1 + \frac{1}{\sqrt{n}}$ in general, and $1 + \frac{1}{n}$ for some specific function classes). More generally, one can consider estimation of parameters other than location; the Cramér-Rao bound still relates the asymptotic behavior to the Fisher information, but a rate of convergence remains elusive. We believe that understanding high-dimensional location estimation is a good step toward understanding the estimation of multiple parameters.

8. Acknowledgments

Shivam Gupta and Eric Price are supported by NSF awards CCF-2008868, CCF-1751040 (CAREER), and the NSF AI Institute for Foundations of Machine Learning (IFML). Some of this work was done while Shivam Gupta was visiting UC Berkeley. Jasper C.H. Lee is supported in part by the generous funding of a Croucher Fellowship for Postdoctoral Research, NSF award DMS-2023239, NSF Medium Award CCF-2107079 and NSF AiTF Award CCF-2006206.

References

Adamczak, R. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electronic Journal of Probability*, 13(none):1000 – 1034, 2008. doi: 10.1214/EJP.v13-521. URL https://doi.org/10.1214/EJP.v13-521.

Bickel, P. J. and Doksum, K. A. *Mathematical statistics: basic ideas and selected topics, volume I*. Chapman and Hall/CRC, 2015.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. ISBN 978-0-19-953525-5. doi: 10.1093/acprof:oso/9780199535255.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199535255.001.0001.

Bousquet, O. A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002.

Bousquet, O. Concentration inequalities for sub-additive functions using the entropy method. In Giné, E., Houdré, C., and Nualart, D. (eds.), *Stochastic Inequalities and Applications*, pp. 213–247, Basel, 2003.

Catoni, O. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148 – 1185, 2012. doi: 10.1214/11-AIHP454. URL https://doi.org/10.1214/11-AIHP454.

Cherapanamjeri, Y., Flammarion, N., and Bartlett, P. L. Fast mean estimation with sub-gaussian rates. In Beygelzimer, A. and Hsu, D. (eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 786–806. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/cherapanamjeri19b.html.

Devroye, L., Lerasle, M., Lugosi, G., and Oliveira, R. I. Sub-Gaussian mean estimators. *Ann. Stat*, 44(6):2695–2725, 2016.

Diakonikolas, I., Kane, D. M., and Pensia, A. Outlier robust mean estimation with subgaussian rates via stability. *Advances in Neural Information Processing Systems*, 33: 1830–1840, 2020.

Gupta, S., Lee, J. C. H., Price, E., and Valiant, P. Finite-sample maximum likelihood estimation of location. In *Proc. NeurIPS'22*, 2022.

Hendeby, G. *Fundamental Estimation and Detection Limits in Linear Non-Gaussian Systems*. PhD thesis, 11 2005.

Hopkins, S. B. Sub-gaussian mean estimation in polynomial time. *ArXiv*, abs/1809.07425, 2018.

Hsu, D., Kakade, S., and Zhang, T. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17:1–6, 2012.

Lee, J. C. H. and Valiant, P. Optimal sub-gaussian mean estimation in $\mathbb{R}$. In *Proc. FOCS'21*, pp. 672–683, 2022a.

Lee, J. C. H. and Valiant, P. Optimal sub-gaussian mean estimation in very high dimensions. In *Proc. ITCS'22*, pp. 98:1–98:21, 2022b.

Lugosi, G. and Mendelson, S. Sub-gaussian estimators of the mean of a random vector. *Annals of Statistics*, 47, 02 2017. doi: 10.1214/17-AOS1639.

Miao, Y. Concentration inequality of maximum likelihood estimator. *Applied Mathematics Letters*, 23(10):1305–1309, 2010.

Pinelis, I. Optimal-order uniform and nonuniform bounds on the rate of convergence to normality for maximum likelihood estimators. *Electronic Journal of Statistics*, 11 (1):1160 – 1179, 2017. doi: 10.1214/17-EJS1264. URL https://doi.org/10.1214/17-EJS1264.

Spokoiny, V. Parametric estimation. finite sample theory. *The Annals of Statistics*, 40(6):2877–2909, 2011. doi: 10.1214/12-AOS1054.

user940. Show $\mathbb{E}[f(x)g(x)] \geq \mathbb{E}[f(x)]\mathbb{E}[g(x)]$ for f, g bounded, nondecreasing. Mathematics Stack Exchange, 2015. URL https://math.stackexchange.com/q/1446526. Downloaded 2023-01.

van de Geer, S. A. and Lederer, J. The bernstein–orlicz norm and deviation inequalities. *Probability Theory and Related Fields*, 157:225–250, 2011.

van der Vaart, A. *Asymptotic Statistics*. Cambridge University Press, 2000. ISBN 9780521784504.

A. Complete analysis of 1-dimensional location estimation

A.1. 1-dimensional local estimation

The following algorithm (Algorithm 1) is the *local* part of the 1-dimensional estimation: it assumes that there is an initial estimate that is close to the true parameter λ.

Algorithm 1 Local smoothed MLE for one dimension

Input Parameters:

- Description of f, smoothing parameter r, samples $x_1, \ldots, x_n \overset{i.i.d.}{\sim} f^\lambda$ and initial estimate λ_1 of λ

1. Let $s(\hat{\lambda})$ be the score function of f_r, the r-smoothed version of f.
2. For each sample x_i, compute a perturbed sample $x_i' = x_i + \mathcal{N}(0, r^2)$ where all the Gaussian noise are drawn independently across all the samples.
3. Compute the empirical score at λ_1, namely $\hat{s}(\lambda_1) = \frac{1}{n}\sum_{i=1}^n s(x_i' - \lambda_1)$.
4. Return $\hat{\lambda} = \lambda_1 - (\hat{s}(\lambda_1)/\mathcal{I}_r)$.

The local algorithm is what uses the simplified view of smoothed MLE and distinguishes our approach from the previous approach of Gupta et al. (2022).

We will show the following guarantee for Algorithm 1. It says that, if the initial estimate λ_1 has distance at most $\epsilon_{\max}$ from true parameter λ, and suppose we choose a sufficiently large smoothing parameter r, then the output of Algorithm 1 will be close to the true parameter λ.

Lemma A.1. *In Algorithm 1, suppose $|\lambda_1 - \lambda| \le \epsilon_{\max}$ for some $\epsilon_{\max} \ge \sqrt{\frac{2\log\frac{2}{\delta}}{n}\frac{1}{\mathcal{I}_r}}$. Suppose also that the smoothing parameter is $r \ge 2\epsilon_{\max}$, and there exists a parameter $\gamma \ge 1$ such that 1) $r^2\sqrt{\mathcal{I}_r} \ge \gamma\epsilon_{\max}$, 2) $r^2\sqrt{\log\frac{2}{\delta}/n} \ge \gamma\epsilon_{\max}^2$ and 3) $(\log\frac{2}{\delta})/n \le 1/\gamma^2$. (For interpretation, γ is supposed to be large and "$\omega(1)$" when the lemma is used.)*

Then, with probability at least $1-\delta$ over n samples from f^λ, the output of Algorithm 1 satisfies

$$|\hat{\lambda} - \lambda| \le \left(1 + O\left(\frac{1}{\gamma}\right)\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n\mathcal{I}_r}}$$

The proof of Lemma A.1 relies on the following facts from (Gupta et al., 2022) about the concentration of the empirical score of the smoothed distribution, when evaluated at an initial parameter estimate that are close to the true parameter.

The first fact is the subgamma concentration of the score.

Fact A.2. *Suppose we take n i.i.d. samples $y_1, \ldots, y_n \leftarrow f_r^\lambda$, and consider the empirical score function $\hat{s}$ mapping a candidate parameter $\hat{\lambda}$ to $\frac{1}{n}\sum_i s_r(y_i - \hat{\lambda})$, where s_r is the score function of f_r.*

Then, for any $|\epsilon| \le r/2$,

$$\mathbb{P}_{y_i \overset{i.i.d.}{\sim} f_r^\lambda}\left(|\hat{s}(\lambda+\epsilon) - \mathbb{E}_{x\leftarrow f_r}[s(x-\epsilon)]| \ge \sqrt{\frac{2\max(\mathbb{E}_x[s_r^2(x-\epsilon)], \mathcal{I}_r)\log\frac{2}{\delta}}{n}} + \frac{15\log\frac{2}{\delta}}{nr}\right) \le \delta$$

The next two facts bound the expectation and second moment of the score.

Fact A.3. *For any $|\epsilon| \le r/2$, the expected score $\mathbb{E}_{x\sim f_r}[s_r(x+\epsilon)]$ satisfies*

$$\mathbb{E}_{x\sim f_r}[s_r(x+\epsilon)] \in \left[-\mathcal{I}_r\epsilon \pm O\left(\sqrt{\mathcal{I}_r}\frac{\epsilon^2}{r^2}\right)\right]$$

Fact A.4. *For any $|\epsilon| \le r/2$, if $r/\epsilon = \Omega(\sqrt{\log e/(r^2\mathcal{I}_r)})$, the second moment of the score satisfies*

$$\mathop{\mathbb{E}}_{x\sim f_r}\left[s_r^2(x+\epsilon)\right] \le \mathcal{I}_r\left(1 + O\left(\frac{\epsilon}{r}\sqrt{\log\frac{e}{r^2\mathcal{I}_r}}\right)\right)$$

Furthermore, we always have $\mathcal{I}_r \le 1/r^2$, and therefore $\sqrt{\log 1/(r^2\mathcal{I}_r)}$ above is well-defined.

We can now prove Lemma A.1 using these facts. The proof strategy is straightforward: we use Facts A.2 and A.4 to show that $\hat{s}(y)$ concentrates close to its expectation with high probability, and we use Fact A.3 to show that the expectation of $\hat{s}(y)$, which is $\mathbb{E}[s(x-\epsilon)]$ for $y = \lambda + \epsilon$, is very close to $\mathcal{I}_r\epsilon$. The triangle inequality then implies that $y - (\hat{s}(y)/\mathcal{I}_r)$ must be close to λ with high probability.

Proof of Lemma A.1. Let $\lambda_1 = \lambda + \epsilon$. By the lemma assumptions, $|\epsilon| \le \epsilon_{\max}$.

First, we show that, under the lemma assumption that $r^2\sqrt{\mathcal{I}_r} \ge \gamma\epsilon_{\max}$, Fact A.4 implies that the second moment of the score at $\lambda - \epsilon$, namely $\mathbb{E}_{x\sim f_r}[s_r^2(x+\epsilon)]$, is upper bounded by $(1 + O(1/\gamma))\mathcal{I}_r$.

To check that the precondition of Fact A.4 holds, note that $r^2\sqrt{\mathcal{I}_r} \ge \gamma\epsilon_{\max} \ge \gamma\epsilon$ is equivalent to $r/\epsilon \ge \gamma/\sqrt{r^2\mathcal{I}_r}$, which implies that

$$\begin{aligned}\frac{r}{\epsilon} &\ge \frac{\gamma}{\sqrt{r^2\mathcal{I}_r}} \\ &= \frac{\gamma}{\sqrt{e}}\sqrt{\frac{e}{r^2\mathcal{I}_r}} \\ &\ge \frac{\gamma}{\sqrt{e}}\sqrt{\log\frac{e}{r^2\mathcal{I}_r}}\end{aligned}$$

satisfying the precondition of Fact A.4.

Then, the fact implies that

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim f_r}[s_r^2(x+\epsilon)] &\le \mathcal{I}_r\left(1 + O\left(\frac{\epsilon}{r}\sqrt{\log\frac{e}{r^2\mathcal{I}_r}}\right)\right) \\ &\le \mathcal{I}_r\left(1 + O\left(\frac{\epsilon_{\max}}{r}\sqrt{\log\frac{e}{r^2\mathcal{I}_r}}\right)\right) \\ &\le \mathcal{I}_r\left(1 + O\left(\frac{\epsilon_{\max}}{r}\sqrt{\frac{e}{r^2\mathcal{I}_r}}\right)\right) \\ &\le \mathcal{I}_r\left(1 + O\left(\frac{\epsilon_{\max}}{r^2\sqrt{\mathcal{I}_r}}\right)\right) \\ &\le \mathcal{I}_r\left(1 + O\left(\frac{1}{\gamma}\right)\right)\end{aligned}$$

Next, we combine the concentration bound of Fact A.2 with the second moment bound for $\mathbb{E}_x[s_r^2(x+\epsilon)]$ we just derived to show that $\hat{s}(\lambda - \epsilon)$ is close to its expectation with high probability.

$$\begin{aligned}\left|\hat{s}(y) - \mathop{\mathbb{E}}_{x\leftarrow f_r}[s(\lambda-\epsilon)]\right| &\le \sqrt{\frac{2\log\frac{2}{\delta}}{n}\mathcal{I}_r}\left(1 + O\left(\frac{1}{\gamma}\right)\right) + \frac{15\log\frac{2}{\delta}}{nr} \\ &\le \left(1 + O\left(\frac{1}{\gamma}\right)\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n}\mathcal{I}_r} + \frac{15}{2\sqrt{\gamma}}\left(\frac{2\log\frac{2}{\delta}}{n}\right)^{\frac{1}{4}}\sqrt{\frac{2\log\frac{2}{\delta}}{n}\mathcal{I}_r} \quad \text{(see below)} \\ &\le \left(1 + O\left(\frac{1}{\gamma}\right)\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n}\mathcal{I}_r} + O\left(\frac{1}{\gamma}\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n}\mathcal{I}_r} \quad \text{since } \log\frac{2}{\delta}/n \le 1/\gamma^2\end{aligned}$$

$$= \left(1 + O\left(\frac{1}{\gamma}\right)\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n}\mathcal{I}_r}$$

where the second inequality is due to the assumption that $r^2\sqrt{\mathcal{I}_r} \geq \gamma\epsilon_{\max} \geq \gamma\sqrt{\frac{2\log\frac{1}{\delta}}{n}\frac{1}{\mathcal{I}_r}}$.

Further using Fact A.3, this implies that $\epsilon = y - \lambda$ is well-approximated by $\hat{s}(y)/\mathcal{I}_r$, as follows.

$$\begin{aligned}
|\epsilon - (\hat{s}(y)/\mathcal{I}_r)| &= \frac{1}{\mathcal{I}_r}|\mathcal{I}_r\epsilon - \hat{s}(y)| \\
&= \frac{1}{\mathcal{I}_r}\left|\hat{s}(y) - \mathop{\mathbb{E}}_{x\leftarrow f_r}[s(\lambda-\epsilon)] + \mathop{\mathbb{E}}_{x\leftarrow f_r}[s(\lambda-\epsilon)] - \mathcal{I}_r\epsilon\right| \\
&\leq \frac{1}{\mathcal{I}_r}\left|\hat{s}(y) - \mathop{\mathbb{E}}_{x\leftarrow f_r}[s(\lambda-\epsilon)]\right| + \frac{1}{\mathcal{I}_r}\left|\mathop{\mathbb{E}}_{x\leftarrow f_r}[s(\lambda-\epsilon)] - \mathcal{I}_r\epsilon\right| \\
&= \left(1 + O\left(\frac{1}{\gamma}\right)\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n\mathcal{I}_r}} + O\left(\frac{\epsilon^2}{r^2\sqrt{\mathcal{I}_r}}\right) \\
&\qquad \text{by the previous bound and Fact A.3}
\end{aligned}$$

By the lemma assumption, we have $\epsilon^2/r^2 \leq \epsilon_{\max}^2/r^2 \leq (1/\gamma)\sqrt{\log\frac{2}{\delta}/n}$, and so we have bounded $|\epsilon - (\hat{s}(y)/\mathcal{I}_r)|$ by

$$|\epsilon - (\hat{s}(y)/\mathcal{I}_r)| \leq \left(1 + O\left(\frac{1}{\gamma}\right)\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n\mathcal{I}_r}}$$

To conclude, we have

$$|\hat{\lambda} - \lambda| = |y - (\hat{s}(y)/\mathcal{I}_r) - \lambda| = |\lambda + \epsilon - (\hat{s}(y)/\mathcal{I}_r) - \lambda| \leq \left(1 + O\left(\frac{1}{\gamma}\right)\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n\mathcal{I}_r}}$$

as desired. □

A.2. 1-dimensional global estimation

We can now state the 1-dimensional global estimation algorithm (Algorithm 2), which first gets a preliminary estimate of the true parameter from a $o(1)$ fraction of the data, before invoking the local Algorithm 1 on the rest of the data.

Algorithm 2 Global smoothed MLE for one dimension

Input Parameters:

- Failure probability δ, description of f, n i.i.d. samples drawn from f^λ for some unknown λ

1. Let q be $\sqrt{2}(\log\frac{2}{\delta}/n)^{2/5}$.
2. Compute an $\alpha \in [q, 1-q]$ to minimize the width of interval defined by the $\alpha \pm q$ quantiles of f.
3. Take the sample α-quantile of the first $(\log\frac{2}{\delta}/n)^{1/10}$ fraction of the n samples.
4. Let $r^* = \Omega((\frac{\log\frac{2}{\delta}}{n})^{1/8})\text{IQR}$.
5. Run Algorithm 1 on the rest of the samples, using initial estimate $\lambda_1 = x_\alpha$ and r^*-smoothing, and return the final estimate $\hat{\lambda}$.

Both the global part of the algorithm and its analysis are essentially identical to what Gupta et al. (2022), up to minor changes in certain parameters. We note again that the algorithmic improvement lies in the local part of the algorithm, in Algorithm 1.

Theorem 1.1 (1-d Smoothed MLE). *Given a model f, let the r-smoothed Fisher information of a distribution f be $\mathcal{I}_r$, and let* IQR *be the interquartile range of f. Fix the failure probability be $\delta \le 0.5$, and assume that $n \ge c \cdot \log \frac{2}{\delta}$ for some sufficiently large constant c.*

Choose $r^* = \Omega((\frac{\log \frac{2}{\delta}}{n})^{1/8})$IQR. Then, with probability at least $1-\delta$, the output $\hat{\lambda}$ of Algorithm 2 satisfies

$$|\hat{\lambda} - \lambda| \le \left(1 + O\left(\frac{\log \frac{2}{\delta}}{n}\right)^{\frac{1}{10}}\right)\sqrt{\frac{2\log \frac{2}{\delta}}{n\mathcal{I}_{r^*}}}$$

The analysis of Algorithm 2 requires one more technical fact from (Gupta et al., 2022), which is a lower bound on smoothed Fisher information.

Fact A.5. *Let $\mathcal{I}_r$ be the Fisher information for f_r, the r-smoothed version of distribution f. Let* IQR *be the interquartile range of f. Then, $\mathcal{I}_r \gtrsim 1/(\text{IQR}+r)^2$. Here, the hidden constant is a universal one independent of the distribution f and independent of r.*

Proof of Theorem 1.1. Step 2 uses $(\log \frac{2}{\delta}/n)^{1/10}\, n$ samples to compute the sample α-quantile. By standard Chernoff bounds, with probability at least $1 - \delta(\log \frac{2}{\delta}/n)^2$, the error of the sample quantile (in terms of its quantile in the true distribution) is at most

$$\begin{aligned}
&\sqrt{\frac{2\log \frac{2}{\delta(\log \frac{2}{\delta}/n)^2}}{(\log \frac{2}{\delta}/n)^{1/10}\, n}} \\
&\le \sqrt{\frac{2(\log \frac{2}{\delta})(\frac{n}{\log \frac{2}{\delta}})^{1/10}}{(\log \frac{2}{\delta}/n)^{1/10}\, n}} \\
&= \sqrt{2}\left(\frac{\log \frac{2}{\delta}}{n}\right)^{2/5}
\end{aligned}$$

Therefore, if the above event happens, Step 2 will yield a sample α-quantile x_α such that $x_\alpha - \lambda$ is within the $\alpha - \sqrt{2}(\log \frac{2}{\delta}/n)^{2/5}$ and $\alpha + \sqrt{2}(\log \frac{2}{\delta}/n)^{2/5}$ quantiles of f. Furthermore, by the minimality condition in the definition of α, the distance between these two quantiles is at most $O((\log \frac{2}{\delta}/n)^{2/5})$IQR.

We will apply Lemma A.1 using failure probability $\delta(1 - (\log \frac{2}{\delta}/n)^2)$. We will check that, **(A)** conditioned on Step 2 succeeding in the above sense, the preconditions of Lemma A.1 will hold for $\lambda_1 = x_\alpha$, the chosen r^* and an appropriate choice of γ, and also that **(B)** the estimation error guaranteed by Lemma A.1 implies the desired error bound. If the above deterministic checks are true, then by a union bound, Algorithm 2 will satisfy the desired intermediate bound guarantees except with probability δ.

For the following calculations, note that $\log \frac{2}{\delta(1-(\log \frac{2}{\delta}/n)^2)} \le 1.1 \log \frac{2}{\delta}$ since $n \gg \log \frac{2}{\delta}$ and $\delta \le 0.5$.

(A): We condition on Step 2 succeeding, and check the preconditions of Lemma A.1.

We now check the precondition that $r^* \ge 2\epsilon_{\max}$, for $\epsilon_{\max} = \max(\sqrt{2\log \frac{2}{\delta(1-(\log \frac{2}{\delta}/n)^2)}/(n\mathcal{I}_{r^*})}, O(\log \frac{2}{\delta}/n)^{2/5}\text{IQR})$. First, $r^* = \Omega((\frac{\log \frac{2}{\delta}}{n})^{1/8})\text{IQR} \gg O((\log \frac{2}{\delta}/n)^{2/5})\text{IQR}$, where the $\gg$ uses the assumption on the size of n. We can also show that $O(\log \frac{1}{\delta}/n)^{2/5}\text{IQR} \ge \sqrt{2\log \frac{2}{\delta}/(n\mathcal{I}_{r^*})}$. Recall by Fact A.5 that $\mathcal{I}_r \ge \Omega(1/(\text{IQR}+r)^2)$ for any $r > 0$. Therefore, $\sqrt{2\log \frac{2}{\delta(1-(\log \frac{2}{\delta}/n)^2)}/(n\mathcal{I}_{r^*})} \le O(\sqrt{\log \frac{2}{\delta}/(n\mathcal{I}_{r^*})}) \le O((\log \frac{2}{\delta}/n)^{1/2}\text{IQR}) \ll O((\log \frac{2}{\delta}/n)^{2/5})\text{IQR}$, where the $\ll$ is due to the theorem assumption on the size of n.

We now need to check the last 3 preconditions of Lemma A.1. Let $n' = (1 - (\log \frac{2}{\delta}/n)^{1/10})n$ be the number of samples used in the call to Algorithm 1, in Step 4. By the theorem assumption, we have $n' = \Theta(n)$. Further, recall by Fact A.5 that $\mathcal{I}_r \ge \Omega(1/(\text{IQR}+r)^2)$. Picking $\gamma = O(\frac{n}{\log \frac{2}{\delta}})^{1/10}$, we check that the following remaining conditions from Lemma A.1 are satisfied when applied to the $n' = \Theta(n)$ points used in Step 4:

1. $(r^*)^2\sqrt{\mathcal{I}_{r^*}} \geq (r^*)^2/(\mathrm{IQR}+r^*) \geq \Omega(\frac{\log\frac{2}{\delta}}{n})^{1/4}\mathrm{IQR} \geq \Omega(\frac{\log\frac{2}{\delta}}{n})^{3/10}\mathrm{IQR} \geq \gamma\epsilon_{\max}$.
2. $(r^*)^2\sqrt{\log\frac{2}{\delta(1-(\log\frac{2}{\delta}/n)^2)}/n'} \geq \Omega(\frac{\log\frac{2}{\delta}}{n})^{1/4}\mathrm{IQR}^2\sqrt{\log\frac{1}{\delta}/n} = \Omega(\frac{\log\frac{2}{\delta}}{n})^{7/10}\mathrm{IQR}^2 = \gamma\epsilon_{\max}^2$
3. $\log\frac{2}{\delta(1-(\log\frac{2}{\delta}/n)^2)}/n' \leq O(\log\frac{2}{\delta}/n') \leq O((\log\frac{2}{\delta}/n)^{1/5}) \leq 1/\gamma^2$.

(B): We check that the guarantees of Lemma A.1 is sufficient to imply the desired bound. To do so, we need a slightly more refined bound on $\log\frac{2}{\delta(1-(\log\frac{2}{\delta}/n)^2)}$:

$$\begin{aligned}
\log\frac{2}{\delta(1-(\log\frac{2}{\delta}/n)^2)} &= \left(1+\frac{\log\frac{1}{1-(\log\frac{2}{\delta}/n)^2}}{\log\frac{2}{\delta}}\right)\log\frac{2}{\delta} \\
&\leq \left(1+O\left(\frac{(\log\frac{2}{\delta}/n)^2}{\log\frac{2}{\delta}}\right)\right)\log\frac{2}{\delta} \quad \text{since } n \gg \log\frac{2}{\delta} \\
&\leq \left(1+O\left(\frac{\log\frac{2}{\delta}}{n}\right)\right)\log\frac{2}{\delta}
\end{aligned}$$

When the preconditions of Lemma A.1, the success of Step 4 implies a final estimate $\hat{\lambda}$ satisfying

$$\begin{aligned}
|\hat{\lambda}-\lambda| &\leq \left(1+O\left(\frac{1}{\gamma}\right)\right)\sqrt{\frac{2\log\frac{2}{\delta(1-(\log\frac{2}{\delta}/n)^2)}}{n'\mathcal{I}_{r^*}}} \\
&\leq \left(1+O\left(\frac{1}{\gamma}\right)+O\left(\frac{\log\frac{2}{\delta}}{n}\right)\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n'\mathcal{I}_{r^*}}} \\
&= \left(1+O\left(\frac{\log\frac{2}{\delta}}{n}\right)^{\frac{1}{10}}+O\left(\frac{\log\frac{2}{\delta}}{n}\right)\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n'\mathcal{I}_{r^*}}} \\
&= \left(1+O\left(\frac{\log\frac{2}{\delta}}{n}\right)^{\frac{1}{10}}\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n'\mathcal{I}_{r^*}}} \\
&= \left(1+O\left(\frac{\log\frac{2}{\delta}}{n}\right)^{\frac{1}{10}}\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n\mathcal{I}_{r^*}}} \\
&\quad \text{since } n' = \left(1-\left(\log\frac{2}{\delta}/n\right)^{1/10}\right)n
\end{aligned}$$

□

B. High dimensional location estimation

This section provides a complete analysis of our main Theorem B.16 for estimating the location of a high-dimensional distribution. We start by providing some important definitions in Appendix B.1. Then, in Appendix B.2, we prove some key properties of the score of our smoothed distribution. In Appendix B.3 we show that our score function is subgamma with appropriate variance and scale parameters. Then, Appendix B.4 shows an error bound for the deviation between the empirical score estimate and true true score. Finally Appendix B.5 and B.6 provide analyses of our Local MLE and Global MLE algorithms respectively.

B.1. Definitions

Let f be an arbitrary distribution on $\mathbb{R}^d$ and let $Y \sim f$. Let our smoothing parameter $R \in \mathbb{R}^{d\times d}$ be the covariance matrix of our noise $Z_R \sim w_R = \mathcal{N}(0,R)$ sampled independently of Y. We define the *R-smoothed* distribution f_R to be such that

$X = Y + Z_R \sim f_R$. Thus, the pdf of f_R is given by

$$f_R(x) = \mathop{\mathbb{E}}_{Z_R \sim w_R}[f(x + Z_r)]$$

Let s_R be the score function of f_R. We have

$$s_R(x) = \nabla \log f_R(x) = \frac{\nabla f_R(x)}{f_R(x)}$$

Let $\mathcal{I}_R$ be the Fisher information matrix of f_R. Then,

$$\mathcal{I}_R = \mathop{\mathbb{E}}_{x \sim f_R}[s_R(x) s_R(x)^T]$$

We define the M-norm of vector x to be

$$\|x\|_M = \sqrt{x^T M x}$$

B.2. Properties of the smoothed score

In this section, we prove some properties of the score function s_R of the R-smoothed distribution f_R that we make use of throughout the paper. First, in Lemma B.1, we provide a useful characterization of s_R. Then, using Lemma B.2 we prove Lemma B.3, which tells us for good initial estimates of our location, say incurring error $\epsilon \in \mathbb{R}^d$ for "small" ϵ, "inverting the score" by left multiplying $s_R(x + \epsilon)$ by $-\mathcal{I}_R^{-1}$ provides a good estimate of the error ϵ in expectation. After this, using Lemma B.4, we prove Lemma B.5, which says that for small ϵ, the shifted score $s_R(x + \epsilon)$ when appropriately transformed has covariance similar to the corresponding transformation of the Fisher information matrix $\mathcal{I}_R$.

We begin by providing a characterization of the score s_R that we make use of throughout.

Lemma B.1. *Let f be an arbitrary distribution on $\mathbb{R}^d$, and let f_R be the R-smoothed version of f. That is, $f_R(x) = \mathbb{E}_{y \sim f}\left[(2\pi)^{-d/2} \det(R)^{-1/2} \exp\left(-\frac{1}{2}(x - Y)^T R^{-1}(x - Y)\right)\right]$. Let s_R be the score function of f_R. Let (X, Y, Z_R) be the joint distribution such that $Y \sim f$, $Z_R \sim \mathcal{N}(0, R)$ are independent, and $X = Y + Z_R \sim f_R$. We have for $\epsilon \in \mathbb{R}^d$,*

$$\frac{f_R(x + \epsilon)}{f_R(x)} = \mathop{\mathbb{E}}_{Z_R | x}\left[e^{\epsilon^T R^{-1} Z_R - \frac{1}{2}\epsilon^T R^{-1} \epsilon}\right]$$

so that

$$s_R(x) = \mathop{\mathbb{E}}_{Z_R | x}\left[R^{-1} Z_R\right]$$

Proof. First, we show that for $\epsilon \in \mathbb{R}$

$$\frac{f_R(x + \epsilon)}{f_R(x)} = \mathop{\mathbb{E}}_{Z_R | x}\left[\frac{w_R(Z_R + \epsilon)}{w_R(Z_R)}\right]$$

Note that

$$p(z|x) = \frac{p(z, x)}{p(x)} = \frac{f(x - z) w_R(z)}{f_R(x)}$$

So,

$$\begin{aligned} f_R(x + \epsilon) &= \int_{[-\infty, \infty]^d} w_R(z) f(x + \epsilon - z) dz \\ &= \int p(z|x) f_R(x) \frac{w_R(z + \epsilon)}{w_R(z)} dz \\ &= f_R(x) \mathop{\mathbb{E}}_{Z_R | x}\left[\frac{w_R(Z_R + \epsilon)}{w_r(Z_R)}\right] \end{aligned}$$

But now, $w_R(x) = (2\pi)^{-d/2} \det(R)^{-1/2} e^{-\frac{1}{2} x^T R^{-1} x}$ So,

$$\frac{f_R(x + \epsilon)}{f_R(x)} = \mathop{\mathbb{E}}_{Z_R | x}\left[e^{\epsilon^T R^{-1} Z_R - \frac{1}{2}\epsilon^T R^{-1} \epsilon}\right]$$

which is the first claim. Now, let $\epsilon = \gamma e_i$. We take the derivative wrt γ, and evaluate at $\gamma = 0$ to get

$$\frac{\nabla_{e_i} f_R(x)}{f_R(x)} = \mathop{\mathbb{E}}_{Z_R|x}\left[(R^{-1}Z_R)_i\right]$$

So,

$$s_R(x) = \frac{\nabla f_R(x)}{f_R(x)} = \mathop{\mathbb{E}}_{Z_R|x}\left[R^{-1}Z_R\right]$$

□

The next Lemma B.2 is a utility result that we make use of in Lemma B.3.

Lemma B.2. *Let f_R be the R-smoothed version of distribution f on $\mathbb{R}^d$. For $\epsilon \in \mathbb{R}^d$, let*

$$\Delta_\epsilon(x) := \frac{f_R(x+\epsilon) - f_R(x) - (\nabla f_R(x))^T\epsilon}{f_R(x)}$$

Then, for any ϵ such that $|\epsilon^T R^{-1}\epsilon| \le \frac{1}{4}$, we have

$$\mathop{\mathbb{E}}_{x\sim f_R}[\Delta_\epsilon(x)^2] \lesssim (\epsilon^T R^{-1}\epsilon)^2$$

Proof. By Lemma B.1, we have

$$\Delta_\epsilon(x) = \frac{f_R(x+\epsilon) - f_R(x) - (\nabla f_R(x))^T\epsilon}{f_R(x)} = \mathop{\mathbb{E}}_{Z_R|x}\left[e^{\epsilon^T R^{-1}Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon} - 1 - Z_R^T R^{-1}\epsilon\right]$$

Let $\alpha_\epsilon : \mathbb{R}^d \to \mathbb{R}$ be such that

$$\alpha_\epsilon(z) = e^{\epsilon^T R^{-1}z - \frac{1}{2}\epsilon^T R^{-1}\epsilon} - 1 - z^T R^{-1}\epsilon$$

We want to bound

$$\begin{aligned}
\mathop{\mathbb{E}}_x[\Delta_\epsilon(x)^2] &= \mathop{\mathbb{E}}_x[\mathop{\mathbb{E}}_{Z_R|X}[\alpha_\epsilon(Z_r)]^2] \\
&\le \mathop{\mathbb{E}}_{x,Z_R}[(\alpha_\epsilon(Z_R))^2] \\
&= \mathop{\mathbb{E}}_{Z_R\sim\mathcal{N}(0,R)}[(\alpha_\epsilon(Z_R))^2]
\end{aligned} \tag{7}$$

For the remaining proof, let $W = \epsilon^T R^{-1}Z_R$. Since $Z_R \sim \mathcal{N}(0,R)$, we have that $W \sim \mathcal{N}(0, \epsilon^T R^{-1}\epsilon)$. When $|W| \le 1$, by a Taylor expansion, we have

$$e^{W - \frac{1}{2}\epsilon^T R^{-1}\epsilon} = 1 + W - \frac{1}{2}\epsilon^T R^{-1}\epsilon + O\left(\left(W - \frac{1}{2}\epsilon^T R^{-1}\epsilon\right)^2\right)$$

so that

$$|\alpha_\epsilon(Z_R)| \lesssim \epsilon^T R^{-1}\epsilon + W^2$$

This implies that $\alpha_\epsilon(Z_R)^2 \lesssim (\epsilon^T R^{-1}\epsilon)^2 + W^4$, meaning that

$$\begin{aligned}
\mathop{\mathbb{E}}_{Z_R\sim\mathcal{N}(0,R)}\left[(\alpha_\epsilon(Z_R))^2 \cdot \mathbb{1}_{|\epsilon^T R^{-1}Z_R|\le 1}\right] &\lesssim \mathop{\mathbb{E}}_{W\sim\mathcal{N}(0,\epsilon^T R^{-1}\epsilon)}\left[\left((\epsilon^T R^{-1}\epsilon)^2 + W^4\right)\cdot \mathbb{1}_{|\epsilon^T R^{-1}\epsilon|\le 1}\right] \\
&\lesssim (\epsilon^T R^{-1}\epsilon)^2 + \mathop{\mathbb{E}}_{W\sim\mathcal{N}(0,\epsilon^T R^{-1}\epsilon)}[W^4] \\
&\lesssim (\epsilon^T R^{-1}\epsilon)^2
\end{aligned} \tag{8}$$

On the other hand, when $|W| \ge 1$,

$$|\alpha_\epsilon(Z_R)| \le e^{|W|}$$

$$\begin{aligned}
\mathop{\mathbb{E}}_{Z_R\sim\mathcal{N}(0,R)}\left[\alpha_\epsilon(Z_R)^2\cdot\mathbb{1}_{|\epsilon^TR^{-1}Z_r|\geq 1}\right] &\leq \mathop{\mathbb{E}}_{W\sim\mathcal{N}(0,\epsilon^TR^{-1}\epsilon)}[e^{2|W|}\mathbb{1}_{|W|\geq 1}] \\
&= 2\int_1^\infty \frac{1}{\sqrt{2\pi\epsilon^TR^{-1}\epsilon}}e^{2|w|}e^{-\frac{w^2}{2\epsilon^TR^{-1}\epsilon}}dw \\
&= 2e^{2|\epsilon^TR^{-1}\epsilon|}\int_1^\infty \frac{1}{\sqrt{2\pi\epsilon^TR^{-1}\epsilon}}e^{-\frac{(w-2|\epsilon^TR^{-1}\epsilon|)^2}{2\epsilon^TR^{-1}\epsilon}}dw \\
&\leq 2\sqrt{e}\mathop{\mathbb{P}}_{W\sim\mathcal{N}(0,\epsilon^TR^{-1}\epsilon)}\left[W\geq 1-2|\epsilon^TR^{-1}\epsilon|\right] \\
&\lesssim e^{-\frac{(1-2|\epsilon^TR^{-1}\epsilon|)^2}{2\epsilon^TR^{-1}\epsilon}} \\
&\leq e^{-\frac{1}{8\epsilon^TR^{-1}\epsilon}} \lesssim (\epsilon^TR^{-1}\epsilon)^2
\end{aligned} \tag{9}$$

which combines with (7) and (8) to give the claim. □

The next Lemma B.3 tells us that for good initial estimates $\epsilon\in\mathbb{R}^d$, "inverting the score" by left multiplying s_R by $-\mathcal{I}_R^{-1}$ provides a good estimate of ϵ in expectation.

Lemma B.3 (Score Inversion)**.** *Let f_R be an R-smoothed distribution with Fisher information matrix $\mathcal{I}_R$. Let $s_R:\mathbb{R}^d\to\mathbb{R}^d$ be the score function of f_R. Let $M\in\mathbb{R}^{d\times d}$ be a symmetric matrix such that $M\succcurlyeq 0$. Then, for any $\epsilon\in\mathbb{R}^d$ with $|\epsilon^TR^{-1}\epsilon|\leq 1/4$, we have*

$$\|\mathop{\mathbb{E}}_{x\sim f_R}[-\mathcal{I}_R^{-1}s_R(x+\epsilon)]-\epsilon\|_M^2\lesssim\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|(\epsilon^TR^{-1}\epsilon)^2$$

Proof. By definition of s_R,

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim f_R}[s_R(x+\epsilon)] &= \int_{[-\infty,\infty]^d} f_R(x)\frac{\nabla f_R(x+\epsilon)}{f_R(x+\epsilon)}dx \\
&= \int\nabla f_R(x)\left(\frac{f_R(x-\epsilon)-f_R(x)}{f_R(x)}\right)dx
\end{aligned}$$

since

$$\int\nabla f_R(x)dx=0$$

Now, by the definition of $\mathcal{I}_R$

$$\mathcal{I}_R=\mathop{\mathbb{E}}_{x\sim f_R}[s_R(x)s_R(x)^T]=\int_{[-\infty,\infty]^d}\frac{\nabla f_R(x)(\nabla f_R(x))^T}{f_R(x)}dx$$

So,

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim f_R}[s_R(x+\epsilon)]+\mathcal{I}_R\epsilon &= \int_{-[\infty,\infty]^d}\frac{\nabla f_R(x)}{f_R(x)}\left(f_R(x-\epsilon)-f_R(x)+(\nabla f_R(x))^T\epsilon\right)dx \\
&= \mathop{\mathbb{E}}_{x\sim f_R}[\Delta_{-\epsilon}(x)s_R(x)]
\end{aligned}$$

where $\Delta_\epsilon(x):=\frac{f_R(x+\epsilon)-f_R(x)-(\nabla f_R(x))^T\epsilon}{f_R(x)}$. Now, left multiplying both sides by $-M^{1/2}\mathcal{I}_R^{-1}$,

$$M^{1/2}\left(\mathop{\mathbb{E}}_{x\sim f_R}[-\mathcal{I}_R^{-1}s_R(x+\epsilon)]-\epsilon\right)=\mathop{\mathbb{E}}_{x\sim f_R}\left[\Delta_{-\epsilon}(x)(-M^{1/2}\mathcal{I}_R^{-1}s_R(x))\right]$$

So, we have

$$\|\mathop{\mathbb{E}}_{x\sim f_R}[-\mathcal{I}_R^{-1}s_R(x+\epsilon)]-\epsilon\|_M^2=\|\mathop{\mathbb{E}}_{x\sim f_R}[\Delta_{-\epsilon}(x)\left(-M^{1/2}\mathcal{I}_R^{-1}s_R(x)\right)]\|^2$$

Now, by Cauchy-Schwarz

$$\begin{aligned}
\| \mathop{\mathbb{E}}_{x\sim f_R}\left[\Delta_{-\epsilon}(x)(-M^{1/2}\mathcal{I}_R^{-1}s_R(x))\right]\|^2 &= \sup_{w\in S^{d-1}} \mathop{\mathbb{E}}_{x\sim f_R}[\Delta_{-\epsilon}(x)(-M^{1/2}\mathcal{I}_R^{-1}s_R(x))^T w]^2 \\
&\le \sup_{w\in S^{d-1}} \mathop{\mathbb{E}}_{x\sim f_R}[\Delta_{-\epsilon}(x)^2] \mathop{\mathbb{E}}_{x\sim f_R}\left[(-M^{1/2}\mathcal{I}_R^{-1}s_R(x))^T w)^2\right] \\
&= \mathop{\mathbb{E}}_{x\sim f_R}\left[\Delta_{-\epsilon}(x)^2\right] \| \mathop{\mathbb{E}}_{x\sim f_R}\left[(-M^{1/2}\mathcal{I}_R^{-1}s_R(x))(s_R(x)^T\mathcal{I}_R^{-1}M^{1/2})\right] \| \\
&= \mathop{\mathbb{E}}_{x\sim f_R}[\Delta_{-\epsilon}(x)^2]\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|
\end{aligned}$$

Using Lemma B.2, we finally have

$$\| \mathop{\mathbb{E}}_{x\sim f_R}[-\mathcal{I}_R^{-1}s_R(x+\epsilon)] - \epsilon\|_M^2 \lesssim \|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|(\epsilon^T R^{-1}\epsilon)^2$$

□

Next, in Lemma B.4 we prove a utility result that we make use of in Lemma B.5.

Lemma B.4. *Let f_R be the R-smoothed version of f on $\mathbb{R}^d$. For $\epsilon \in \mathbb{R}^d$, let*

$$\zeta_\epsilon(x) = \frac{f_R(x-\epsilon) - f_R(x)}{f_R(x)}$$

Then, for any ϵ such that $|\epsilon^T R^{-1}\epsilon| \le 1/4$, and for any α such that $\alpha^2(\epsilon^T R^{-1}\epsilon) \lesssim 1$ we have

$$\mathop{\mathbb{E}}_{x\sim f_R}[\zeta_\epsilon(x)^2] \lesssim (\epsilon^T R^{-1}\epsilon)(\alpha^2 e^{-\Omega(\alpha^2)} + e^{-\Omega(\alpha^2)})$$

Proof. By Lemma B.1, we have

$$\zeta_\epsilon(x) = \frac{f_R(x-\epsilon) - f_R(x)}{f_R(x)} = \mathop{\mathbb{E}}_{Z_R|x}\left[e^{-\epsilon^T R^{-1}Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon} - 1\right]$$

For the remaining proof, let $W = \epsilon^T R^{-1} Z_R$. Since $Z_R \sim \mathcal{N}(0, R)$, we have that $W \sim \mathcal{N}(0, \epsilon^T R^{-1}\epsilon)$. So, we have that

$$\zeta_\epsilon(x) = \mathop{\mathbb{E}}_{W|x}\left[e^{-W - \frac{1}{2}\epsilon^T R^{-1}\epsilon} - 1\right]$$

Let α be a parameter such that $\alpha^2(\epsilon^T R^{-1}\epsilon) \lesssim 1$. Now, we have

$$\zeta_\epsilon(x) \le O\left(\alpha\sqrt{\epsilon^T R^{-1}\epsilon}\right) + \mathop{\mathbb{E}}_{W|x}\left[\mathbb{1}_{|W|>\alpha\sqrt{\epsilon^T R^{-1}\epsilon}}\left(e^{-W} - 1\right)\right]$$

So,

$$\zeta_\epsilon(x)^2 \lesssim \alpha^2(\epsilon^T R^{-1}\epsilon) + \mathop{\mathbb{E}}_{W|x}\left[\mathbb{1}_{|W|>\alpha\sqrt{\epsilon^T R^{-1}\epsilon}}(e^{-W} - 1)\right]^2$$

Now, to bound the second term, by Jensen's inequality, we have

$$\mathop{\mathbb{E}}_{W|x}\left[\mathbb{1}_{|W|>\alpha\sqrt{\epsilon^T R^{-1}\epsilon}}(e^{-W} - 1)\right]^2 \le \mathop{\mathbb{E}}_{W|x}\left[\mathbb{1}_{|W|>\alpha\sqrt{\epsilon^T R^{-1}\epsilon}}(e^{-W} - 1)^2\right]$$

So, we have

$$\mathop{\mathbb{E}}_{x\sim f_R}\left[\zeta_\epsilon(x)^2\right] \lesssim \alpha^2(\epsilon^T R^{-1}\epsilon) + \mathop{\mathbb{E}}_{W}\left[\mathbb{1}_{|W|>\alpha\sqrt{\epsilon^T R^{-1}\epsilon}}(e^{-W} - 1)^2\right]$$

We will now bound the second term above, $\mathbb{E}_W\left[\mathbb{1}_{|W|>\alpha\sqrt{\epsilon^T R^{-1}\epsilon}}(e^{-W} - 1)^2\right]$, in two separate cases, when

1. $|W| \le 1$

2. $|W| > 1$

When $|W| \le 1$, by linear approximations to the exponential function, we have

$$(e^{-W} - 1)^2 \lesssim W^2$$

So,

$$\begin{aligned}\mathop{\mathbb{E}}_{W}\left[\mathbb{1}_{|W|>\alpha\sqrt{\epsilon^T R^{-1}\epsilon}}\mathbb{1}_{|W|\le 1}(e^{-W}-1)^2\right] &\lesssim \mathop{\mathbb{E}}_{W\sim\mathcal{N}(0,\epsilon^T R^{-1}\epsilon)}\left[\mathbb{1}_{|W|>\alpha\sqrt{\epsilon^T R^{-1}\epsilon}}\cdot W^2\right] \\ &\lesssim \alpha^2(\epsilon^T R^{-1}\epsilon)e^{-\Omega(\alpha^2)}\end{aligned}$$

On the other hand, when $|W| > 1$

$$\begin{aligned}&\mathop{\mathbb{E}}_{W}\left[\mathbb{1}_{|W|>\max(1,\alpha\sqrt{\epsilon^T R^{-1}\epsilon})}(e^{-W}-1)^2\right] \\ &\le \int_{-\infty}^{-(1+\alpha\sqrt{\epsilon^T R^{-1}\epsilon})} \frac{1}{\sqrt{2\pi\epsilon^T R^{-1}\epsilon}} e^{-\frac{w^2}{2\epsilon^T R^{-1}\epsilon}}(e^{-w}-1)^2 dw + \int_{1+\alpha\sqrt{\epsilon^T R^{-1}\epsilon}}^{\infty} \frac{1}{\sqrt{2\pi\epsilon^T R^{-1}\epsilon}} e^{-\frac{w^2}{2\epsilon^T R^{-1}\epsilon}}(e^{-w}-1)^2 dw \\ &\lesssim \int_{1+\alpha\sqrt{\epsilon^T R^{-1}\epsilon}}^{\infty} \frac{1}{\sqrt{2\pi\epsilon^T R^{-1}\epsilon}} e^{-\frac{w^2}{2\epsilon^T R^{-1}\epsilon}}(e^{w}-1)^2 dw \\ &\lesssim \int_{1+\alpha\sqrt{\epsilon^T R^{-1}\epsilon}}^{\infty} \frac{1}{\sqrt{2\pi\epsilon^T R^{-1}\epsilon}} e^{-\frac{w^2}{2\epsilon^T R^{-1}\epsilon}} e^{2w} dw \\ &= e^{2(\epsilon^T R^{-1}\epsilon)} \int_{1+\alpha\sqrt{\epsilon^T R^{-1}\epsilon}}^{\infty} \frac{1}{\sqrt{2\pi\epsilon^T R^{-1}\epsilon}} e^{-\frac{(w-2\epsilon^T R^{-1}\epsilon)^2}{2\epsilon^T R^{-1}\epsilon}} dw \\ &\lesssim e^{-\Omega\left(\frac{1}{\epsilon^T R^{-1}\epsilon}+\alpha^2\right)} \lesssim (\epsilon^T R^{-1}\epsilon)e^{-\Omega(\alpha^2)} \quad \text{since } |\epsilon^T R^{-1}\epsilon| \le 1/4\end{aligned}$$

Thus, we have shown that

$$\mathop{\mathbb{E}}_{x\sim f_R}\left[\zeta_\epsilon(x)^2\right] \lesssim \alpha^2(\epsilon^T R^{-1}\epsilon)e^{-\Omega(\alpha^2)} + (\epsilon^T R^{-1}\epsilon)e^{-\Omega(\alpha^2)}$$

The claim follows. □

The next Lemma B.5 shows that for small ϵ, the covariance of the appropriately transformed version of the shifted score $s_R(x+\epsilon)$ is similar to the corresponding transformation of the Fisher information matrix $\mathcal{I}_R$.

Lemma B.5. *Suppose f_R is a R-smoothed distribution on $\mathbb{R}^d$ with Fisher information matrix $\mathcal{I}_R$. Let $M \in \mathbb{R}^{d\times d}$ be a symmetric matrix such that $M \succcurlyeq 0$. Then for any $\epsilon \in \mathbb{R}^d$ with $|\epsilon^T R^{-1}\epsilon| \le 1/4$, we have, for every $v \in \mathbb{R}^d$ with $\|v\| = 1$,*

$$\begin{aligned}&\left| v^T \left(\mathop{\mathbb{E}}_{x\sim f_R}\left[M^{1/2}\mathcal{I}_R^{-1} s_R(x+\epsilon)s_R(x+\epsilon)^T \mathcal{I}_R^{-1} M^{1/2}\right] - M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\right) v \right| \\ &\lesssim \sqrt{\epsilon^T R^{-1}\epsilon}\cdot (v^T M^{1/2}\mathcal{I}_R^{-1}M^{1/2}v)\sqrt{\log\left(\sup_{w\in S^{d-1}} \frac{w^T R^{-1} w}{w^T \mathcal{I}_R w}\right)}\end{aligned}$$

Proof. We have, by definition of score,

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim f_R}[s_R(x+\epsilon)s_R(x+\epsilon)^T] &= \int_{[-\infty,\infty]^d} f_R(x)\frac{\nabla f_R(x+\epsilon)(\nabla f_R(x+\epsilon))^T}{f_R(x+\epsilon)^2}dx \\
&= \int f_R(x-\epsilon)\frac{\nabla f_R(x)(\nabla f_R(x))^T}{f_R(x)^2}dx \\
&= \mathcal{I}_R + \int (f_R(x-\epsilon)-f_R(x))\left(\frac{\nabla f_R(x)(\nabla f_R(x))^T}{f_R(x)^2}\right)dx \\
&= \mathcal{I}_R + \int \zeta_\epsilon(x)\left(\frac{\nabla f_R(x)(\nabla f_R(x))^T}{f_R(x)}\right)dx \\
&= \mathcal{I}_R + \mathop{\mathbb{E}}_{x\sim f_R}\left[\zeta_\epsilon(x)\frac{\nabla f_R(x)(\nabla f_R(x))^T}{f_R(x)^2}\right]
\end{aligned}$$

where $\zeta_\epsilon(x) = \frac{f_R(x-\epsilon)-f_R(x)}{f_R(x)}$. Now, since $s_R(x) = \frac{\nabla f_R(x)}{f_R(x)}$, the above is equivalent to

$$\mathop{\mathbb{E}}_{x\sim f_R}\left[s_R(x+\epsilon)s_R(x+\epsilon)^T\right] - \mathcal{I}_R = \mathop{\mathbb{E}}_{x\sim f_R}\left[\zeta_\epsilon(x)s_R(x)s_R(x)^T\right]$$

Left and right multiplying both sides by $M^{1/2}\mathcal{I}_R^{-1}$, this is

$$\mathop{\mathbb{E}}_{x\sim f_R}\left[M^{1/2}\mathcal{I}_R^{-1}s_R(x+\epsilon)s_R(x+\epsilon)^T\mathcal{I}_R^{-1}M^{1/2}\right] - M^{1/2}\mathcal{I}_R^{-1}M^{1/2} = M^{1/2}\mathcal{I}_R^{-1}\mathop{\mathbb{E}}_{x\sim f_R}\left[\zeta_\epsilon(x)s_R(x)s_R(x)^T\right]\mathcal{I}_R^{-1}M^{1/2}$$

Then, for $v\in\mathbb{R}^d$ with $\|v\|=1$

$$v^T\left(\mathop{\mathbb{E}}_{x\sim f_R}\left[M^{1/2}\mathcal{I}_R^{-1}s_R(x+\epsilon)s_R(x+\epsilon)^T\mathcal{I}_R^{-1}M^{1/2}\right] - M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\right)v = \mathop{\mathbb{E}}_{x\sim f_R}\left[\zeta_\epsilon(x)(v^TM^{1/2}\mathcal{I}_R^{-1}s_R(x))^2\right]$$

Then, using Cauchy-Schwarz,

$$\begin{aligned}
&\left|v^T\left(\mathop{\mathbb{E}}_{x\sim f_R}\left[M^{1/2}\mathcal{I}_R^{-1}s_R(x+\epsilon)s_R(x+\epsilon)^T\mathcal{I}_R^{-1}M^{1/2}\right] - M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\right)v\right| \\
&\le \sqrt{\mathop{\mathbb{E}}_{x\sim f_R}[\zeta_\epsilon(x)^2]\mathop{\mathbb{E}}_{x\sim f_R}\left[(v^TM^{1/2}\mathcal{I}_R^{-1}s_R(x))^4\right]}
\end{aligned} \tag{10}$$

To bound the second term inside the square root, recall that by Lemma B.1, we have

$$s_R(x) = \mathop{\mathbb{E}}_{Z_R|x}[R^{-1}Z_R]$$

So, by Jensen's inequality, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim f_R}[(v^TM^{1/2}\mathcal{I}_R^{-1}s_R(x))^4] &= \mathop{\mathbb{E}}_{x\sim f_R}\left[(v^TM^{1/2}\mathcal{I}_R^{-1}\mathop{\mathbb{E}}_{Z_R|x}[R^{-1}Z_R])^4\right] \\
&= \mathop{\mathbb{E}}_{x\sim f_R}\left[\mathop{\mathbb{E}}_{Z_R|x}[v^TM^{1/2}\mathcal{I}_R^{-1}R^{-1}Z_R]^4\right] \\
&\le \mathop{\mathbb{E}}_{x\sim f_R}\left[\mathop{\mathbb{E}}_{Z_R|x}[(v^TM^{1/2}\mathcal{I}_R^{-1}R^{-1}Z_R)^4]\right] \\
&= \mathop{\mathbb{E}}_{Z_R}\left[(v^TM^{1/2}\mathcal{I}_R^{-1}R^{-1}Z_R)^4\right]
\end{aligned}$$

Now, since $Z_R\sim\mathcal{N}(0,R)$, we have that $v^\top M^{1/2}\mathcal{I}_R^{-1}R^{-1}Z_R\sim\mathcal{N}(0, v^\top M^{1/2}\mathcal{I}_R^{-1}R^{-1}\mathcal{I}_R^{-1}M^{1/2}v)$ is a 1-dimensional Gaussian. Thus, using the standard fact about the 4^{th} moment of a 1-dimensional Gaussian, we have

$$\mathop{\mathbb{E}}_{x\sim f_R}\left[(v^TM^{1/2}\mathcal{I}_R^{-1}s_R(x))^4\right] \le \mathop{\mathbb{E}}_{Z_R}\left[(v^TM^{1/2}\mathcal{I}_R^{-1}R^{-1}Z_R)^4\right] = 3(v^TM^{1/2}\mathcal{I}_R^{-1}R^{-1}\mathcal{I}_R^{-1}M^{1/2}v)^2$$

For the first term under the square root in (10), by Lemma B.4, for any $\alpha \in \mathbb{R}$ such that $\alpha^2(\epsilon^T R^{-1}\epsilon) \lesssim 1$, we have

$$\mathop{\mathbb{E}}_{x\sim f_R}[\zeta_\epsilon(x)^2] \lesssim (\epsilon^T R^{-1}\epsilon)(\alpha^2 e^{-\Omega(\alpha^2)} + e^{-\Omega(\alpha^2)})$$

So, combining the above with (10), we have

$$\begin{aligned}
&\left| v^T \left(\mathop{\mathbb{E}}_{x\sim f_R}\left[M^{1/2}\mathcal{I}_R^{-1} s_R(x+\epsilon) s_R(x+\epsilon)^T \mathcal{I}_R^{-1} M^{1/2} \right] - M^{1/2}\mathcal{I}_R^{-1}M^{1/2} \right) v \right| \\
&\lesssim (v^T M^{1/2}\mathcal{I}_R^{-1} R^{-1} \mathcal{I}_R^{-1} M^{1/2} v) \sqrt{(\epsilon^T R^{-1}\epsilon)(\alpha^2 e^{-\Omega(\alpha^2)} + e^{-\Omega(\alpha^2)})} \\
&\lesssim (v^T M^{1/2}\mathcal{I}_R^{-1} R^{-1} \mathcal{I}_R^{-1} M^{1/2} v) \sqrt{\epsilon^T R^{-1}\epsilon}(\alpha e^{-\Omega(\alpha^2)})
\end{aligned}$$

Setting $\alpha = O\left(\sqrt{\log \frac{v^T M^{1/2}\mathcal{I}_R^{-1} R^{-1} \mathcal{I}_R^{-1} M^{1/2} v}{v^T M^{1/2}\mathcal{I}_R^{-1} M^{1/2} v}}\right)$ yields

$$\begin{aligned}
&\left| v^T \left(\mathop{\mathbb{E}}_{x\sim f_R}\left[M^{1/2}\mathcal{I}_R^{-1} s_R(x+\epsilon) s_R(x+\epsilon)^T \mathcal{I}_R^{-1} M^{1/2} \right] - M^{1/2}\mathcal{I}_R^{-1}M^{1/2} \right) v \right| \\
&\lesssim \sqrt{\epsilon^T R^{-1}\epsilon} \cdot (v^T M^{1/2}\mathcal{I}_R^{-1} M^{1/2} v) \sqrt{\log \frac{v^T M^{1/2}\mathcal{I}_R^{-1} R^{-1} \mathcal{I}_R^{-1} M^{1/2} v}{v^T M^{1/2}\mathcal{I}_R^{-1} M^{1/2} v}}
\end{aligned}$$

Since

$$\frac{v^T M^{1/2}\mathcal{I}_R^{-1} R^{-1} \mathcal{I}_R^{-1} M^{1/2} v}{v^T M^{1/2}\mathcal{I}_R^{-1} M^{1/2} v} \le \sup_{w\in S^{d-1}} \frac{w^T R^{-1} w}{w^T \mathcal{I}_R w}$$

the claim follows. □

B.3. Subgamma concentration of score

In this section, we establish that every one-dimensional projection of the score function s_R after applying a symmetric PSD linear transformation is subgamma with appropriate variance and scale parameters. We begin by showing a bound on the Jacobian of the score, which we make use of in future lemmas.

Lemma B.6. *Let $s_R : \mathbb{R}^d \to \mathbb{R}^d$ be the score function of f_R, the R-smoothed version of distribution f. Let $\boldsymbol{J}_{s_R}$ be the Jacobian of s_R. We have that*

$$\boldsymbol{J}_{s_R} \succcurlyeq -R^{-1}$$

Proof. Taking the gradient in Lemma B.1 wrt ϵ, we have

$$\frac{\nabla f_R(x+\epsilon)}{f_R(x)} = \mathop{\mathbb{E}}_{Z_R|x}\left[e^{\epsilon^T R^{-1} Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon} \left(R^{-1}Z_R - R^{-1}\epsilon \right) \right]$$

So,

$$s_R(x+\epsilon) = \frac{\nabla f_R(x+\epsilon)}{f_R(x+\epsilon)} \cdot \frac{f_R(x+\epsilon)}{f_R(x)} = \frac{\mathbb{E}_{Z_R|x}\left[e^{\epsilon^T R^{-1} Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon} \left(R^{-1}Z_R - R^{-1}\epsilon \right) \right]}{\mathbb{E}_{Z_R|x}\left[e^{\epsilon^T R^{-1} Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon} \right]}$$

Now, let $\epsilon = \gamma v$ for $\gamma \in \mathbb{R}$, $\gamma > 0$ so that $\|v\| = 1$. Now, $e^{\epsilon^T R^{-1} Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon}$ and $v^T R^{-1} Z_R - v^T R^{-1}\epsilon$ are monotonically non-decreasing in $v^T R^{-1} Z_R$. So, by Lemma C.1, they are positively correlated. That is,

$$v^T \mathop{\mathbb{E}}_{Z_R|x}\left[e^{\epsilon^T R^{-1} Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon} \left(R^{-1}Z_R - R^{-1}\epsilon \right) \right] \ge \mathop{\mathbb{E}}_{Z_R|x}\left[e^{\epsilon^T R^{-1} Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon} \right] \cdot \left(v^T \mathop{\mathbb{E}}_{Z_R|x}\left[R^{-1}Z_R - R^{-1}\epsilon \right] \right)$$

So,

$$v^T s_R(x+\epsilon) \geq v^T \mathop{\mathbb{E}}_{Z_R|x}\left[R^{-1}Z_R - R^{-1}\epsilon\right] \tag{11}$$

Now, by definition of Jacobian

$$\mathbf{J}_{s_R} v = \left[\frac{\partial}{\partial\gamma} s_R(x+\gamma v)\right]_{\gamma=0}$$

So, in (11), taking the derivative wrt γ and setting $\gamma = 0$, we get

$$v^T \mathbf{J}_{s_R} v \geq -v^T R^{-1} v$$

as required. □

The next lemma shows that every 1-dimensional projection of the score $s_R(x)$ is subgamma with appropriate variance and scale parameters. As a corollary (Corollary B.8) we obtain that every 1-dimensional projection of the score when transformed using a symmetric PSD matrix is also subgamma, with appropriately transformed variance and scale.

Lemma B.7. *Let $s_R : \mathbb{R}^d \to \mathbb{R}^d$ be the score function of an R-smoothed distribution f_R with Fisher information matrix $\mathcal{I}_R$. For any fixed $v \in \mathbb{R}^d$ with $\|v\| = 1$, we have*

$$\mathop{\mathbb{E}}_{x\sim f_R}[|v^T R^{1/2} s_R(x)|^k] \leq (1.6)^{k-2} k^{k/2} (v^T R^{1/2} \mathcal{I}_R R^{1/2} v)$$

Equivalently, for any $v \in \mathbb{R}^d$, $v^T R^{1/2} s_R(x)$ is a subgamma random variable.

$$v^T R^{1/2} s_R(x) \in \Gamma(v^T R^{1/2} \mathcal{I}_R R^{1/2} v, 1.6\|v\|)$$

Proof. For $x, \gamma \in \mathbb{R}^d$, by Lemma B.1, and Jensen's inequality,

$$f_R(x+\gamma) \geq f_R(x) e^{\gamma^T s_R(x) - \frac{1}{2}\gamma^T R^{-1}\gamma}$$

Set $\gamma = R^{1/2} v$. Then,

$$f_R\left(x + \gamma \cdot \text{sign}(\gamma^T s_R(x))\right) \geq f_R(x) e^{|\gamma^T s_R(x)|}/\sqrt{e}$$

Now, by Lemma B.6, we have,

$$\begin{aligned}
\gamma^T s_R(x+\gamma) &= \gamma^T s_R(x) + \gamma^T \mathbf{J}_{s_R}\gamma \\
&= \gamma^T s_R(x) + v^T R^{1/2} \mathbf{J}_{s_R} R^{1/2} v \\
&\geq \gamma^T s_R(x) - 1
\end{aligned}$$

Similarly,

$$\gamma^T s_R(x-\gamma) \leq \gamma^T s_R(x) + 1$$

Combining these two, we have

$$|\gamma^T s_R(x + \gamma \cdot \text{sign}(\gamma^T s_R(x))| \geq |\gamma^T s_R(x)| - 1$$

So, for any $k \geq 2$, and $|\gamma^T s_R(x)| > \alpha$ for $\alpha := 2 + 1.2\sqrt{k}$

$$\begin{aligned}
&f_R(x + \gamma \cdot \text{sign}(\gamma^T s_R(x)))|\gamma^T s_R(x + \gamma \cdot \text{sign}(\gamma^T s_R(x)))|^k \\
&\geq \frac{1}{\sqrt{e}} f_R(x) e^{|\gamma^T s_R(x)|} \left(|\gamma^T s_R(x)| - 1\right)^k \\
&= f_R(x)|\gamma^T s_R(x)|^k \cdot \left(\frac{1}{\sqrt{e}} e^{|\gamma^T s_R(x)|}\left(1 - \frac{1}{|\gamma^T s_R(x)|}\right)^k\right) \\
&\geq f_R(x)|\gamma^T s_R(x)|^k \cdot \left(\frac{1}{\sqrt{e}} e^{\alpha - 1.4\frac{k}{\alpha}}\right) \\
&\geq f_R(x)|\gamma^T s_R(x)|^k \cdot 4
\end{aligned}$$

Thus,

$$f_R(x)|\gamma^T s_R(x)|^k \le \frac{1}{4}\left(f_R(x-\gamma)|\gamma^T s_R(x-\gamma)|^k + f_R(x+\gamma)|\gamma^T s_R(x+\gamma)|^k\right)$$

when $k \ge 2$ and $|\gamma^T s_R(x)| \ge \alpha$. Integrating this,

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim f_R}\left[|\gamma^T s_R(x)|^k\right] &= \int_{[-\infty,\infty]^d} f_R(x)|\gamma^T s_R(x)|^k dx \\
&\le 2\int f_R(x)|\gamma^T s_R(x)|^k - \frac{1}{4}f_R(x-\gamma)|\gamma^T s_R(x-\gamma)|^k - \frac{1}{4}f_R(x+\gamma)|\gamma^T s_R(x+\gamma)|dx \\
&\le 2\int f_R(x)|\gamma^T s_R(x)|^k \mathbb{1}_{|\gamma^T s_R(x)|<\alpha} dx \\
&\le 2\int f_R(x)|\gamma^T s_R(x)|^2 \alpha^{k-2} \mathbb{1}_{|\gamma^T s_R(x)|<\alpha} dx \\
&\le 2\alpha^{k-2}\,\mathbb{E}[|\gamma^T s_R(x)|^2] = 2\alpha^{k-2}\gamma^T \mathcal{I}_R \gamma
\end{aligned}$$

Finally, for any $k \ge 2$,

$$2\alpha^{k-2} = 2(1.2\sqrt{k}+2)^{k-2} \le k^{k/2}\cdot 1.6^{k-2}$$

The claim follows. □

Corollary B.8. *Let $s_R : \mathbb{R}^d \to \mathbb{R}^d$ be the score function of an R-smoothed distribution f_R with Fisher information matrix $\mathcal{I}_R$. Let $M \in \mathbb{R}^{d\times d}$ be a symmetric matrix such that $M \succcurlyeq 0$. For any fixed $v \in \mathbb{R}^d$ with $\|v\| = 1$, we have*

$$\mathop{\mathbb{E}}_{x\sim f_R}[|v^T M^{1/2}\mathcal{I}_R^{-1} s_R(x)|^k] \le (1.6\|M^{1/2}\mathcal{I}_R^{-1}R^{-1/2}v\|)^{k-2}k^{k/2}(v^T M^{1/2}\mathcal{I}_R^{-1}M^{1/2}v)$$

Equivalently, $v^T M^{1/2}\mathcal{I}_R^{-1}s_R(x)$ is subgamma.

$$|v^T M^{1/2}\mathcal{I}_R^{-1}s_R(x)| \in \Gamma(v^T M^{1/2}\mathcal{I}_R^{-1}M^{1/2}v, 1.6\|M^{1/2}\mathcal{I}_R^{-1}R^{-1/2}v\|)$$

Lemmas B.9 and B.10 proved next are helper lemmas that we make use of to prove the main result of this section, Lemma B.11, which shows that every one dimensional projection of $s_R(x+\epsilon)$ for $x \sim f_R$ is subgamma.

Lemma B.9. *Let $s_R : \mathbb{R}^d \to \mathbb{R}^d$ be the score function of an R-smoothed distribution f_R with Fisher information matrix $\mathcal{I}_R$. For any fixed $v \in \mathbb{R}^d$ with $\|v\| = 1$, $x \in \mathbb{R}^d$, $k \ge 3$, and $\epsilon \in \mathbb{R}^d$ with $0 \le \epsilon^T R^{-1}\epsilon \le 1/4$, if $v^T R^{1/2} s_R(x+\epsilon) \ge \max(2\sqrt{k}+2, 9.5)$, then, for $\gamma = R^{1/2}v$,*

$$f_R(x)|\gamma^T s_R(x+\epsilon)|^k \le \frac{1}{5}\max\left(f_R(x-\epsilon)|\gamma^T s_R(x-\epsilon)|^k, f_R(x+\epsilon+\gamma)|\gamma^T s_R(x+\epsilon+\gamma)|^k\right)$$

Proof. Let $\alpha := \frac{f_R(x)}{f_R(x+\epsilon)}$. By Lemma B.1, we have

$$\alpha = \mathop{\mathbb{E}}_{Z_R|x+\epsilon}\left[e^{-\epsilon^T R^{-1}Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon}\right] \tag{12}$$

Let $\gamma = R^{1/2}v$. We will consider two cases

When $\log\alpha < \frac{3}{4}\gamma^T s_R(x+\epsilon) - 2$. First, by Lemma B.1 and Jensen's inequality, we have

$$\frac{f_R(x+\epsilon+\gamma)}{f_R(x+\epsilon)} \ge e^{\gamma^T s_R(x+\epsilon) - 1/2}$$

Also, by Lemma B.6, we have

$$\gamma^T s_R(x+\epsilon+\gamma) \ge \gamma^T s_R(x+\epsilon) - 1$$

So,

$$\begin{aligned}
f_R(x+\epsilon+\gamma)|\gamma^T s_R(x+\epsilon+\gamma)|^k &\ge f_R(x+\epsilon)|\gamma^T s_R(x+\epsilon)|^k e^{\gamma^T s_R(x+\epsilon)-\frac{1}{2}}\left(1-\frac{1}{\gamma^T(s_R(x+\epsilon)}\right)^k \\
&\ge f_R(x+\epsilon)|\gamma^T s_R(x+\epsilon)|^k e^{\gamma^T s_R(x+\epsilon) - \frac{k}{\gamma^T s_R(x+\epsilon)-1} - \frac{1}{2}}
\end{aligned}$$

Since $\gamma^T s_R(x+\epsilon) \geq 2\sqrt{k}+2$,

$$f_R(x+\epsilon+\gamma)|s_R(x+\epsilon+\gamma)|^k \geq f_R(x+\epsilon)|s_R(x+\epsilon)|^k e^{\frac{3}{4}\gamma^T s_R(x+\epsilon)}$$

So, since

$$\alpha = \frac{f_R(x)}{f_R(x+\epsilon)} \leq e^{\frac{3}{4}\gamma^T s_R(x+\epsilon)-2}$$

we have

$$f(x)|s_R(x+\epsilon)|^k = \alpha f_R(x+\epsilon)|s_R(x+\epsilon)|^k \leq \frac{1}{5} f_R(x+\epsilon+\gamma)|s_R(x+\epsilon+\gamma)|^k$$

When $\log\alpha > \frac{3}{4}\gamma^T s_R(x+\epsilon) - 2$**.** Evaluating (12) at $x-\epsilon$ gives

$$\frac{f_R(x-\epsilon)}{f_R(x)} = \mathop{\mathbb{E}}_{Z_R|x}\left[e^{-\epsilon^T R^{-1} Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon}\right]$$

Taking the gradient wrt ϵ, we have

$$\frac{\nabla f_R(x-\epsilon)}{f_R(x)} = \mathop{\mathbb{E}}_{Z_R|x}\left[R^{-1}(Z_R+\epsilon)e^{-\epsilon^T R^{-1} Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon}\right]$$

so evaluating at $x+\epsilon$,

$$\frac{\nabla f_R(x)}{f_R(x+\epsilon)} = \mathop{\mathbb{E}}_{Z_R|x+\epsilon}\left[R^{-1}(Z_R+\epsilon)e^{-\epsilon^T R^{-1} Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon}\right]$$

In particular,

$$\epsilon^T \frac{\nabla f_R(x)}{f_R(x+\epsilon)} = \mathop{\mathbb{E}}_{Z_R|x+\epsilon}\left[\epsilon^T R^{-1}(Z_R+\epsilon)e^{-\epsilon^T R^{-1} Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon}\right]$$

Define $y = e^{-\epsilon^T R^{-1} Z_R - \epsilon^T R^{-1}\epsilon}$ so that $\mathbb{E}_{Z_R|x+\epsilon}[y] = \alpha e^{-\frac{1}{2}\epsilon^T R^{-1}\epsilon}$, and

$$\epsilon^T R^{-1}(Z_R+\epsilon)e^{-\epsilon^T R^{-1} Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon} = -e^{\frac{1}{2}\epsilon^T R^{-1}\epsilon} y\log y$$

is concave, so by Jensen's inequality,

$$\epsilon^T \frac{\nabla f_R(x)}{f_R(x+\epsilon)} \leq -e^{\frac{1}{2}\epsilon^T R^{-1}\epsilon}\left(e^{-\frac{1}{2}\epsilon^T R^{-1}\epsilon}\alpha\right)\log\left(e^{-\frac{1}{2}\epsilon^T R^{-1}\epsilon}\alpha\right) = -\alpha\log\alpha + \frac{1}{2}\alpha\epsilon^T R^{-1}\epsilon$$

So,

$$\epsilon^T s_R(x) = \epsilon^T \frac{\nabla f_R(x)}{f_R(x)} \leq -\log\alpha + \frac{1}{2}\epsilon^T R^{-1}\epsilon$$

Finally we consider the move to $x-\epsilon$. By Lemma B.6, we have

$$\epsilon^T s_R(x-\epsilon) \leq s_R(x) + \epsilon^T R^{-1}\epsilon \leq -\log\alpha + \frac{3}{2}\epsilon^T R^{-1}\epsilon$$

By Lemma B.1,

$$\frac{f_R(x-\epsilon)}{f_R(x+\epsilon)} = \mathop{\mathbb{E}}_{Z_R|x+\epsilon}\left[e^{-2\epsilon^T R^{-1} Z_R - 2\epsilon^T R^{-1}\epsilon}\right] = \mathop{\mathbb{E}}_{Z_R|x+\epsilon}[y^2] \geq \mathop{\mathbb{E}}_{Z_R|x+\epsilon}[y]^2 = \alpha^2 e^{-\epsilon^T R^{-1}\epsilon}$$

Since $\log\alpha \geq \frac{3}{4}\gamma^T s_R(x+\epsilon) - 2$,

$$-\epsilon^T s_R(x-\epsilon) \geq \frac{3}{4}\gamma^T s_R(x+\epsilon) - 2 - \frac{3}{2}\epsilon^T R^{-1}\epsilon \geq \frac{3}{4}\gamma^T s_R(x+\epsilon) - \frac{19}{8} \geq \gamma^T s_R(x)$$

where the second inequality comes from the fact that $\frac{3}{4}\gamma^T s_R(x+\epsilon) - 2 > 0$, so that the function is decreasing in $\epsilon^T R^{-1}\epsilon$, and $\epsilon^T R^{-1}\epsilon \leq 1/4$. Thus,

$$f_R(x-\epsilon)|\gamma^T s_R(x-\epsilon)|^k \geq \alpha e^{-\epsilon^T R^{-1}\epsilon} f_R(x)|s_R(x+\epsilon)|^k$$

Since our assumptions give $\alpha e^{-\epsilon^T R^{-1}\epsilon} \geq 5$, we get the result. □

Lemma B.10. *Let $s_R : \mathbb{R}^d \to \mathbb{R}^d$ be the score function of an R-smoothed distribution f_R with Fisher information matrix $\mathcal{I}_R$.*

For any fixed $v \in \mathbb{R}^d$ with $\|v\| = 1$, $x \in \mathbb{R}^d, k \geq 3$ and $\epsilon \in \mathbb{R}^d$ with $1/4 \leq \epsilon^T R^{-1} \epsilon \leq 0$, if $v^T R^{1/2} s_R(x+\epsilon) \geq \alpha$ for $\alpha = 2 + 1.2\sqrt{k}$, then we have for $\gamma = R^{1/2} v$,

$$f_R(x)|\gamma^T s_R(x+\epsilon)|^k \leq \frac{1}{4}\left(f_R(x-\gamma)|\gamma^T s_R(x+\epsilon-\gamma)|^k + f_R(x+\gamma)|s_R(x+\epsilon+\gamma)|^k\right)$$

As an immediate corollary, the statement is also true when $0 \leq \epsilon^T R^{-1}\epsilon \leq 1/4$ and $v^T R^{1/2} s_R(x) \leq -\alpha$.

Proof. By Lemma B.1 and Jensen's inequality,

$$f_R(x+\gamma) \geq f_R(x) e^{\gamma^T s_R(x)}/\sqrt{e}$$

By Lemma B.6, we have that

$$\gamma^T s_R(x+\epsilon+\gamma) \geq \gamma^T s_R(x+\epsilon) - 1$$

Since the right hand side is positive by assumption, we have

$$|\gamma^T s_R(x+\epsilon+\gamma)| \geq |\gamma^T s_R(x+\epsilon)| - 1$$

Now, when $\epsilon^T R^{-1}\epsilon < 0$, we have by Lemma B.6, and since $|\epsilon^T R^{-1}\epsilon| \leq 1$ that

$$\gamma^T s_R(x) \geq \gamma^T s_R(x+\epsilon) - 1$$

So,

$$\begin{aligned}
f_R(x+\gamma)|\gamma^T s_R(x+\epsilon+\gamma)|^k &\geq \frac{1}{\sqrt{e}} f_R(x) e^{\gamma^T s_R(x)} \left(|\gamma^T s_R(x+\epsilon)| - 1\right)^k \\
&\geq \frac{1}{\sqrt{e}} f_R(x) e^{\gamma^T s_R(x+\epsilon) - 1} \left(|\gamma^T s_R(x+\epsilon)| - 1\right)^k \\
&\geq f_R(x)|\gamma^T s_R(x+\epsilon)|^k \left(\frac{1}{\sqrt{e}} e^{\gamma^T s_R(x+\epsilon)-1}\left(1 - \frac{1}{|\gamma^T s_R(x+\epsilon)|}\right)^k\right) \\
&\geq f_R(x)|\gamma^T s_R(x+\epsilon)|^k \cdot \left(e^{-3/2} e^{\alpha - 1.4k/\alpha}\right) \\
&\geq f_R(x)|\gamma^T s_R(x+\epsilon)|^k \cdot 4
\end{aligned}$$

□

We are now ready to prove that every 1-dimensional projection of $s_R(x+\epsilon)$ for $x \sim f_R$ is subgamma with appropriate variance and scale. As a corollary (Corollary B.12), we obtain that every 1-dimensional projection of $s_R(x+\epsilon)$ when transformed by applying a symmetric PSD matrix is also subgamma, with appropriately transformed variance and scale.

Lemma B.11. *Let s_R be the score function of an R-smoothed distribution f_R with Fisher information matrix $\mathcal{I}_R$. For $k \geq 3$ and $\epsilon \in \mathbb{R}^d$ such that $|\epsilon^T R^{-1}\epsilon| \leq 1/4$, we have that for any $v \in \mathbb{R}^d$ with $\|v\| = 1$,*

$$\mathop{\mathbb{E}}_{x \sim f_R}\left[|v^T R^{1/2} s_R(x+\epsilon)|^k\right] \leq (15)^{k-2} k^{k/2} \max\left(\mathop{\mathbb{E}}_{x \sim f_R}[v^T R^{1/2} s_R(x+\epsilon) s_R(x+\epsilon)^T R^{1/2} v], v^T R^{1/2} \mathcal{I}_R R^{1/2} v\right)$$

Equivalently, $v^T R^{1/2} s_R(x+\epsilon)$ is a subgamma random variable.

$$v^T R^{1/2} s_R(x+\epsilon) \in \Gamma\left(\max\left(\mathop{\mathbb{E}}_{x \sim f_R}[v^T R^{1/2} s_R(x+\epsilon) s_R(x+\epsilon)^T R^{1/2} v], v^T R^{1/2} \mathcal{I}_R R^{1/2} v\right), 15\right)$$

Proof. Without loss of generality, we only show the $\epsilon^T R^{-1}\epsilon \geq 0$ case. As before, let $\gamma = R^{1/2}v$. Using Lemma B.9 and Lemma B.7, we have

$$\begin{aligned}
&\int_{[-\infty,\infty]^d} f_R(x-\epsilon)|\gamma^T s_R(x)|^k \mathbb{1}_{\gamma^T s_R(x) > \max(2\sqrt{k}+2, 9.5)} dx \\
&\leq \int_{[-\infty,\infty]^d} \frac{1}{5}\max\left(f_R(x-2\epsilon)|\gamma^T s_R(x-2\epsilon)|^k, f_R(x+\gamma)|\gamma^T s_R(x+\gamma)|^k\right) dx \\
&= \frac{2}{5}\mathop{\mathbb{E}}_{x\sim f_R}[|\gamma^T s_R(x)|^k] \\
&\leq \frac{2}{5}(1.6)^{k-2}k^{k/2}(\gamma^T \mathcal{I}_R \gamma)
\end{aligned}$$

Then, we can start bounding the k^{th} moment quantity in the lemma. Using Lemma B.10, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x\sim f_R}\left[|\gamma^T s_R(x+\epsilon)|^k\right] = \int_{[-\infty,\infty]^d} f_R(x-\epsilon)|\gamma^T s_R(x)|^k dx \\
&= 2\int f_R(x-\epsilon)|\gamma^T s_R(x)|^k - \frac{1}{4}f_R(x-\epsilon-\gamma)|\gamma^T s_R(x-\gamma)|^k - \frac{1}{4}f_R(x-\epsilon+\gamma)|\gamma^T s_R(x+\gamma)|^k dx \\
&\leq \int f_R(x-\epsilon)|\gamma^T s_R(x)|^k \mathbb{1}_{\gamma^T s_R(x) \geq -\max(2\sqrt{k}+2, 9.5)} dx
\end{aligned}$$

Now, using the previous claim, we get

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x\sim f_R}\left[|\gamma^T s_R(x+\epsilon)|^k\right] \\
&\leq 2\int f_R(x-\epsilon)|\gamma^T s_R(x)|^k \mathbb{1}_{|\gamma^T s_R(x)| \leq \max(2\sqrt{k}+2, 9.5)} dx + \frac{4}{5}(1.6)^{k-2}k^{k/2}(\gamma^T \mathcal{I}_R \gamma) \\
&\leq 2\int f_R(x-\epsilon)|\gamma^T s_R(x)|^2 (\max(2\sqrt{k}+2, 9.5))^{k-2} \mathbb{1}_{|\gamma^T s_R(x)| \leq \max(2\sqrt{k}+2, 9.5)} dx + \frac{4}{5}(1.6)^{k-2}k^{k/2}(\gamma^T \mathcal{I}_R \gamma) \\
&\leq 2\max(2\sqrt{k}+2, 9.5)^{k-2} \mathop{\mathbb{E}}_{x\sim f_R}[|\gamma^T s_R(x+\epsilon)|^2] + \frac{4}{5}(1.6)^{k-2}k^{k/2}(\gamma^T \mathcal{I}_R \gamma) \\
&\leq 2k^{k/2}(2.5)^{k-2} \mathop{\mathbb{E}}_{x\sim f_R}[|\gamma^T s_R(x+\epsilon)|^2] + \frac{4}{5}(1.6)^{k-2}k^{k/2}(\gamma^T \mathcal{I}_R \gamma) \\
&\leq 3k^{k/2}(2.5)^{k-2}\max(\mathop{\mathbb{E}}_{x\sim f_R}[|\gamma^T s_R(x+\epsilon)|^2], \gamma^T \mathcal{I}_R \gamma) \\
&\leq k^{k/2}(15)^{k-2}\max\left(\mathop{\mathbb{E}}_{x\sim f_R}[\gamma^T s_R(x+\epsilon)s_R(x+\epsilon)^T\gamma], \gamma^T \mathcal{I}_R \gamma\right)
\end{aligned}$$

as required. □

Corollary B.12. *Let s_R be the score function of an R-smoothed distribution f_R with Fisher information matrix $\mathcal{I}_R$. Let $M \in \mathbb{R}^{d\times d}$ be a symmetric matrix such that $M \succcurlyeq 0$. For $k \geq 3$ and $\epsilon \in \mathbb{R}^d$ such that $|\epsilon^T R^{-1}\epsilon| \leq 1/4$, we have that for any $v \in \mathbb{R}^d$ with $\|v\| = 1$,*

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x\sim f_R}\left[|v^T M^{1/2}\mathcal{I}_R^{-1}s_R(x+\epsilon)|^k\right] \\
&\leq (15\|M^{1/2}\mathcal{I}_R^{-1}R^{-1/2}v\|)^{k-2}k^{k/2}v^T\left(M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\left(1+O\left(\sqrt{\epsilon^T R^{-1}\epsilon}\sqrt{\log\sup_{w\in S^{d-1}}\frac{w^T R^{-1}w}{w^T\mathcal{I}_R w}}\right)\right)\right)v
\end{aligned}$$

In other words,

$$M^{1/2}\mathcal{I}_R^{-1}s_R(x+\epsilon) \in \Gamma\left(M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\left(1+O\left(\sqrt{\epsilon^T R^{-1}\epsilon}\sqrt{\log\sup_{w\in S^{d-1}}\frac{w^T R^{-1}w}{w^T\mathcal{I}_R w}}\right)\right), M^{1/2}\mathcal{I}_R^{-1}R^{-1/2}\right)$$

Proof. By the Lemma,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x\sim f_R}\left[|v^T M^{1/2}\mathcal{I}_R^{-1}s_R(x+\epsilon)|^k\right] \\
&\leq (15\|M^{1/2}\mathcal{I}_R^{-1}R^{-1/2}v\|)^{k-2}k^{k/2} \\
&\max\left(\mathop{\mathbb{E}}_{x\sim f_R}\left[v^T M^{1/2}\mathcal{I}_R^{-1}s_R(x+\epsilon)s_R(x+\epsilon)^T\mathcal{I}_R^{-1}M^{1/2}v\right], v^T M^{1/2}\mathcal{I}_R^{-1}M^{1/2}v\right)
\end{aligned}$$

Then, using Lemma B.5, the claim follows. □

B.4. Estimation of inverted score

In this section, we use the subgamma bound on 1-dimensional projections of $s_R(x+\epsilon)$ for $x \sim f_R$ from Corollary B.12, as well as our norm concentration bound for subgamma vectors from Theorem 5.1 to establish a bound on the deviation of our inverted empirical score at $x+\epsilon$ from its expectation.

Lemma B.13. *Let f be an arbitrary distribution on $\mathbb{R}^d$ and let f_R be the R-smoothed version of f. Let $\mathcal{I}_R$ be the Fisher information matrix of f_R. Let $\epsilon \in \mathbb{R}^d$ be such that $\epsilon^T R^{-1}\epsilon \leq 1/4$. Consider the parametric family of distributions $f_R^\lambda(x) = f_R(x-\lambda)$. Suppose we have n i.i.d. samples $x_1, \ldots, x_n \sim f_R^\lambda$. Let $M \in \mathbb{R}^{d\times d}$ be a symmetric matrix with $M \succcurlyeq 0$. Let $\hat{\epsilon} = \frac{1}{n}\sum_{i=1}^n \mathcal{I}_R^{-1}s_R(x_i-\lambda-\epsilon)$. Let*

$$T := M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\left(1 + O\left(\sqrt{\epsilon^T R^{-1}\epsilon}\sqrt{\log \sup_{w\in S^{d-1}} \frac{w^T R^{-1} w}{w^T \mathcal{I}_R w}}\right)\right)$$

Then, with probability $1-\delta$, we have

$$\begin{aligned}
&\|\hat{\epsilon} - \mathop{\mathbb{E}}_{x\sim f_R}[\mathcal{I}_R^{-1}s_R(x-\epsilon)]\|_M \\
&\leq \sqrt{\frac{\mathrm{Tr}(T)}{n}} + 4\sqrt{\frac{\|T\|\log\frac{2}{\delta}}{n}} + 16\frac{\|M^{1/2}\mathcal{I}_R^{-1}R^{-1/2}\|\log\frac{2}{\delta}}{n} + 8\frac{\|M^{1/2}\mathcal{I}_R^{-1}R^{-1/2}\|_F^2}{n^{3/2}\sqrt{\mathrm{Tr}(T)}}\log\frac{2}{\delta}
\end{aligned}$$

Proof. By Corollary B.12, $M^{1/2}\mathcal{I}_R^{-1}s_R(x)$ is $(T, M^{1/2}\mathcal{I}_R^{-1}R^{-1/2})$-subgamma. Then, applying our subgamma norm concentration bound from Theorem 5.1 gives

$$\begin{aligned}
&\|\hat{\epsilon} - \mathop{\mathbb{E}}_{x\sim f_R}\left[\mathcal{I}_R^{-1}s_R(x-\epsilon)\right]\|_M \\
&= \left\|M^{1/2}\left(\frac{1}{n}\sum_{i=1}^n \mathcal{I}_R^{-1}s_R(x_i-\lambda-\epsilon)\right) - M^{1/2}\mathop{\mathbb{E}}_{x\sim f_R}\left[\mathcal{I}_R^{-1}s_R(x-\epsilon)\right]\right\| \\
&= \left\|\left(\frac{1}{n}\sum_{i=1}^n M^{1/2}\mathcal{I}_R^{-1}s_R(x_i-\lambda-\epsilon)\right) - \mathop{\mathbb{E}}_{x\sim f_R}\left[M^{1/2}\mathcal{I}_R^{-1}s_R(x-\epsilon)\right]\right\| \\
&\leq \sqrt{\frac{\mathrm{Tr}(T)}{n}} + 4\sqrt{\frac{\|T\|\log\frac{2}{\delta}}{n}} + 16\frac{\|M^{1/2}\mathcal{I}_R^{-1}R^{-1/2}\|\log\frac{2}{\delta}}{n} + 8\frac{\|M^{1/2}\mathcal{I}_R^{-1}R^{-1/2}\|_F^2}{n^{3/2}\sqrt{\mathrm{Tr}(T)}}\log\frac{2}{\delta}
\end{aligned}$$

□

B.5. Local MLE

In this section, we show how to estimate our location λ at rate that depends on $\mathcal{I}_R$ when given samples from f_λ, along with an initial uncertainty region S that is guaranteed to contain λ.

Algorithm 3 High-dimensional Local MLE

Input Parameters:

- Description of distribution f on $\mathbb{R}^d$, smoothing R, samples $x_1, \ldots, x_n \overset{i.i.d.}{\sim} f^\lambda$, and initial estimate λ_1

1. Let $\mathcal{I}_R$ be the Fisher information matrix of f_R, the R-smoothed version of f. Let s_R be the score function of f_R.
2. For each sample x_i, compute a perturbed sample $x_i' = x_i + \mathcal{N}(0, R)$ where all the Gaussian noise are drawn independently across all the samples.
3. Let $\hat{\epsilon} = \frac{1}{n}\sum_{i=1}^n \mathcal{I}_R^{-1} s_R(x_i' - \lambda_1)$ and return $\hat{\lambda} = \lambda_1 - \hat{\epsilon}$.

Lemma B.14 (Local MLE)**.** *Suppose we have a known model f on $\mathbb{R}^d$, and that f_R is the R-smoothed version of f, for $R = r^2 I_d$ for scalar $r > 0$. Suppose f_R has Fisher information matrix $\mathcal{I}_R$. Further, suppose that the unknown true parameter is λ, and that we have access to an initial estimate $\lambda_1 = \lambda + \epsilon$ with the guarantee that $\epsilon^T R^{-1} \epsilon \le \tau$ for $\tau \le 1/4$. Suppose there exists a large parameter $\gamma \ge 1$ such that $\tau \le \frac{1}{\gamma^2 \log^2 \frac{\|\mathcal{I}_R^{-1}\|}{r^2}}$. Further, suppose $r^2 \ge 4\gamma^2 \|\mathcal{I}_R^{-1}\| \frac{\log \frac{2}{\delta}}{n}$ Then, with probability $1 - \delta$ over n samples from f^λ, the output of Algorithm 3 satisfies*

$$\|\hat{\lambda} - \lambda\|_M \le \left(1 + O\left(\frac{1}{\gamma}\right)\right)\left(\sqrt{\frac{\mathrm{Tr}(M^{1/2}\mathcal{I}_R^{-1}M^{1/2})}{n}} + 4\sqrt{\frac{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|\log\frac{2}{\delta}}{n}}\right) + O\left(\tau\sqrt{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|}\right)$$

Proof. By the guarantee on $\lambda_1 = \lambda + \epsilon$, we have that $\epsilon^T R^{-1}\epsilon \le \tau$. Let T be as defined in Lemma B.13. Now

$$\sup_{w \in S^{d-1}} \frac{w^T R^{-1} w}{w^T \mathcal{I}_R w} = \frac{\|\mathcal{I}_R^{-1}\|}{r^2}$$

so that since $\tau \le \frac{1}{\gamma^2 \log^2 \frac{\|\mathcal{I}_R^{-1}\|}{r^2}}$,

$$\sqrt{\tau}\log\left(\sup_{w \sim S^{d-1}} \frac{w^T R^{-1} w}{w^T \mathcal{I}_R w}\right) \le \frac{1}{\gamma}$$

So, we have

$$\mathrm{Tr}(T) \le \mathrm{Tr}(M^{1/2}\mathcal{I}_R^{-1}M^{1/2})\left(1 + \frac{1}{\gamma}\right)$$

and

$$\|T\| \le \|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|\left(1 + \frac{1}{\gamma}\right)$$

So, by Lemma B.13

$$\begin{aligned}
&\|\hat{\epsilon} - \mathop{\mathbb{E}}_{x \sim f_R}[\mathcal{I}_R^{-1} s_R(x - \epsilon)]\|_M \\
&\le \left(1 + O\left(\frac{1}{\gamma}\right)\right)\left(\sqrt{\frac{\mathrm{Tr}(M^{1/2}\mathcal{I}_R^{-1}M^{1/2})}{n}} + 4\sqrt{\frac{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|\log\frac{2}{\delta}}{n}} + \frac{8\|M^{1/2}\mathcal{I}_R^{-1}\|_F^2}{r^2 n^{3/2}\sqrt{\mathrm{Tr}(M^{1/2}\mathcal{I}_R^{-1}M^{1/2})}}\log\frac{2}{\delta}\right) \\
&+ 16\frac{\|M^{1/2}\mathcal{I}_R^{-1}\|\log\frac{2}{\delta}}{rn}
\end{aligned}$$

Since $r^2 \geq 4\gamma^2 \|\mathcal{I}_R^{-1}\| \frac{\log \frac{2}{\delta}}{n}$, $\frac{1}{r} \leq \left(\frac{n}{\log \frac{2}{\delta}}\right)^{1/2} \frac{1}{2\gamma\sqrt{\|\mathcal{I}_R^{-1}\|}}$. So,

$$\begin{aligned}
16\frac{\|M^{1/2}\mathcal{I}_R^{-1}\| \log \frac{2}{\delta}}{rn} &\leq \frac{8\|M^{1/2}\mathcal{I}_R^{-1}\|}{\gamma\sqrt{\|\mathcal{I}_R^{-1}\|}} \left(\frac{\log \frac{2}{\delta}}{n}\right)^{1/2} \\
&\leq \frac{8}{\gamma}\sqrt{\frac{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\| \log \frac{2}{\delta}}{n}} \\
&\qquad \text{since } \frac{(\|M^{1/2}\mathcal{I}_R^{-1}\|)^2}{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|} = \left(\frac{\|M^{1/2}\mathcal{I}_R^{-1}\|}{\|M^{1/2}\mathcal{I}_R^{-1/2}\|}\right)^2 \leq \|\mathcal{I}_R^{-1}\|
\end{aligned}$$

Similarly, $\frac{1}{r^2} \leq \frac{n}{4\gamma^2\|\mathcal{I}_R^{-1}\| \log \frac{2}{\delta}}$. So,

$$\begin{aligned}
\frac{8\|M^{1/2}\mathcal{I}_R^{-1}\|_F^2}{r^2 n^{3/2}\sqrt{\operatorname{Tr}(M^{1/2}\mathcal{I}_R^{-1}M^{1/2})}} \log\frac{2}{\delta} &\leq \frac{8\operatorname{Tr}(M\mathcal{I}_R^{-2})}{r^2 n^{3/2}\sqrt{\operatorname{Tr}(M\mathcal{I}_R^{-1})}} \log\frac{2}{\delta} \\
&\leq \frac{2\operatorname{Tr}(M\mathcal{I}_R^{-2})}{\gamma\|\mathcal{I}_R^{-1}\| n\sqrt{\operatorname{Tr}(M\mathcal{I}_R^{-1})}} \sqrt{\log\frac{2}{\delta}} \\
&\leq \frac{8}{\gamma}\sqrt{\frac{\operatorname{Tr}(M^{1/2}\mathcal{I}_R^{-1}M^{1/2})}{n}} \quad \text{using Lemma C.3}
\end{aligned}$$

So, we have

$$\left\|\hat{\epsilon} - \mathop{\mathbb{E}}_{x\sim f_R}\left[\mathcal{I}_R^{-1}s_R(x-\epsilon)\right]\right\|_M \leq \left(1 + O\left(\frac{1}{\gamma}\right)\right)\left(\sqrt{\frac{\operatorname{Tr}(M^{1/2}\mathcal{I}_R^{-1}M^{1/2})}{n}} + \sqrt{\frac{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\| \log\frac{2}{\delta}}{n}}\right)$$

Now, using Lemma B.3

$$\left\|\epsilon - \mathop{\mathbb{E}}_{x\sim f_R}\left[\mathcal{I}_R^{-1}s_R(x-\epsilon)\right]\right\|_M \lesssim \sqrt{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|}(\epsilon^T R^{-1}\epsilon) \leq \tau\sqrt{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|}$$

So, we have

$$\begin{aligned}
\|\hat{\epsilon} - \epsilon\|_M &\leq \|\hat{\epsilon} - \mathop{\mathbb{E}}_{x\sim f_R}\left[\mathcal{I}_R^{-1}s_R(x-\epsilon)\right]\|_M + \|\epsilon - \mathop{\mathbb{E}}_{x\sim f_R}\left[\mathcal{I}_R^{-1}s_R(x-\epsilon)\right]\|_M \\
&\leq \left(1 + O\left(\frac{1}{\gamma}\right)\right)\left(\sqrt{\frac{\operatorname{Tr}(M^{1/2}\mathcal{I}_R^{-1}M^{1/2})}{n}} + 4\sqrt{\frac{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\| \log\frac{2}{\delta}}{n}}\right) \\
&\quad + O\left(\tau\sqrt{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|}\right)
\end{aligned}$$

Now, since $\hat{\lambda} = \lambda_1 - \hat{\epsilon}$ and $\lambda = \lambda_1 - \epsilon$, $\hat{\lambda} - \lambda = \hat{\epsilon} - \epsilon$. The claim follows. □

B.6. Global MLE

In this section, we state and prove our main theorem, which shows how to estimate the location λ on rate that depends on $\mathcal{I}_R$, given n samples from f^λ.

We begin by stating a result from the heavy-tailed estimation literature, which we will make use of to generate an initial estimate $\lambda + \epsilon$. We will then apply the result from the previous section to refine this estimate in order to recover our final estimate.

Theorem B.15 ((Hopkins, 2018; Cherapanamjeri et al., 2019; Diakonikolas et al., 2020)). *There are universal constants C_0, C_1, C_2 such that for every $n, d \in \mathbb{N}$ and $\delta > 2^{-n/C_2}$, there is an algorithm which runs in time $O(nd) + (d\log(1/\delta))^{C_0}$ such that for every random variable X on $\mathbb{R}^d$, given i.i.d. copies $X_1, \dots, X_n$ of X, outputs a vector $\hat{\mu}_\delta(X_1, \dots, X_n)$ such that*

$$\mathbb{P}\left[\|\mu - \hat{\mu}_\delta\| > C_1\left(\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} + \sqrt{\frac{\|\Sigma\|\log(1/\delta)}{n}}\right)\right] \le \delta$$

where $\mathbb{E}[X] = \mu$ and $\mathbb{E}\left[(X-\mu)(X-\mu)^T\right] = \Sigma$

Algorithm 4 High-dimensional Global MLE

Input Parameters:

- Failure probability δ, description of distribution f, n samples from f^λ, Smoothing R, Approximation parameter γ

1. Let Σ be the covariance matrix of f. Compute an initial estimate λ_1 using the first $1/\gamma$ fraction of of the n samples, using an estimator from Theorem B.15.
2. Run Algorithm 3 using the remaining $1 - 1/\gamma$ fraction of samples using R-smoothing and our initial estimate λ_1, returning the final estimate $\hat{\lambda}$.

Theorem B.16 (Global MLE). *Let f be a given model on $\mathbb{R}^d$, and suppose we are given n samples from f^λ for unknown λ. Let $R = r^2 I_d$ for $0 < r^2 < \|\Sigma\|$ so that $\mathcal{I}_R$ is the R-smoothed Fisher information matrix of f, and let Σ be the covariance of f. Let $M \in \mathbb{R}^{d\times d}$ be any symmetric matrix with $M \succcurlyeq 0$ and let $d_R := d_{\mathit{eff}}(M^{1/2}\mathcal{I}_R^{-1}M^{1/2})$. Fix failure probability $\delta > 0$ and let $2 \le \gamma \le \left(\frac{n}{d_R + \log\frac{1}{\delta}}\right)^{1/8-\alpha}$ for some $\alpha > 0$. Let $n \ge C\gamma^4(\frac{\|\Sigma\|}{r^2})^2\left(\log\frac{4}{\delta} + d_R + \left(\frac{d_{\mathit{eff}}(\Sigma)^2}{d_R}\right)\right)$ for large enough constant $C > 0$. Then, with probability $1-\delta$, the output $\hat{\lambda}$ of Algorithm 4 satisfies*

$$\|\hat{\lambda} - \lambda\|_M \le \left(1 + O\left(\frac{1}{\gamma}\right)\right)\left(\sqrt{\frac{\mathrm{Tr}(M^{1/2}\mathcal{I}_R^{-1}M^{1/2})}{n}} + 4\sqrt{\frac{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|\log\frac{4}{\delta}}{n}}\right)$$

Proof. By the guarantee from Theorem B.15, our initial estimate $\lambda_1 = \lambda + \epsilon$ from Step 1 has the property that with probability $1 - \delta/2$,

$$\|\epsilon\|^2 \lesssim \frac{\mathrm{Tr}(\Sigma)}{n/\gamma} + \frac{\|\Sigma\|\log\frac{2}{\delta}}{n/\gamma}$$

We condition on the success of Step 1. Let $n' = n(1-1/\gamma) \ge n/2$ be the number of samples used in Step 2 to call the Local MLE Algorithm 3. By our lower bound on n,

$$r^2 \ge \sqrt{C}\gamma^2\frac{\|\Sigma\|}{\sqrt{n}}\left(\frac{d_{\mathrm{eff}}(\Sigma)}{\sqrt{d_R}} + d_R + \sqrt{\log\frac{4}{\delta}}\right) \ge \sqrt{C}\gamma^2\frac{\mathrm{Tr}(\Sigma) + \|\Sigma\|\log\frac{4}{\delta}}{\sqrt{2}n'} \cdot \frac{(n')^{1/2}}{\sqrt{d_R + \log\frac{4}{\delta}}}$$

So, for large enough C, since $\gamma > 1$, setting

$$\tau = \frac{1}{\gamma}\sqrt{\frac{d_R + \log\frac{4}{\delta}}{n'}}$$

yields that

$$\epsilon^T R^{-1}\epsilon = \frac{\|\epsilon\|^2}{r^2} \le \tau$$

Also $\tau \le 1/4$ since $n' \ge \frac{C}{2}(\log\frac{4}{\delta} + d_R)$. So, the condition on the confidence set S in Lemma B.14 is satisfied.

By the constraint on n, we have

$$r^2 \geq \sqrt{C}\gamma^2\|\Sigma\|\sqrt{\frac{\log\frac{4}{\delta} + d_R + \left(\frac{d_{\text{eff}}(\Sigma)^2}{d_R}\right)}{n}} \geq \sqrt{\frac{C}{2}}\gamma^3\|\Sigma\|\tau$$

We also have by Lemma C.2 that $\|\mathcal{I}_R^{-1}\|/\|\Sigma\| \leq \frac{\|\Sigma+R\|}{\|\Sigma\|} \leq 2$, so

$$\log\frac{\|\mathcal{I}_R^{-1}\|}{r^2} \leq \log\frac{2\sqrt{2}}{\sqrt{C}\gamma^3\tau}$$

so the τ constraint is that

$$\gamma^2\tau\log^2\frac{2\sqrt{2}}{\sqrt{C}\gamma^3\tau} \leq 1$$

the LHS is at most

$$O(\tau^{0.99}\gamma^{2.01}) < 1$$

since $\tau < 1/\gamma^5$, with a constant that is arbitrarily small with C. So the constraint on τ of Lemma B.14 is satisfied.

Using the fact that $n' \geq \frac{C}{2}\left(\log\frac{4}{\delta}\right)$,

$$\begin{aligned}
\frac{r^2 n'}{\|\mathcal{I}_R^{-1}\|\log\frac{4}{\delta}} &\geq \frac{C}{2}\gamma^2\frac{\|\Sigma\|\sqrt{n}}{\|\mathcal{I}_R^{-1}\|\log\frac{4}{\delta}}\left(\frac{d_{\text{eff}}(\Sigma)}{\sqrt{d_R}} + \sqrt{\log\frac{4}{\delta}}\right) \\
&\geq \frac{C}{2}\gamma^2\frac{\|\Sigma\|\sqrt{n}}{\|\Sigma+R\|\log\frac{4}{\delta}}\sqrt{\log\frac{4}{\delta}} \quad \text{by Lemma C.2} \\
&\geq \gamma^2 \quad \text{since } R = r^2 I_d \text{ so that } \|R\| = r^2 < \|\Sigma\|
\end{aligned}$$

So the conditions of Lemma B.14 are satisfied, and with probability $1-\delta/2$,

$$\begin{aligned}
\|\hat{\lambda}-\lambda\|_M &\leq \left(1+O\left(\frac{1}{\gamma}\right)\right)\left(\sqrt{\frac{\text{Tr}(M^{1/2}\mathcal{I}_R^{-1}M^{1/2})}{n'}} + 4\sqrt{\frac{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|\log\frac{4}{\delta}}{n'}}\right) \\
&\quad + O\left(\tau\sqrt{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|}\right) \\
&\leq \left(1+O\left(\frac{1}{\gamma}\right)\right)\left(\sqrt{\frac{\text{Tr}(M^{1/2}\mathcal{I}_R^{-1}M^{1/2})}{n}} + 4\sqrt{\frac{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|\log\frac{4}{\delta}}{n}}\right)
\end{aligned}$$

since $n' = n(1-1/\gamma)$ and $\tau = \frac{1}{\gamma}\sqrt{\frac{d_R+\log\frac{4}{\delta}}{n'}}$. So, our total failure probability is δ. The claim follows. □

Theorem B.17 (Global MLE, Informal)**.** *Let f have covariance matrix Σ. For any $r^2 \leq \|\Sigma\|$, let $R = r^2 I_d$ and $\mathcal{I}_R$ be the R-smoothed Fisher information of the distribution. For any constant $0 < \epsilon < 1$,*

$$\|\widehat{\lambda}-\lambda\|_2 \leq (1+\epsilon)\sqrt{\frac{\text{Tr}(\mathcal{I}_R^{-1})}{n}} + 5\sqrt{\frac{\|\mathcal{I}_R^{-1}\|\log\frac{4}{\delta}}{n}}$$

with probability $1-\delta$, for $n > O_\epsilon\left(\left(\frac{\|\Sigma\|}{r^2}\right)^2\left(\log\frac{2}{\delta} + d_{\text{eff}}(\mathcal{I}_R^{-1}) + \frac{d_{\text{eff}}(\Sigma)^2}{d_{\text{eff}}(\mathcal{I}_R^{-1})}\right)\right)$.

Proof. First, if $\epsilon > 1/4$, we reset $\epsilon = 1/4$. Setting $M = I_d$ so that $d_R = d_{\text{eff}}(\mathcal{I}_R^{-1})$, and setting $\gamma = \frac{C_0}{\epsilon}$ for sufficiently large constant C_0 in Theorem B.16 gives the claim. □

C. Useful Results

The following is a continuous version of the rearrangement inequality (user940, 2015):

Lemma C.1. *Let $f, g : \mathbb{R} \to \mathbb{R}$ be monotonically non-decreasing functions, and X be a random variable over $\mathbb{R}$. Then*

$$\mathbb{E}[f(X)]\,\mathbb{E}[g(X)] \le \mathbb{E}[f(X)g(X)]$$

Proof. Let Y be an independent copy of X. By monotonicity,

$$(f(X) - f(Y))(g(X) - g(Y)) \ge 0$$

always. Taking the expectation of both sides,

$$2\,\mathbb{E}[f(X)g(X)] - 2\,\mathbb{E}[f(X)g(Y)] \ge 0.$$

Since Y is independent of X, this gives the result. □

Lemma C.2. *Let f be an arbitrary distribution on $\mathbb{R}^d$, and let Σ be its covariance matrix. Let f_R be the R-smoothed version of f, with Fisher information matrix $\mathcal{I}_R$. Then,*

$$\mathcal{I}_R \succcurlyeq (\Sigma + R)^{-1}$$

Proof. Follows from the fact that the covariance of f_R is $\Sigma + R$, and using Theorem 1.2 from (Hendeby, 2005). □

Lemma C.3. *Let A, B be symmetric PSD matrices. Then*

$$\operatorname{Tr}(AB) \le \operatorname{Tr}(A)\|B\|$$

Proof. Let the eigenvectors of B be $v_1, \ldots, v_d$. Then

$$\operatorname{Tr}(AB) = \sum_{i=1}^{d} v_i^T A(Bv_i) \le \|B\| \sum_{i=1}^{d} v_i^T A v_i = \|B\| \operatorname{Tr}(A).$$

□

D. Computing the high-dimensional R-smoothed local MLE (Algorithm 3)

A precise bound on the complexity of Algorithm 3 depends on how the high-dimensional distribution is represented. However, the algorithm is polynomial time (in n, d) as long as (a) sampling from the distribution and (b) computing the score of the smoothed distribution at a point are both efficient operations.

Reading Algorithm 3, the only non-trivial computation is for $\mathcal{I}_R$, the R-smoothed Fisher information matrix. Since $\mathcal{I}_R$ is the covariance matrix of the score vector, and the score vector has nice tails (e.g. Lemma B.7), we can estimate it by the empirical covariance of *simulated* samples from the smoothed model. With enough samples, we get a spectral approximation to $\mathcal{I}_R$ and hence $\mathcal{I}_R^{-1}$. This approximation can then be used in place of the true $\mathcal{I}_R$ in Algorithm 3. The number of simulated samples required, such that the approximation adds minimal extra estimation error, will be $\operatorname{poly}(1/\eta, d, \|\Sigma\|/r^2)$, where η is the approximation parameter in Theorem 4.1. The bounds on η and r imply this is $\operatorname{poly}(nd)$.

A popular representation of high-dimensional distributions is Gaussian mixture models with different covariances, which appear also in our experiments in Section 6. For a GMM with k components, the runtime of Algorithm 3 is dominated by $O(knd^2)$ time to compute the scores and $\tilde{O}(kn^{1.5}d^4 + d^3)$ time to estimate $\mathcal{I}_R$ and invert it.