Finite-Sample Maximum Likelihood Estimation of Location

Shivam Gupta
The University of Texas at Austin
shivamgupta@utexas.edu

Jasper C.H. Lee
University of Wisconsin–Madison
jasper.lee@wisc.edu

Eric Price
The University of Texas at Austin
ecprice@cs.utexas.edu

Paul Valiant
Purdue University
pvaliant@gmail.com

Abstract

We consider 1-dimensional location estimation, where we estimate a parameter λ from n samples $\lambda + \eta_i$, with each η_i drawn i.i.d. from a known distribution f. For fixed f the maximum-likelihood estimate (MLE) is well-known to be optimal *in the limit* as $n \to \infty$: it is asymptotically normal with variance matching the Cramér-Rao lower bound of $\frac{1}{n\mathcal{I}}$, where $\mathcal{I}$ is the Fisher information of f. However, this bound does not hold for finite n, or when f varies with n. We show for arbitrary f and n that one can recover a similar theory based on the Fisher information of a *smoothed* version of f, where the smoothing radius decays with n.

1 Introduction

We revisit a fundamental problem in statistics: consider a translation-invariant parametric model $\{f^\theta\}_{\theta\in\mathbb{R}}$ of distributions, where $f^\theta(x) = f(x-\theta)$. Suppose there is an arbitrarily chosen unknown true parameter λ, and we get i.i.d. samples from f^λ. The task is to accurately estimate λ from the samples. This problem is known as *location parameter* estimation in the statistics literature.

Location estimation is a well-studied and general model, including as a special case the important setting of Gaussian mean estimation. In contrast to general mean estimation (where we want to estimate the mean of a distribution given minimal assumptions such as moment conditions), in location estimation we are given the shape of the distribution up to shift. This advantage lets us handle some distributions where mean estimation is impossible (e.g., the mean may not exist), and lets us aim for higher accuracy than is possible without knowing the distribution.

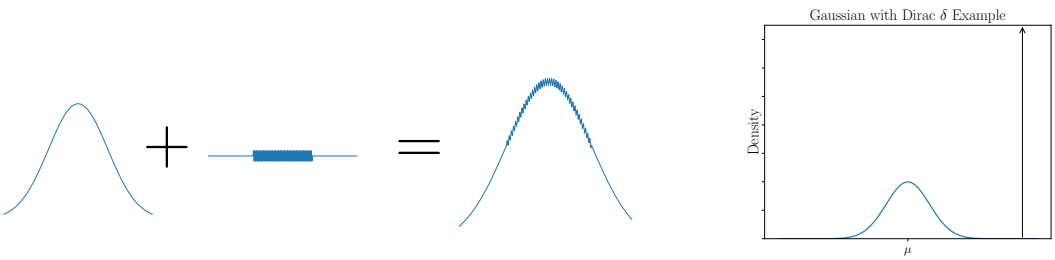
The classic theory of location estimation is asymptotic, see [vdV00] for a detailed background. On the algorithmic side, it is well-known that the maximum likelihood estimator (MLE) is asymptotically normal. Specifically, as the number of samples n tends to infinity, the distribution of the MLE converges to a Gaussian centered at the true parameter with variance $1/(n\mathcal{I})$, where $\mathcal{I}$ is the *Fisher information* of the distribution f:

$$\mathcal{I} := \int \frac{(f'(x))^2}{f(x)}\,\mathrm{d}x = \mathop{\mathbb{E}}_{x\sim f}\left[(\frac{\partial}{\partial x}\log f(x))^2\right]. \tag{1}$$

Conversely, the celebrated Cramér-Rao bound states that the variance of any unbiased estimator must be at least $1/(n\mathcal{I})$, meaning that the MLE has mean-squared error that is asymptotically at least as good as any unbiased estimator.

In the last few decades, motivated by an increasing dependence on data for high-stakes applications, the statistics and computer science communities have shifted focus towards *finite-sample* and *high*

Preprint. Under review.



(a) Sawtooth example

(b) Dirac δ example

Figure 1: Bad examples for MLE

probability theories: 1) asymptotic theories assume access to an infinite amount of data, and can in certain cases fail to predict the performance of an algorithm with only a finite number of samples—see the next section for bad examples for the MLE— 2) in high-stakes applications where failure can be catastrophic, it is crucial for predictions to hold except with exponentially small probability. Yet, classic results bounding the variance or mean-squared error of estimators, such as the Cramér-Rao bound, do not readily translate to (tight) high probability bounds.

The goal of this paper is to establish a finite-sample and high probability theory for the location estimation problem, in both the algorithmic bound and the estimation lower bound. Our algorithmic theory includes a simple yet crucial modification of perturbing samples by Gaussian noise—corresponding to drawing samples from a Gaussian-smoothed version of the underlying distribution—before performing MLE. We show that this smoothed MLE has finite-n high-probability performance analogous to the Gaussian tail in the classic asymptotic theory, but replacing the usual Fisher information with a *smoothed Fisher information*. The amount of smoothing required decreases with n.

Complementing our upper bound result, we prove a high probability version of the Cramér-Rao bound for Gaussian-smoothed distributions, showing that for these distributions our sub-Gaussian accuracy bound, with variance determined by the Fisher information, is optimal to within a $1 + o(1)$ factor.

1.1 Obstacles to a Finite Sample Theory

Before discussing our results in detail, we examine two simple distributions where the asymptotic theory predictions for the MLE do not hold in finite samples. The first example highlights an information-theoretic barrier, that no algorithm can attain the performance predicted by the Gaussian with variance $1/\mathcal{I}$. The second example, on the other hand, demonstrates how the MLE can be "tricked" by the distribution, and how an algorithmic remedy is needed to improve its accuracy.

Gaussian with sawtooth noise Our first example takes a standard Gaussian, and adds a fine-grained sawtooth perturbation to it over a bounded region at the center of the Gaussian, as shown in Figure 1a. The fine-grained sawtooth has slope either $+\Delta$ or $-\Delta$, alternating over "teeth" of width w, for $\Delta \gg 1$ and $w \ll 1/\Delta$. This sawtooth perturbation barely changes the pdf, but significantly changes its derivatives, so the Fisher information $\int (f')^2/f$ grows from 1 to $\Theta(\Delta^2)$.

Essentially, the sawtooth perturbation makes the distribution easier to align *within* a tooth, but not *across* teeth. The asymptotic analysis reflects that: for $n \gg 1/w^2$ where we can align the teeth correctly, the MLE is much more accurate on the perturbed distribution. But for $n \ll 1/w^2$, no algorithm can do better on the perturbed distribution than on a regular Gaussian.[1] Thus, the normalized estimation accuracy depends on n: for large enough n, it has variance $\frac{\Theta(1)}{\Delta^2 n}$, but for smaller n it has variance $\frac{1}{n}$. Our finite sample theory should reflect this.

Gaussian with a Dirac δ spike Our next example adds an even simpler noise to the standard Gaussian: a Dirac δ spike with minimal mass ε, placed sufficiently far away from the Gaussian mean 0 (Figure 1b). This distribution has infinite Fisher information, reflecting that for $n \gg 1/\varepsilon$, we will

[1] This can be shown by the KL divergence from shifting an integer number of teeth of distance about $1/\sqrt{n}$.

probably see the spike multiple times, identify it, and get zero estimation error. But for $n < 1/\varepsilon$, we probably will not see the spike, so our error bound will not reflect the overall Fisher information.

Moreover, the MLE performs remarkably badly when $n \ll 1/\varepsilon$. Most likely the Dirac δ is not sampled, yet the Dirac δ has infinite density, and so the MLE will match a Gaussian sample to the spike to get a maximium likelihood solution. As the spike is placed far from the true mean, this leads to much higher error than (say) the empirical mean.

Fortunately, there is a simple modification to the MLE that solves this issue almost completely: we add a small amount of Gaussian noise to the distribution. This means we convolve the PDF with a small Gaussian, and add independent Gaussian noise to each individual sample to effectively draw from the convolved PDF. The crucial effect of this Gaussian smoothing is that the density of the Dirac δ will be reduced from infinity down to some constant, and hence MLE will no longer "fit it at any cost". On the other hand, the smoothing increases the variance of the distribution and decreases the Fisher information. This raises the question of determining the *best* amount of smoothing, which we address in this paper.

1.2 Our Results and Approach

Based on the discussion of the spiked Gaussian model in the previous section, we propose a finite sample theory for MLE that uses Gaussian smoothing. As we described, the algorithmic approach is simple: we perturb all the samples by independent Gaussian noise of variance r^2, and convolve the known model f by the same Gaussian to yield the model f_r, before performing MLE.

We state below the basic MLE algorithm (Algorithm 1) in this paper, which adopts the above approach. Algorithm 1 is a *local* algorithm, in that it assumes as input an initial uncertainty region guaranteed to contain the true parameter λ, and performs MLE only over this domain. Furthermore, Algorithm 1 attempts only to find a local optimum in the likelihood: it computes the derivative of the log likelihood function, also known as the *score* function, and returns any root of the score function.

Algorithm 1 Local MLE for known parametric model

Input Parameters: Description of distribution f, smoothing parameter r, samples $x_1, \ldots, x_n \overset{i.i.d.}{\sim} f^\lambda$, uncertainty region $[\ell, u]$ containing the unknown λ

1. Let $s_r(\hat{\lambda})$ 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 $\hat{\lambda}$ that is a root of the empirical score function $\hat{s}(\hat{\lambda}) = \sum_{i=1}^n s_r(x_i' - \hat{\lambda})$ inside the domain $[\ell, u]$. A root should exist and picking any root is sufficient.
4. Return $\hat{\lambda}$.

Theorem 1.1 states our guarantees on Algorithm 1. Locally around the true parameter—that is, within $r/2$ for the r-smoothed distribution—any root of the score function (i.e., local optimum of likelihood) must be very close to the true parameter λ with high probability over the n samples. In particular, the estimation error is within a $1 + o(1)$ factor of the Gaussian deviation with variance $\frac{1}{n\mathcal{I}_r}$ and failure probability δ, where $\mathcal{I}_r$ is the Fisher information of f_r.

Theorem 1.1 (Local Convergence)**.** *Suppose we have a known model f_r that is the result of r-smoothing, with Fisher information $\mathcal{I}_r$, and a given width parameter $\varepsilon_{\max}$ and failure probability δ. Further suppose that r satisfies $r \geq 2\varepsilon_{\max}$ and there is a sufficiently large parameter γ such that 1) $r^2\sqrt{\mathcal{I}_r} \geq \gamma\varepsilon_{\max}$, 2) $(\log \frac{1}{\delta})/n \leq \frac{1}{\gamma^2}$ and 3) $\log 1/(r\sqrt{\mathcal{I}_r}) \leq \frac{1}{\gamma} \log \frac{1}{\delta}/\log\log \frac{1}{\delta}$.*

Then, with probability at least $1 - \delta$, for all $\varepsilon \in \left((1 + O(\frac{1}{\gamma}))\sqrt{\frac{2\log\frac{1}{\delta}}{n\mathcal{I}_r}}, \varepsilon_{\max}\right]$, $\hat{s}(\lambda - \varepsilon)$ is strictly negative and $\hat{s}(\lambda + \varepsilon)$ is strictly positive.

In the scenario where we do have an initial uncertainty region for the true parameter λ, we would use Theorem 1.1 to compute the minimal smoothing amount r satisfying the assumptions in the theorem, then use Algorithm 1 with this parameter r, to obtain an accurate estimate of λ.

In general, however, we may not have an initial uncertainty region for λ. In Section 5, we present Algorithm 2, a global two-stage MLE algorithm, which first infers an initial uncertainty region by using quantile information from the distribution f before invoking Algorithm 1, the local MLE algorithm. The guarantees of Algorithm 2 are summarized here.

Theorem 1.2 (Global MLE guarantees, informal version of Theorem 5.1)**.** *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. When $n \gg \log \frac{1}{\delta} \gg 1$, there exists an $r^* = o(\mathrm{IQR})$ such that, with probability at least $1-\delta$, the output $\hat{\lambda}$ of Algorithm 2 satisfies*

$$|\hat{\lambda} - \lambda| \leq (1 + o(1)) \sqrt{\frac{2 \log \frac{1}{\delta}}{n I_{r^*}}}$$

In addition to the theoretical framework, Section 7 gives experimental evidence demonstrating that r-smoothed Fisher information does capture the empirical performance of (smoothed) MLE.

We also prove new estimation lower bounds for the location estimation problem for r-smoothed distributions. The lower bound statement below (Theorem 1.3) shows that the estimation error $(1 + o(1)) \sqrt{\frac{2 \log \frac{1}{\delta}}{n I_r}}$ is optimal to within a $1 + o(1)$ factor.

Theorem 1.3. *Suppose f_r is an r-smoothed distribution with Fisher information $\mathcal{I}_r$. Given failure probability δ and sample size n, no algorithm can distinguish f_r and $f_r^{2\varepsilon}$ with probability $1-\delta$, where $\varepsilon = (1 - o(1)) \sqrt{2 \log \frac{1}{\delta} / (n \mathcal{I}_r)}$. Here, the $o(1)$ term tends to 0 as $\delta \to 0$ and $\log \frac{1}{\delta} / n \to 0$, for a fixed $r^2 \mathcal{I}_r$.*

This lower bound is the standard "two-point" statement that, with n samples, it is statistically impossible to distinguish between the distributions f and f shifted by a small error (in the x-axis) with probability $1-\delta$. Even though there are known standard inequalities on distribution distances and divergences for proving lower bounds of this form, the technical challenge is that they generally yield estimation lower bounds that are only tight to within constant factors, instead of the $1 + o(1)$ tightness we desire. This paper presents new analysis to derive a $1 + o(1)$-tight lower bound, which may be of independent interest.

1.3 Notations

We denote the shift-invariant model we consider by the distribution f, and the distribution with parameter λ by $f^\lambda(x) = f(x - \lambda)$. Denote by Z_r the Gaussian with mean 0 and variance r^2. The r-smoothed model for f is denoted by f_r (and similarly, for parameter λ, f_r^λ) which is distributed as $Y = Z_r + X$ where $X \leftarrow f$ independently from the Gaussian perturbation Z_r.

The log-likelihood function of f is denoted by $l = \log f$. The *score* function is the derivative of l, denoted by $s = l' = f'/f$. We use the notation s_r to denote the score function of f_r. The Fisher Information of f is denoted by $I = \mathbb{E}_{x \leftarrow f}[s^2(x)]$. Similarly, the Fisher Information of f_r is denoted by I_r.

2 Related Work

Location estimation and MLE in general has been extensively studied under the lens of asymptotic statistics. See [vdV00] for an in-depth treatment. The MLE has also been studied under the finite-sample setting [Spo11, Pin17, Mia10], but these prior works impose restrictive regularity conditions and also loses (at least) multiplicative constants in the estimation accuracy. In contrast, our work modifies the MLE to include smoothing, and we give analyses that are tight to within $1 + o(1)$ factors.

There has also been a flurry of recent interest in the related mean estimation problem, in the finite-sample and high-probability setting. Recall that mean estimation does not assume knowledge of the shape of the distribution, but instead imposes mild moment conditions, for example the finiteness of the variance. Catoni [Cat12] initiated a line of work studying the statistical limits of univariate mean estimation to within a $1 + o(1)$ factor, ending recently with the work of Lee and Valiant [LV22], which proposed and analyzed an estimator with accuracy optimal to within a $1 + o(1)$ factor for all distributions with finite variance. See also the recent work of Minsker [Min22] for an alternative solution.

Beyond the differences in assumptions, the main distinction between location and mean estimation lies in their statistical limits. In mean estimation, the optimal accuracy is captured by the variance of the underlying distribution, scaling linearly with the standard deviation. On the other hand, the classic asymptotic theory suggests that the Fisher information captures the optimal accuracy for location estimation, scaling with the reciprocal of the square root of the Fisher information. It is a well-known fact that the Fisher information is always lower bounded by the reciprocal of the variance [SV11], which shows that location estimation is always easier than mean estimation in the infinite-sample regime. In this work, we refine this understanding, showing that in finite samples, the optimal accuracy for location estimation is instead given by the r-smoothed Fisher information in place of the unsmoothed Fisher information.

3 Tails and boundedness of r-smoothed score and Fisher information

Recall that given a distribution f, its r-smoothed version f_r is distributed as $Y = X + Z_r$ where $X \sim f$ and $Z_r \sim \mathcal{N}(0, r^2)$ and X, Z_r are independent.

Both our algorithmic and lower bound theories are centered around r-smoothed distributions. Therefore, we state here basic concentration and boundedness properties of r-smoothed score function and Fisher information, which we use in the rest of the paper. We prove all these lemmas in Appendix A.

First, we show that the r-smoothed Fisher information $\mathcal{I}_r$ is upper bounded by $1/r^2$ and can be lower bounded using the interquartile range of f.

Lemma 3.1. *Let $\mathcal{I}_r$ be the Fisher information of an r-smoothed distribution f_r. Then, $\mathcal{I}_r \le 1/r^2$.*

Lemma 3.2. *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, $\mathcal{I}_r \gtrsim 1/(\mathrm{IQR} + r)^2$. Here, the hidden constant is a universal one independent of the distribution f and independent of r.*

Next, we show that, fixing a point close to the true parameter λ, the empirical score function evaluated at that point will concentrate around its expectation for smoothed distributions.

Corollary 3.3. *Let f be an arbitrary distribution and let f_r be the r-smoothed version of f. That is, $f_r(x) = \mathbb{E}_{y \leftarrow f}[\frac{1}{\sqrt{2\pi r^2}} e^{-\frac{(x-y)^2}{2r^2}}]$. Consider the parametric family of distributions $f_r^\lambda(x) = f_r(x - \lambda)$. 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 $|\varepsilon| \le r/2$,

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

4 A Finite Sample Analysis of r-smoothed Local MLE

In this section, we analyze Algorithm 1, which is our version of local MLE with r-smoothing applied. Algorithm 1 takes an initial uncertainty region that the true parameter is guaranteed to lie in, and uses the model and the initial interval to refine the estimate to high accuracy. We first present a simpler and easier-to-interpret version of our result, Theorem 1.1, which we stated in Section 1.2.

Recall that Algorithm 1 computes the empirical score function, and returns any of its roots. The theorem thus states that, with high probability, for any point $\lambda + \varepsilon$ with $|\varepsilon|$ too large, the empirical score function must be non-zero and thus $\lambda + \varepsilon$ will not returned as the estimate. More precisely, given an initial interval of length $\varepsilon_{\max}$ as well as the failure probability δ, the theorem assumes that the smoothing parameter r is sufficiently large (conditions 1 and 3 in the theorem) and that the sample size n is sufficiently large, and guarantees an estimation error that is within a $1 + o(1)$ factor of the error predicted by the Gaussian with variance $1/\mathcal{I}_r$, where $\mathcal{I}_r$ is the Fisher information of f_r.

Theorem 1.1 (Local Convergence)**.** *Suppose we have a known model f_r that is the result of r-smoothing, with Fisher information $\mathcal{I}_r$, and a given width parameter $\varepsilon_{\max}$ and failure probability δ.*

Further suppose that r satisfies $r \geq 2\varepsilon_{\max}$ and there is a sufficiently large parameter γ such that 1) $r^2\sqrt{\mathcal{I}_r} \geq \gamma\varepsilon_{\max}$, 2) $(\log \frac{1}{\delta})/n \leq \frac{1}{\gamma^2}$ and 3) $\log 1/(r\sqrt{\mathcal{I}_r}) \leq \frac{1}{\gamma}\log\frac{1}{\delta}/\log\log\frac{1}{\delta}$.

Then, with probability at least $1-\delta$, for all $\varepsilon \in \left((1+O(\frac{1}{\gamma}))\sqrt{\frac{2\log\frac{1}{\delta}}{n\mathcal{I}_r}}, \varepsilon_{\max}\right]$, $\hat{s}(\lambda-\varepsilon)$ is strictly negative and $\hat{s}(\lambda+\varepsilon)$ is strictly positive.

The above theorem follows from the following theorem, which makes the "$o(1)$" term (the $O(1/\gamma)$ term) in the theorem explicit. Assumptions 2 and 3 in the theorem statement essentially bounds various multiplicative terms in the estimation error and makes sure that they are "$1+o(1)$" terms.

Theorem 4.1. *Suppose we have a known model f_r that is the result of r-smoothing, and a given parameter $\varepsilon_{\max}$. Let β and η be the hidden multiplicative constants in Lemmas B.2 and B.3. Further suppose that r satisfies $r \geq 2\varepsilon_{\max}$ and $r^2\sqrt{\mathcal{I}_r} \geq \gamma\varepsilon_{\max}$ for some parameter $\gamma \geq \beta$.*

Now define the notation ρ_r by

$$1+\rho_r = \sqrt{1+\frac{\eta}{\gamma}} + \frac{15}{2\sqrt{\gamma}}\left(\frac{2\log\frac{4\log\frac{1}{\delta}}{r^2\mathcal{I}_r(1-\frac{\beta}{\gamma})\delta}}{n}\right)^{\frac{1}{4}}$$

Then, for sufficiently small $\delta > 0$, with probability at least $1-\delta$, for all $\varepsilon \in \left((1+\frac{1}{\log\frac{1}{\delta}})\frac{1+\rho_r}{1-\frac{\beta}{\gamma}}\sqrt{1+\frac{\log\frac{4\log\frac{1}{\delta}}{r^2\mathcal{I}_r(1-\frac{\beta}{\gamma})}}{\log\frac{1}{\delta}}}\sqrt{\frac{2\log\frac{1}{\delta}}{n\mathcal{I}_r}}, \varepsilon_{\max}\right]$, $\hat{s}(\lambda-\varepsilon) < 0$ and $\hat{s}(\lambda+\varepsilon) > 0$.

To prove Theorem 4.1, it suffices to show the following lemma. The theorem follows directly by reparameterizing δ and choosing ξ to be $1/\log\frac{1}{\delta}$.

Lemma 4.2. *Suppose we have a known model f_r that is the result of r-smoothing with Fisher information $\mathcal{I}_r$, and a given parameter $\varepsilon_{\max}$. Let β and η be the hidden multiplicative constants in Lemmas B.2 and B.3. Further suppose that r satisfies $r \geq 2\varepsilon_{\max}$ and $r^2\sqrt{\mathcal{I}_r} \geq \gamma\varepsilon_{\max}$ for some parameter $\gamma \geq \beta$. Also define the notation $\tilde{\rho}$ (a "$o(1)$" term) by*

$$1+\tilde{\rho} = \sqrt{1+\frac{\eta}{\gamma}} + \frac{15}{2\sqrt{\gamma}}\left(\frac{2\log\frac{1}{\delta}}{n}\right)^{\frac{1}{4}}$$

Then, for every $\xi \ll 1$, with probability at least $1-\delta\cdot\frac{2}{\xi r^2\mathcal{I}_r(1-\frac{\beta}{\gamma})(1-\delta)}$, for all $\varepsilon \in \left((1+\xi)\frac{1+\tilde{\rho}}{1-\frac{\beta}{\gamma}}\sqrt{\frac{2\log\frac{1}{\delta}}{n\mathcal{I}_r}}, \varepsilon_{\max}\right]$, $\hat{s}(\lambda-\varepsilon)$ is strictly negative and $\hat{s}(\lambda+\varepsilon)$ is strictly positive.

We prove Lemma 4.2 in Appendix B, and here we give a proof sketch.

Proof sketch for Lemma 4.2. First, recall that Corollary 3.3 from Section 3 shows that fixing a candidate input value $\lambda+\varepsilon$ for some small ε, the value of the empirical score function at $\lambda+\varepsilon$ is well-concentrated around its expectation. In Lemmas B.2 and B.3, we calculate and bound the expectation and second moment of the empirical score function at $\lambda+\varepsilon$ for all sufficiently small ε. This allows us to derive tail bounds for the empirical score function at each point $\lambda+\varepsilon$, to show that it is bounded away from 0. Next, we need to show that with high probability, the empirical score function is *simultaneously* bounded away from 0 for all ε with magnitude greater than the desired estimation accuracy. We achieve this via a straightforward net argument, crucially utilizing the fact that the expectation of the empirical score function is bounded away from 0 by an essentially linear function in ε, and that the variance is essentially constant in ε. This means that the probability for the empirical score function at $\lambda+\varepsilon$ to hit 0 is decreasing exponentially in ε, which allows us to complete the net argument. □

5 Global Two-Stage MLE Algorithm

Algorithm 1, which we stated in the introduction and analyzed in Section 4, is a *local* algorithm that assumes we have knowledge of a non-trivially small uncertainty region containing the true parameter

λ. The smoothing parameter r can then be computed from the assumptions of Theorems 4.1 or 1.1, and we run Algorithm 1 to obtain an accurate estimate of the true parameter λ, with accuracy predicted by the r-smoothed Fisher information $\mathcal{I}_r$.

However, in general, we might not have a-priori knowledge of where the true parameter λ lies. In this section, we propose a *global* maximum likelihood algorithm (Algorithm 2) which first estimates a preliminary interval containing λ, before choosing the smoothing parameter r^* using an *easily calculable* expression that is $o(1)$ times smaller than the interquartile range of the distribution, and finally applies the local MLE algorithm (Algorithm 1) to obtain a final estimate. Theorem 5.1 states that the accuracy of Algorithm 2 is always within a $1 + o(1)$ times the accuracy predicted by the r^*-smoothed Fisher information $\mathcal{I}_{r^*}$.

Algorithm 2 Global MLE for known parametric model

Input Parameters: Failure probability δ, description of distribution f, n i.i.d. samples drawn from f^λ for some unknown λ

1. Compute an $\alpha \in [\sqrt{2\log\frac{4}{\delta}/n}, 1 - \sqrt{2\log\frac{4}{\delta}/n}]$ such that the interval defined by the $\alpha - \sqrt{2\log\frac{4}{\delta}/n}$ and $\alpha + \sqrt{2\log\frac{4}{\delta}/n}$ quantiles of f is the smallest.
2. By standard Chernoff bounds, with probability at least $1 - \frac{\delta}{2}$, the sample α-quantile x_α will be such that $x_\alpha - \lambda$ is within the $\alpha - \sqrt{2\log\frac{4}{\delta}/n}$ and $\alpha + \sqrt{2\log\frac{4}{\delta}/n}$ quantiles of f. Based on this, compute an initial confidence interval $[\ell, u]$ for λ.
3. Let $r^* = \Omega(\max((\frac{\log\frac{1}{\delta}}{n})^{1/8}, 2^{-O(\sqrt{\log\frac{1}{\delta}})}))\mathrm{IQR}$.
4. Run Algorithm 1 on the interval $[\ell, u]$, using r^*-smoothing and failure probability $\delta/2$, returning the final estimate $\hat{\lambda}$.

Theorem 5.1 (Global MLE Theorem)**.** *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 δ.*

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

$$|\hat{\lambda} - \lambda| \leq \left(1 + O\left(\frac{\log\frac{1}{\delta}}{n}\right)^{\frac{1}{4}} + O\left(\frac{1}{\sqrt{\log\frac{1}{\delta}}}\right)\right)\sqrt{\frac{2\log\frac{1}{\delta}}{n\mathcal{I}_{r^*}}}$$

Proof. The total failure probability of the steps is at most δ. Thus, in this proof we condition on the success of Algorithm 1 in all probabilistic steps.

By the minimality condition in the definition of α, the length $\varepsilon_{\max}$ of the interval $[\ell, u]$ from Step 2 is at most $O(\sqrt{\log\frac{1}{\delta}/n})\mathrm{IQR}$.

Further, recall by Lemma 3.2 that $\mathcal{I}_r \geq \Omega(1/(\mathrm{IQR} + r)^2)$. Picking $r^* = \Omega(\max((\frac{\log\frac{1}{\delta}}{n})^{1/8}, 2^{-O(\sqrt{\log\frac{1}{\delta}})}))\mathrm{IQR}$ and $\gamma_1 = O(\frac{n}{\log\frac{1}{\delta}})^{1/4}$ and $\gamma_2 = O(\sqrt{\log\frac{1}{\delta}})$, we check that the following conditions are satisfied:

1. $(r^*)^2\sqrt{\mathcal{I}_{r^*}} \geq (r^*)^2/(\mathrm{IQR} + r^*) \geq \Omega(\frac{\log\frac{1}{\delta}}{n})^{1/4}\mathrm{IQR} = \gamma_1\varepsilon_{\max}$.
2. $\log\frac{1}{\delta}/n \leq O(\sqrt{\log\frac{1}{\delta}/n}) \leq 1/\gamma_1^2$.
3. $\log 1/(r^*\sqrt{\mathcal{I}_{r^*}}) \leq O(\log 2^{O(\sqrt{\log\frac{1}{\delta}})}) = O(\frac{1}{\gamma_2}\log\frac{1}{\delta})$.

Further note that $\log\log\frac{1}{\delta}/\log\frac{1}{\delta} \leq 1/\sqrt{\log\frac{1}{\delta}} = O(1/\gamma_2)$.

Thus, using Theorem 1.1, Step 4 returns an estimate $\hat{\lambda}$ satisfying

$$|\hat{\lambda} - \lambda| \leq \left(1 + O\left(\frac{1}{\gamma_1}\right) + O\left(\frac{1}{\gamma_2}\right)\right)\sqrt{\frac{2\log\frac{1}{\delta}}{n\mathcal{I}_{r^*}}}$$

which is equivalent to the theorem statement. □

6 High Probability Cramér-Rao Bound

Complementing our algorithmic results, we also give new results on *lower bounding* the estimation error in the location parameter model. The celebrated Cramér-Rao bound lower bounds the variance of estimators, which does not readily translate to (tight) lower bounds on the distribution tail of the estimation error. In this section, we show that it is possible to derive a high probability version of the Cramér-Rao lower bound for r-smoothed distributions, where, given a failure probability δ, we lower bound the estimation error to within a $1 + o(1)$-factor of the error predicted by the asymptotic normality of the standard maximum likelihood algorithm, namely the Gaussian with the true parameter as the mean, and variance $1/(n\mathcal{I}_r)$ for estimation using n samples.

Theorem 1.3. *Suppose f_r is an r-smoothed distribution with Fisher information $\mathcal{I}_r$. Given failure probability δ and sample size n, no algorithm can distinguish f_r and $f_r^{2\varepsilon}$ with probability $1 - \delta$, where $\varepsilon = (1 - o(1))\sqrt{2\log\frac{1}{\delta}/(n\mathcal{I}_r)}$. Here, the $o(1)$ term tends to 0 as $\delta \to 0$ and $\log\frac{1}{\delta}/n \to 0$, for a fixed $r^2\mathcal{I}_r$.*

We prove Theorem 1.3 in Appendix C. The high-level technique we use is a standard one, showing that, it is statistically impossible to distinguish two slightly shifted copies of f_r with probability $1 - \delta$, using n samples. The shift corresponds to (twice) the estimation accuracy lower bound. The difficulty lies in getting the right constant.

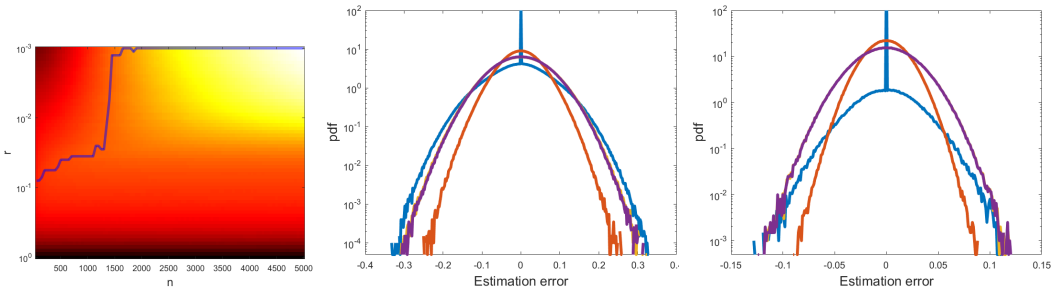
Standard inequalities for showing indistinguishability results rely on calculating either the squared Hellinger distance [BY02] or the KL-divergence [BH79] between the two distributions. While these inequalities are straightforward to apply, given the calculated bounds on these statistical distances/divergences, the inequalities only yield constant-factor tightness in the estimation accuracy lower bound. On the other hand, in this work, we aim to give accuracy upper and lower bounds that are matching strongly, to within $1 + o(1)$ factors. As such, our proof of Theorem 1.3 involves delicate and non-standard bounding techniques which may be of independent interest. The proof techniques are currently slightly ad-hoc, and for future work, we hope to improve on these techniques to make them more general and more usable.

7 Experimental Results

In this section, we give experimental evidence supporting our proposed algorithmic theory. Our goals are to demonstrate that 1) r-smoothing is a beneficial pre-processing to the MLE, that r-smoothed Fisher information does capture the algorithmic performance in location estimation and 2) r-smoothed MLE can outperform the standard MLE, as well as standard mean estimation algorithms which do not leverage information about the distribution shape.

The version of smoothed MLE we use for experimentation is even simpler than Algorithm 1: use Gaussian smoothing before performing actual maximum likelihood finding over the entire real line, instead of returning a root of the empirical score function. This is closest to what statisticians would do in practice, and further does not require any initial uncertainty region on the true parameter λ.

We use the Gaussian-spiked *Laplace* model for experiments, with a Laplace distribution of density proportional to $e^{-|x|}$, and a Gaussian of mass 0.001 and width roughly 0.002 (the discretization granularity) added at $x = 4$. The reason we choose the Laplace over the Gaussian as the "body" of the distribution is because, fixing the variance of the distribution, the Laplace has twice the Fisher information as the Gaussian. Given that standard mean estimation algorithms only aim to achieve sub-Gaussian concentration, choosing the Laplace as the core distribution lets us demonstrate that the smoothed MLE can outperform mean estimation algorithms even in finite samples. We also note that this example is crucially different from the Dirac δ-spiked example in Section 1.1. The Dirac δ spike



(a) Heat map of mean squared error

(b) Estimation error distribution, n = 500

(c) Estimation error distribution, n = 3000

Figure 2: Experimental results - Spiked Laplace model

has infinity density, whereas the narrow Gaussian spike only has somewhat large, but finite density. Given the finite and not too large density in the spike, in our experiments, the basic MLE algorithm will *not* "fit it at any cost", and instead has a smoother error distribution whenever we do not observe samples from the Gaussian spike. Nonetheless, even in this milder setting, we demonstrate that our smoothed MLE algorithm performs better than the original MLE.

Figure 2a is a heat map of the mean squared error of the smoothed MLE. The x-axis varies the number of samples n from 50 to 5000, and the y-axis varies the smoothing parameter r from 0.001 to 1 in log scale. Lighter color indicates a smaller mean squared error. The line overlaid on the heat map indicates, for each value of n, the value of r with the smallest mean squared error. As n increases, the optimal value of r decreases, as predicted by our theory.

For small values of n—below about 1000—the mean squared error first decreases then increases again as we increase the smoothing parameter r. This confirms the theory in the paper: for small n, it is unlikely that we see any samples from the spike, in which case too small values for r cause MLE to overfit. On the other hand, too large values of r simply add too much noise, and also yields a sub-optimal mean squared error. The optimal value of r is thus somewhere in between.

The situation changes when $n \gg 1000$, which is 1 over the mass of the spike. In this case, we expect to typically see samples from the spike, which allows us to estimate the mean highly accurately. Any smoothing just adds noise, and hence the optimal value of r is close to 0.

Figure 2b picks $n = 500$ and compares the distribution of estimation errors across different algorithms: unsmoothed MLE (blue), 0.05-smoothed MLE (orange), empirical mean (yellow), Lee-Valiant (LV) estimator (purple) using $\delta = e^{-5}$. With only 500 samples, the unsmoothed MLE occasionally sees a sample from the spike, and attains high accuracy, but otherwise has large variance in error, compared with the empirical and LV estimators (which have essentially identical performance, so yellow is overlapped by purple on the plot). With just 0.05-smoothing, MLE outperforms all other estimators.

Figure 2c picks $n = 3000$ and compares the same algorithms. The unsmoothed MLE sees samples from the spike most of the time, and attains high accuracy, vastly outperforming the empirical and LV estimators (again, yellow is overlapped by purple). The 0.05-smoothed MLE performs worse than the unsmoothed MLE in the typical case, but has better tail behavior. This plot suggests that the optimal smoothing parameter r in the high probability regime depends on the desired failure probability δ.

8 Future Directions

One natural goal is to extend these techniques to estimate the mean of *unknown* distributions by means of a kernel density estimate (KDE), with accuracy dependent on the true distribution's Fisher information. In general this cannot work, a bias independent of the Fisher information is unavoidable, but for *symmetric* distributions the bias is zero and one can hope for good results. An asymptotic version of this was shown by Stone [Sto75], and we believe our techniques could get a finite-sample guarantee here.

A second direction is to investigate ways to generalize and simplify our lower bound analysis techniques. Recall that, while standard "indistinguishability" bounds based on squared Hellinger distance and KL-divergence are relatively straightforward to apply, they generally lose constant factors. Our analysis is tight to within a $1 + o(1)$-factor, but it requires analyzing several different parameters of the distribution. One can hope to extend and generalize these techniques to yield a new easy-to-apply bound, similar to those based on Hellinger distance and KL-divergence, that gives $1 + o(1)$-factor tightness.

Acknowledgements

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). Jasper C.H. Lee is supported in part by the generous funding of a Croucher Fellowship for Postdoctoral Research and by NSF award DMS-2023239. Paul Valiant is supported by NSF award CCF-2127806.

Bibliography

[BH79] Jean Bretagnolle and Catherine Huber. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(2):119–137, 1979.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[BY02] Ziv Bar-Yossef. *The complexity of massive data set computations*. University of California, Berkeley, 2002.

[Cat12] Olivier Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148 – 1185, 2012.

[LV22] Jasper C.H. Lee and Paul Valiant. Optimal sub-gaussian mean estimation in $\mathbb{R}$. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 672–683, 2022.

[Mia10] Yu Miao. Concentration inequality of maximum likelihood estimator. *Applied Mathematics Letters*, 23(10):1305–1309, 2010.

[Min22] S. Minsker. U-statistics of growing order and sub-gaussian mean estimators with sharp constants. *arXiv preprint arXiv:2202.11842*, 2022.

[Pin17] Iosif Pinelis. 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.

[Spo11] Vladimir Spokoiny. Parametric estimation. finite sample theory. *The Annals of Statistics*, 40, 11 2011.

[Sto75] Charles J Stone. Adaptive maximum likelihood estimators of a location parameter. *The Annals of Statistics*, pages 267–284, 1975.

[SV11] G.L. Shevlyakov and N.O. Vilchevski. *Robustness in Data Analysis: Criteria and Methods*. Modern Probability and Statistics. De Gruyter, 2011.

[vdV00] A.W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

A Proofs from Section 3

We first prove a utility lemma, Lemma A.1, which we use throughout the rest of the paper.

Lemma A.1. *Let f be an arbitrary distribution and let f_r be the r-smoothed version of f. That is, $f_r = \mathbb{E}_{y \leftarrow f}\left[\frac{1}{\sqrt{2\pi r^2}} e^{-\frac{(x-y)^2}{2r^2}}\right]$. Let s_r be the score function of f_r. Let (X, Y, Z_r) be the joint distribution such that $Y \sim f$, $Z_r \sim \mathcal{N}(0, r^2)$ are independent, and $X = Y + Z_r \sim f_r$. We have, for every $\varepsilon > 0$,*

$$\frac{f_r(x+\varepsilon)}{f_r(x)} = \mathop{\mathbb{E}}_{Z_r|x}\left[e^{\frac{2\varepsilon Z_r - \varepsilon^2}{2r^2}}\right] \quad \textit{and in particular} \quad s_r(x) = \mathop{\mathbb{E}}_{Z_r|x}\left[\frac{Z_r}{r^2}\right]$$

Proof. For simplicity of exposition, we only show the case where f has a density. The general case can be proven by, for example, a limit argument. Let w_r be the pdf of $\mathcal{N}(0, r^2)$. First, we show that for any x, ε we have

$$\frac{f_r(x+\varepsilon)}{f_r(x)} = \mathop{\mathbb{E}}_{Z_r|x}\left[\frac{w_r(Z_r+\varepsilon)}{w_r(Z_r)}\right] \tag{2}$$

Denote the density of $(x, z, \widehat{x})$ by $p(\cdot)$. Note that

$$p(z \mid x) = \frac{p(x,z)}{p(x)} = \frac{f(x-z)w_r(z)}{f_r(x)}$$

and hence

$$\begin{aligned} f_r(x+\varepsilon) &= \int_{-\infty}^{\infty} w_r(z) f(x+\varepsilon-z)\,\mathrm{d}z = \int_{-\infty}^{\infty} w_r(z+\varepsilon) f(x-z)\,\mathrm{d}z \\ &= \int_{-\infty}^{\infty} p(z \mid x) f_r(x) \frac{w_r(z+\varepsilon)}{w_r(z)}\,\mathrm{d}z \\ &= f_r(x) \mathop{\mathbb{E}}_{Z|x}\left[\frac{w_r(Z_r+\varepsilon)}{w_r(Z_r)}\right] \end{aligned}$$

proving (2).

Since $w_r(z) = \frac{1}{\sqrt{2\pi r^2}} e^{-\frac{z^2}{2r^2}}$, this gives

$$\frac{f_r(x+\varepsilon)}{f_r(x)} = \mathop{\mathbb{E}}_{Z|x}\left[e^{\frac{2\varepsilon Z_r - \varepsilon^2}{2r^2}}\right].$$

Taking the derivative with respect to ε and evaluating at $\varepsilon = 0$,

$$\frac{f_r'(x)}{f_r(x)} = \mathop{\mathbb{E}}_{Z|x} \frac{Z_r}{r^2}.$$

□

We now prove Lemmas 3.1 and 3.2, which upper and lower bound the r-smoothed Fisher information $\mathcal{I}_r$ respectively.

Lemma 3.1. *Let $\mathcal{I}_r$ be the Fisher information of an r-smoothed distribution f_r. Then, $\mathcal{I}_r \le 1/r^2$.*

Proof. Using Lemma A.1 and Jensen's inequality,

$$I_r = \mathop{\mathbb{E}}_x[s_r^2(x)] = \mathop{\mathbb{E}}_x[(\mathop{\mathbb{E}}_{Z_r|x} Z_r/r^2)^2] \le \mathop{\mathbb{E}}_x[\mathop{\mathbb{E}}_{Z_r|x} Z_r^2/r^4] = 1/r^2$$ □

Lemma 3.2. *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, $\mathcal{I}_r \gtrsim 1/(\mathrm{IQR}+r)^2$. Here, the hidden constant is a universal one independent of the distribution f and independent of r.*

Proof. First, observe that f_r is a smooth distribution in the sense that it is differentiable, and furthermore, its derivative is continuous. Thus, letting R be the 30^{th}-70^{th} percentile range of f_r. Then, by a known result [SV11] (Section 3.1), $I_r \gtrsim 1/R^2$.

Furthermore, it is easy to see via a coupling argument that $R = \text{IQR} + O(r)$. The lemma statement follows. □

Next, we prove another utility lemma, which states that the derivative of the score function cannot be too small for an r-smoothed distribution. Phrased differently, the score function of an r-smoothed distribution cannot decrease fast.

Lemma A.2. *$s_r'(x) \geq -1/r^2$ for all x, where s_r is the score function of f_r, the r-smoothed version of distribution f.*

Proof. By taking the derivative of Lemma A.1 in ε,

$$\frac{f_r'(x+\varepsilon)}{f_r(x)} = \mathop{\mathbb{E}}_{Z|x}\left[e^{\frac{2\varepsilon Z - \varepsilon^2}{2r^2}}\frac{Z_r - \varepsilon}{r^2}\right]$$

Hence

$$s_r(x+\varepsilon) = \frac{f_r'(x+\varepsilon)}{f_r(x+\varepsilon)} = \frac{f_r'(x+\varepsilon)}{f_r(x)}\frac{f_r(x)}{f_r(x+\varepsilon)} = \frac{\mathbb{E}_{Z_r|x}\left[e^{\frac{2\varepsilon Z_r - \varepsilon^2}{2r^2}}\frac{Z_r-\varepsilon}{r^2}\right]}{\mathbb{E}_{Z_r|x}\left[e^{\frac{2\varepsilon Z_r - \varepsilon^2}{2r^2}}\right]}.$$

For $\varepsilon > 0$, since $e^{\frac{2\varepsilon Z_r - \varepsilon^2}{2r^2}}$ and $\frac{Z_r - \varepsilon}{r^2}$ are monotonically increasing in Z_r, and the former is nonnegative, they are positively correlated:

$$\mathop{\mathbb{E}}_{Z_r|x}\left[e^{\frac{2\varepsilon Z_r - \varepsilon^2}{2r^2}}\frac{Z_r-\varepsilon}{r^2}\right] \geq \mathop{\mathbb{E}}_{Z_r|x}\left[e^{\frac{2\varepsilon Z_r - \varepsilon^2}{2r^2}}\right]\mathop{\mathbb{E}}_{Z_r|x}\left[\frac{Z_r-\varepsilon}{r^2}\right]$$

Hence

$$s_r(x+\varepsilon) \geq \mathop{\mathbb{E}}_{Z_r|x}\left[\frac{Z_r-\varepsilon}{r^2}\right] = s_r(x) - \frac{\varepsilon}{r^2}.$$

or (taking $\varepsilon \to 0$), $s_r'(x) \geq -\frac{1}{r^2}$. □

Lastly, we prove the concentration of empirical score function. The way we do so is to show (Lemma A.6) that the k^{th} absolute moment of the score function is upper bounded according to the standard moment bounds for sub-Gamma distributions. As a corollary (Corollary 3.3), we get that the scores have sub-Gamma concentration.

As a utility lemma, we bound the moments of the score function when the score function is aligned with the distribution, instead of being misaligned by some ε distance.

Lemma A.3. *Let s_r be the score function of an r-smoothed distribution f_r with Fisher information $\mathcal{I}_r$. Then, for $k \geq 3$,*

$$\mathop{\mathbb{E}}_{x}[|s_r(x)|^k] \leq (1.6/r)^{k-2}k^{k/2}\mathcal{I}_r$$

Proof. For any x, ε, by Lemma A.1 and Jensen's inequality,

$$f_r(x+\varepsilon) \geq f_r(x)e^{\varepsilon s_r(x) - \frac{\varepsilon^2}{2r^2}}.$$

Setting $\varepsilon = \pm r$ with sign matching $s_r(x)$, we have that

$$f_r(x + r\text{sign}(s_r(x))) \geq f_r(x)e^{r|s_r(x)|}/\sqrt{e}.$$

We also have, from Lemma A.2, that

$$s_r(x-r) \leq s_r(x) + 1/r$$

and

$$s_r(x+r) \geq s_r(x) - 1/r.$$

In other words,

$$|s_r(x + r\text{sign}(s_r(x)))| \geq |s_r(x)| - 1/r.$$

Therefore, for any $k \geq 2$, and $|s_r(x)| > \alpha/r$ for $\alpha := 2 + 1.2\sqrt{k}$,

$$\begin{aligned} f_r(x + r\text{sign}(s_r(x)))|s_r(x + r\text{sign}(s_r(x)))|^k &\geq \frac{1}{\sqrt{e}} f_r(x) e^{r|s_r(x)|} (|s_r(x)| - 1/r)^k \\ &= f_r(x)|s_r(x)|^k \cdot \left(\frac{1}{\sqrt{e}} e^{r|s_r(x)|} (1 - \frac{1}{r|s_r(x)|})^k \right) \\ &\geq f_r(x)|s_r(x)|^k \cdot \left(\frac{1}{\sqrt{e}} e^{\alpha - 1.4\frac{k}{\alpha}} \right) \\ &\geq f_r(x)|s_r(x)|^k \cdot 4. \end{aligned}$$

Therefore

$$f_r(x)|s_r(x)|^k \leq \frac{1}{4} \left(f_r(x - r)|s_r(x - r)|^k + f_r(x + r)|s_r(x + r)|^k \right) \tag{3}$$

whenever $k \geq 2$ and $|s_r(x)| \geq \alpha/r$. Integrating this,

$$\begin{aligned} \mathbb{E}[s_r^k(x)] = \int_{-\infty}^{\infty} f_r(x)|s_r(x)|^k \, \mathrm{d}x &= 2 \int_{-\infty}^{\infty} f_r(x)|s_r(x)|^k - \frac{1}{4} f_r(x - r)|s_r(x - r)|^k - \frac{1}{4} f_r(x + r)|s_r(x + r)|^k \, \mathrm{d}x \\ &\leq 2 \int_{-\infty}^{\infty} f_r(x)|s_r(x)|^k 1_{|s_r(x)| < \alpha/r} \, \mathrm{d}x \\ &\leq 2 \int_{-\infty}^{\infty} f_r(x)|s_r(x)|^2 (\alpha/r)^{k-2} 1_{|s_r(x)| < \alpha/r} \, \mathrm{d}x \\ &\leq 2(\alpha/r)^{k-2} \mathop{\mathbb{E}}_{x}[s_r^2(x)] = 2(\alpha/r)^{k-2} \mathcal{I}_r \end{aligned}$$

Finally, we observe for any $k \geq 2$ that

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

giving the lemma.

□

The proof of Lemma A.6 has the same logical structure as the proof of Lemma A.3, but has further subtleties. The following two lemmas generalize the first step in the proof of Lemma A.3.

Lemma A.4. *Let s_r be the score function of an r-smoothed distribution f_r with Fisher information $\mathcal{I}_r$. For any x, $k \geq 3$ and $0 \leq \varepsilon \leq r/2$, if $s_r(x + \varepsilon) \geq \max(2\sqrt{k} + 2, 9.5)/r$, then*

$$f_r(x)|s_r(x + \varepsilon)|^k \leq \frac{1}{5} \max \left(f_r(x - \varepsilon)|s_r(x - \varepsilon)|^k, f_r(x + \varepsilon + r)|s_r(x + \varepsilon + r)|^k \right)$$

Proof. Let $\alpha := \frac{f_r(x)}{f_r(x+\varepsilon)}$. By Lemma A.1, we have

$$\alpha = \mathop{\mathbb{E}}_{Z_r | x+\varepsilon} \left[e^{\frac{-2\varepsilon Z_r - \varepsilon^2}{2r^2}} \right] \tag{4}$$

We will consider two cases.

When $\log \alpha \leq \frac{3}{4} r s_r(x + \varepsilon) - 2$. First, by Lemma A.1 and Jensen's inequality, we have

$$\frac{f_r(x + \varepsilon + r)}{f_r(x + \varepsilon)} \geq e^{r s_r(x+\varepsilon) - 1/2}$$

We also have, by Lemma A.2,

$$s_r(x + \varepsilon + r) \geq s_r(x + \varepsilon) - 1/r$$

So,

$$f_r(x+\varepsilon+r)|s_r(x+\varepsilon+r)|^k \geq f_r(x+\varepsilon)|s_r(x+\varepsilon)|^k e^{rs_r(x)-1/2}\left(1-\frac{1}{rs_r(x+\varepsilon)}\right)^k$$
$$\geq f_r(x+\varepsilon)|s_r(x+\varepsilon)|^k e^{rs_r(x+\varepsilon)-\frac{k}{rs_r(x+\varepsilon)-1}-1/2}$$

Since $s_r(x+\varepsilon) \geq (2\sqrt{k}+2)/r$,

$$f_r(x+\varepsilon+r)|s_r(x+\varepsilon+r)|^k \geq f_r(x+\varepsilon)|s_r(x+\varepsilon)|^k e^{\frac{3}{4}rs_r(x+\varepsilon)}$$

So, since

$$\alpha = \frac{f_r(x)}{f_r(x+\varepsilon)} \leq e^{\frac{3}{4}rs_r(x+\varepsilon)-2}$$

we have

$$f(x)|s_r(x+\varepsilon)|^k = \alpha f_r(x+\varepsilon)|s_r(x+\varepsilon)|^k \leq \frac{1}{5}f(x+\varepsilon+r)|s_r(x+\varepsilon+r)|^k$$

When $\log\alpha > \frac{3}{4}rs_r(x+\varepsilon)-2$ Evaluating (4) at $x-\varepsilon$ gives

$$\frac{f_r(x-\varepsilon)}{f_r(x)} = \mathop{\mathbb{E}}_{Z_r \mid x}\left[e^{\frac{-2\varepsilon Z_r-\varepsilon^2}{2r^2}}\right]$$

Taking derivative with respect to ε, we have

$$\frac{f_r'(x-\varepsilon)}{f_r(x)} = \mathop{\mathbb{E}}_{Z_r|x}\left[\frac{(Z_r+x)}{r^2}e^{\frac{-2\varepsilon Z_r-\varepsilon^2}{2r^2}}\right]$$

and so by evaluating at $x+\varepsilon$ (to "shift back")

$$\frac{f_r'(x)}{f_r(x+\varepsilon)} = \mathop{\mathbb{E}}_{Z_r|x+\varepsilon}\left[\frac{(Z_r+x+\varepsilon)}{r^2}e^{\frac{-2\varepsilon Z_r-\varepsilon^2}{2r^2}}\right]$$

Define $y = e^{\frac{-2\varepsilon Z_r-2\varepsilon^2}{2r^2}}$, so that $\mathbb{E}_{Z_r|x+\varepsilon}[y] = \alpha e^{\frac{\varepsilon^2}{2r^2}}$, and

$$\frac{(Z_r+\varepsilon)}{r^2}e^{\frac{-2\varepsilon Z_r-\varepsilon^2}{2r^2}} = -\frac{e^{\frac{\varepsilon^2}{2r^2}}}{\varepsilon}y\log y$$

is concave, so by Jensen's inequality

$$\frac{f'(x)}{f(x+\varepsilon)} \leq -\frac{e^{\varepsilon^2/(2r^2)}}{\varepsilon}(e^{-\frac{\varepsilon^2}{2r^2}}\alpha)\log(e^{-\frac{\varepsilon^2}{2r^2}}\alpha) = -\frac{\alpha\log\alpha}{\varepsilon}+\frac{\alpha\varepsilon}{2r^2}$$

So,

$$s_r(x) = \frac{f_r'(x)}{f_r(x)} \leq -\frac{\log\alpha}{\varepsilon}+\frac{\varepsilon}{2r^2}$$

Finally, we move to consider the point $x-\varepsilon$. By Lemma A.2, we have

$$s_r(x-\varepsilon) \leq s_r(x)+\varepsilon/r^2 \leq -\frac{\log\alpha}{\varepsilon}+\frac{3\varepsilon}{2r^2}$$

By Lemma A.1,

$$\frac{f(x-\varepsilon)}{f(x+\varepsilon)} = \mathop{\mathbb{E}}_{Z_r|x+\varepsilon}\left[e^{-\frac{4\varepsilon Z_r-4\varepsilon^2}{2r^2}}\right] = \mathop{\mathbb{E}}_{Z_r|x+\varepsilon}[y^2] \geq \mathop{\mathbb{E}}_{Z_r|x+\varepsilon}[y]^2 = \alpha^2 e^{-\frac{\varepsilon^2}{r^2}}$$

Since $\log\alpha \geq \frac{3}{4}s_r(x+\varepsilon)-2$,

$$-s_r(x-\varepsilon) \geq \frac{\frac{3}{4}rs_r(x+\varepsilon)-2}{\varepsilon}-\frac{3\varepsilon}{2r^2} \geq \frac{3}{2}s_r(x+\varepsilon)-\frac{4.75}{r} \geq s_r(x)$$

where the second inequality comes from the fact that $\frac{3}{4}rs_r(x+\varepsilon)-2>0$ and so the function is decreasing in ε, with minimum evaluated at $\varepsilon = r/2$.

Thus, we have

$$f_r(x-\varepsilon)|s_r(x-\varepsilon)|^k \geq \alpha e^{-\varepsilon^2/r^2}f_r(x)|s_r(x+\varepsilon)|^k$$

Since our assumptions give $\alpha e^{-\varepsilon^2/r^2} \geq e^{5.125}e^{-1/4} \geq 5$, we get the result. □

Lemma A.5. *Let s_r be the score function of an r-smoothed distribution f_r with Fisher information $\mathcal{I}_r$.*

For any x, $k \geq 3$ and $-r/2 \leq \varepsilon \leq 0$, if $s_r(x+\varepsilon) \geq \alpha/r$ for $\alpha = 2 + 1.2\sqrt{k}$, then we have

$$f_r(x)|s_r(x+\varepsilon)|^k \leq \frac{1}{4}\left(f_r(x-r)|s_r(x+\varepsilon-r)|^k + f_r(x+r)|s_r(x+\varepsilon+r)|^k\right)$$

As an immediate corollary, the statement is true also when $\varepsilon \in [0, r/2)$ and $s_r(x) \leq -\alpha/r$.

Proof. For any x, κ, by Lemma A.1 and Jensen's inequality,

$$f_r(x+\kappa) \geq f_r(x) e^{\kappa s_r(x) - \frac{\kappa^2}{2r^2}}.$$

So, setting $\kappa = r$, we have

$$f_r(x+r) \geq f_r(x) e^{r s_r(x)}/\sqrt{e}$$

By Lemma A.2, we have that

$$s_r(x+\varepsilon+r) \geq s_r(x+\varepsilon) - 1/r$$

Since our right hand side is positive by assumption, this is equivalently stated as

$$|s_r(x+\varepsilon+r))| \geq |s_r(x+\varepsilon)| - 1/r$$

When $\varepsilon < 0$, we have, by Lemma A.2, and since $|\varepsilon| \leq r$, $s_r(x) \geq s_r(x+\varepsilon) - 1/r$. So,

$$\begin{aligned}
f_r(x+r)|s_r(x+\varepsilon+r)|^k &\geq \frac{1}{\sqrt{e}} f_r(x) e^{r s_r(x)} (|s_r(x+\varepsilon)| - 1/r)^k \\
&\geq \frac{1}{\sqrt{e}} f_r(x) e^{r(s_r(x+\varepsilon) - 1/r)} (|s_r(x+\varepsilon)| - 1/r)^k \\
&\geq f_r(x)|s_r(x+\varepsilon)|^k \left(\frac{1}{\sqrt{e}} e^{r|s_r(x+\varepsilon)| - 1} \left(1 - \frac{1}{r|s_r(x+\varepsilon)|}\right)^k\right) \\
&\geq f_r(x)|s_r(x+\varepsilon)|^k \cdot \left(e^{-3/2} e^{\alpha - 1.4k/\alpha}\right) \\
&\geq f_r(x)|s_r(x+\varepsilon)|^k \cdot 4 \quad \text{since } k \geq 3
\end{aligned}$$

□

We are now ready to prove Lemma A.6, which states that the distribution of the score function $s_r(x+\varepsilon)$ where $x \sim f_r$ is a sub-Gamma distribution. Corollary 3.3 then states that the average of many score function samples is well-concentrated, following sub-Gamma concentration.

Lemma A.6. *Let s_r be the score function of an r-smoothed distribution f_r with Fisher information $\mathcal{I}_r$. Then, for $k \geq 3$ and $|\varepsilon| \leq r/2$,*

$$\mathop{\mathbb{E}}_{x}[|s_r(x+\varepsilon)|^k] \leq \frac{k!}{2}(15/r)^{k-2} \max(\mathop{\mathbb{E}}_{x}[s_r^2(x+\varepsilon)], \mathcal{I}_r)$$

Equivalently, $s_r(x+\varepsilon)$ is a sub-Gamma random variable.

$$s_r(x+\varepsilon) \in \Gamma(\max(\mathop{\mathbb{E}}_{x}[s_r^2(x+\varepsilon)], \mathcal{I}_r), 15/r).$$

Proof. Without loss of generality we only show the $\varepsilon \geq 0$ case.

Using Lemma A.4 and Lemma A.3, we have

$$\begin{aligned}
&\int_{-\infty}^{\infty} f_r(x-\varepsilon)|s_r(x)|^k \mathbb{1}_{s_r(x) > \max(2\sqrt{k}+2, 9.5)/r} \, \mathrm{d}x \\
&\leq \frac{1}{5}\int_{-\infty}^{\infty} f_r(x-2\varepsilon)|s_r(x-2\varepsilon)|^k + f_r(x+r)|s_r(x+r)|^k \, \mathrm{d}x \\
&= \frac{2}{5}\mathbb{E}[|s_r(x)|^k] \\
&\leq \frac{2}{5}(1.6/r)^{k-2} k^{k/2} \mathcal{I}_r
\end{aligned}$$

We can start bounding the k^{th} moment quantity in the lemma:

$$\begin{aligned}
&\mathbb{E}[|s_r(x+\varepsilon)|^k] \\
&= \int_{-\infty}^{\infty} f_r(x-\varepsilon)|s_r(x)|^k \,\mathrm{d}x \\
&= 2\int_{-\infty}^{\infty} f_r(x-\varepsilon)|s_r(x)|^k - \frac{1}{4} f_r(x-\varepsilon-r)|s_r(x-r)|^k - \frac{1}{4} f_r(x-\varepsilon+r)|s_r(x+r)|^k \,\mathrm{d}x \\
&\le 2\int_{-\infty}^{\infty} f_r(x-\varepsilon)|s_r(x)|^k \mathbb{1}_{s_r(x)\ge -\max(2\sqrt{k}+2,9.5)/r} \,\mathrm{d}x
\end{aligned}$$

where the last inequality follows from (a slight weakening of) Lemma A.5. Now, by the previous claim, we get that

$$\begin{aligned}
&\mathbb{E}[|s_r(x+\varepsilon)|^k] \\
&\le 2\int_{-\infty}^{\infty} f_r(x-\varepsilon)|s_r(x)|^k \mathbb{1}_{|s_r(x)|\le \max(2\sqrt{k}+2,9.5)/r} \,\mathrm{d}x + \frac{4}{5}(1.6/r)^{k-2}k^{k/2}\mathcal{I}_r \\
&\le 2\int_{-\infty}^{\infty} f_r(x-\varepsilon)|s_r(x)|^2 (\max(2\sqrt{k}+2,9.5)/r)^{k-2} \mathbb{1}_{|s_r|\le \max(2\sqrt{k}+2,9.5)/r} \,\mathrm{d}x + \frac{4}{5}(1.6/r)^{k-2}k^{k/2}\mathcal{I}_r \\
&\le 2(\max(2\sqrt{k}+2,9.5)/r)^{k-2} \mathop{\mathbb{E}}_x[|s_r(x+\varepsilon)|^2] + \frac{4}{5}(1.6/r)^{k-2}k^{k/2}\mathcal{I}_r \\
&\le 2k^{k/2}(2.5/r)^{k-2} \mathop{\mathbb{E}}_x[|s_r(x+\varepsilon)|^2] + \frac{4}{5}(1.6/r)^{k-2}k^{k/2}\mathcal{I}_r \\
&\le 3k^{k/2}(2.5/r)^{k-2} \max(\mathop{\mathbb{E}}_x[|s_r(x+\varepsilon)|^2], \mathcal{I}_r) \\
&\le \frac{k!}{2}(15/r)^{k-2} \max(\mathop{\mathbb{E}}_x[|s_r(x+\varepsilon)|^2], \mathcal{I}_r)
\end{aligned}$$

□

Corollary 3.3. *Let f be an arbitrary distribution and let f_r be the r-smoothed version of f. That is, $f_r(x) = \mathbb{E}_{y\leftarrow f}[\frac{1}{\sqrt{2\pi r^2}} e^{-\frac{(x-y)^2}{2r^2}}]$. Consider the parametric family of distributions $f_r^\lambda(x) = f_r(x-\lambda)$. 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 $|\varepsilon| \le r/2$,

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

Proof. Since $\hat{s}(\lambda+\varepsilon) = \frac{1}{n}\sum_{i=1}^n s_r(y_i - \lambda - \varepsilon) = \frac{1}{n}\sum_{i=1}^n s_r(x_i - \varepsilon)$, we know that by Lemma A.6 and the standard algebra of sub-Gamma distributions that $\hat{s}(\lambda+\varepsilon) \in \Gamma(\frac{1}{n}\max(\mathbb{E}_x[s_r^2(x+\varepsilon)], \mathcal{I}_r), 15/r)$. The corollary then follows from the standard Bernstein inequality for sub-Gamma distributions [BLM13]. □

B Proofs omitted in Section 4

We first give the proof of Theorem 1.1, assuming Theorem 4.1.

Proof of Theorem 1.1. It suffices to show that conditions 2) and 3) in the corollary statement implies that each of the following terms from Theorem 4.1 is $1 + O(1/\gamma)$:

- $1 + 1/\log\frac{1}{\delta}$: Note that $\mathcal{I}_r \le \frac{1}{r^2}$ by Lemma 3.1 and so condition 3) implies that $1/\log\frac{1}{\delta} \le (\log\log\frac{1}{\delta})/\log\frac{1}{\delta} \le \frac{1}{\gamma}$

- $1+\rho_r$: $\sqrt{1+O(1/\gamma)} = 1+O(1/\gamma)$. It suffices to check that $\left(\frac{2\log\frac{4\log\frac{1}{\delta}}{r^2\mathcal{I}_r(1-\frac{\beta}{\gamma})\delta}}{n}\right)^{\frac{1}{4}} = O(1/\sqrt{\gamma})$. The fact that $\log\frac{\log\frac{1}{\delta}}{\delta} = O(\log\frac{1}{\delta})$ together with condition 3) imply that the quantity is bounded by $O\left(\frac{(1+O(\gamma))\log\frac{1}{\delta}}{n}\right)^{\frac{1}{4}}$, which in turn is bounded by $O(1/\sqrt{\gamma})$ by condition 2).
- $1/(1-\beta/\gamma) \le 1+O(1/\gamma)$ since β is a constant
- $\sqrt{1+\frac{\log\frac{4\log\frac{1}{\delta}}{r^2\mathcal{I}_r(1-\frac{\beta}{\gamma})}}{\log\frac{1}{\delta}}}$: Note that $\frac{\log\frac{4\log\frac{1}{\delta}}{1-\frac{\beta}{\gamma}}}{\log\frac{1}{\delta}} = O(\frac{\log\log\frac{1}{\delta}}{\log\frac{1}{\delta}}) = O(1/\gamma)$ as before. Also, condition 3) implies that $(\log\frac{1}{r^2\mathcal{I}_r})/(\log\frac{1}{\delta}) \le (\log\frac{1}{r^2\mathcal{I}_r})(\log\log\frac{1}{\delta})/(\log\frac{1}{\delta}) \le 1/\gamma$.

□

The rest of this appendix is on proving Lemma 4.2, which via reparameterization gives Theorem 4.1.

We first show a utility lemma (Lemma B.1), before using it to prove Lemmas B.2 and B.3, which bound the expectation and variance of the empirical score function. After that, we prove Lemma 4.2.

Lemma B.1. *Let w_r be a Gaussian with standard deviation r, f be an arbitrary probability distribution, and f_r be the r-smoothed version of f. Define*

$$\Delta_\varepsilon(x) := \frac{f_r(x+\varepsilon) - f_r(x) - \varepsilon f_r'(x)}{f_r(x)}.$$

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

$$\mathop{\mathbb{E}}_{x\sim f_r}\left[\Delta_\varepsilon(x)^2\right] \lesssim \frac{\varepsilon^4}{r^4}.$$

Proof. By Lemma A.1, we have

$$\Delta_\varepsilon(x) = \frac{f_r(x+\varepsilon) - f_r(x) - \varepsilon f_r'(x)}{f_r(x)} = \mathop{\mathbb{E}}_{Z_r|x}(e^{\frac{2\varepsilon Z_r - \varepsilon^2}{2r^2}} - 1 - \frac{\varepsilon Z_r}{r^2}).$$

Define

$$\alpha_\varepsilon(z) := e^{\frac{2\varepsilon z - \varepsilon^2}{2r^2}} - 1 - \frac{\varepsilon z}{r^2}.$$

We want to bound

$$\begin{aligned}
&\mathop{\mathbb{E}}_{X}\left[\Delta_\varepsilon(x)^2\right] \\
&= \mathop{\mathbb{E}}_{X}\left[\mathop{\mathbb{E}}_{Z_r|X}\left[\alpha_\varepsilon(Z_r)\right]^2\right] \\
&\le \mathop{\mathbb{E}}_{X,Z_r}\left[\left(\alpha_\varepsilon(Z_r)\right)^2\right] \\
&= \mathop{\mathbb{E}}_{Z_r\sim N(0,r^2)}\left(\alpha_\varepsilon(Z_r)\right)^2. \qquad (5)
\end{aligned}$$

Finally, we bound this term (5).

When $|\varepsilon z| \le r^2$, we have by a Taylor expansion that

$$e^{\frac{2\varepsilon z - \varepsilon^2}{2r^2}} = 1 + \frac{\varepsilon z}{r^2} - \frac{\varepsilon^2}{2r^2} + O\left(\left(\frac{2\varepsilon z - \varepsilon^2}{2r^2}\right)^2\right)$$

and so

$$|\alpha_\varepsilon(z)| \lesssim \frac{\varepsilon^2}{r^2} + \left(\frac{\varepsilon z}{r^2}\right)^2$$

or

$$\mathop{\mathbb{E}}_{Z_r\sim N(0,r^2)}\left(\alpha_\varepsilon(Z_r)^2\cdot \mathbb{1}_{|\varepsilon Z_r|\le r^2}\right)^2 \lesssim \frac{\varepsilon^4}{r^4}. \tag{6}$$

On the other hand, for $|\varepsilon z|\ge r^2$,

$$|\alpha_\varepsilon(z)|\le e^{|\frac{\varepsilon z}{r^2}|}$$

so

$$\begin{aligned}
\mathop{\mathbb{E}}_{Z_r\sim N(0,r^2)}\left(\alpha_\varepsilon(Z_r)^2\cdot \mathbb{1}_{|\varepsilon Z_r|\ge r^2}\right)^2 &\le 2\int_{|r^2/\varepsilon|}^\infty \frac{1}{\sqrt{2\pi r^2}} e^{\frac{2|\varepsilon z|}{r^2}} e^{-\frac{z^2}{2r^2}}\,\mathrm{d}z \\
&= 2e^{2\varepsilon^2/r^2}\int_{|r^2/\varepsilon|}^\infty \frac{1}{\sqrt{2\pi r^2}} e^{-\frac{(z-2|\varepsilon|)^2}{2r^2}}\,\mathrm{d}z \\
&\le 2\sqrt{e}\,\Pr[z\ge r^2/|\varepsilon|-2|\varepsilon|] \\
&\lesssim e^{-\frac{(|r^2/\varepsilon|-2|\varepsilon|)^2}{2r^2}} \\
&\le e^{-\frac{r^2}{8\varepsilon^2}} \lesssim \frac{\varepsilon^4}{r^4}.
\end{aligned}$$

Which combines with (5) and (6) to give the result. □

We are now ready to prove Lemma B.2, which bounds the expectation of the empirical score function.

Lemma B.2. *Suppose f_r is an r-smoothed distribution with Fisher information $\mathcal{I}_r$. Then, for any $|\varepsilon|\le r/2$, the expected score $\mathbb{E}_{x\sim f_r}[s_r(x+\varepsilon)]$ satisfies*

$$\mathop{\mathbb{E}}_{x\sim f_r}[s_r(x+\varepsilon)] = -\mathcal{I}\varepsilon + \Theta\left(\sqrt{\mathcal{I}_r}\frac{\varepsilon^2}{r^2}\right)$$

Proof. By definition of s_r,

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim f_r}[s_r(x+\varepsilon)] &= \int_{-\infty}^\infty \frac{f_r(x)f_r'(x+\varepsilon)}{f_r(x+\varepsilon)}\,\mathrm{d}x \\
&= \int_{-\infty}^\infty f_r'(x)\frac{f_r(x-\varepsilon)-f_r(x)}{f_r(x)}\,\mathrm{d}x
\end{aligned}$$

Since by definition of $\mathcal{I}_r$,

$$\mathcal{I}_r := \int_{-\infty}^\infty \frac{f_r'(x)^2}{f_r(x)}\,\mathrm{d}x,$$

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim f_r}[s_r(x+\varepsilon)] + \varepsilon\mathcal{I}_r &= \int_{-\infty}^\infty \frac{f_r'(x)}{f_r(x)}(f_r(x-\varepsilon)-f_r(x)+\varepsilon f_r'(x))\,\mathrm{d}x \\
&= \mathbb{E}\left[s_r(x)\cdot\Delta_{-\varepsilon}(x)\right]
\end{aligned}$$

where $\Delta_\varepsilon(x) := \frac{f_r(x+\varepsilon)-f_r(x)-\varepsilon f_r'(x)}{f_r(x)}$. Thus

$$\begin{aligned}
\left(\mathop{\mathbb{E}}_{x\sim f_r}[s_r(x+\varepsilon)] + \varepsilon\mathcal{I}_r\right)^2 &\le \mathbb{E}\left[s_r(x)^2\right]\mathbb{E}\left[\Delta_{-\varepsilon}(x)^2\right] \\
&= \mathcal{I}_r\,\mathbb{E}\left[\Delta_{-\varepsilon}(x)^2\right]
\end{aligned}$$

By Lemma B.1, we have that

$$\mathbb{E}\left[\Delta_\varepsilon(x)^2\right] \lesssim \frac{\varepsilon^4}{r^4}.$$

and so

$$\left|\mathop{\mathbb{E}}_{x\sim f_r}[s_r(x+\varepsilon)] + \varepsilon\mathcal{I}_r\right| \lesssim \sqrt{\mathcal{I}_r}\frac{\varepsilon^2}{r^2}$$

as desired. □

Lemma B.3. *Suppose f_r is an r-smoothed distribution with Fisher information $\mathcal{I}_r$. Then, for any $|\varepsilon| \le r/2$, the second moment of the score satisfies*

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

Proof. We have that

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim f_r}\left[s_r^2(x+\varepsilon)\right] &= \int_{-\infty}^{\infty} f_r(x)\left(\frac{f_r'(x+\varepsilon)}{f_r(x+\varepsilon)}\right)^2 \mathrm{d}x \\
&= \int_{-\infty}^{\infty} f_r(x-\varepsilon)\left(\frac{f_r'(x)}{f_r(x)}\right)^2 \mathrm{d}x \\
&= \mathcal{I}_r + \int_{-\infty}^{\infty} (f_r(x-\varepsilon)-f_r(x))\left(\frac{f_r'(x)}{f_r(x)}\right)^2 \mathrm{d}x
\end{aligned}$$

By Lemma A.1, we have

$$\frac{f_r(x-\varepsilon)-f_r(x)}{f_r(x)} = \mathop{\mathbb{E}}_{Z_r|x}\left(e^{\varepsilon Z_r/r^2-\varepsilon^2/2r^2}-1\right).$$

We have that

$$\frac{f_r'(x)}{f_r(x)} = \mathop{\mathbb{E}}_{Z_r|x}\frac{Z_r}{r^2}.$$

so that we need to bound

$$\mathop{\mathbb{E}}_{x\sim f_r}\left[s_r^2(x+\varepsilon)\right] - \mathcal{I}_r = \mathop{\mathbb{E}}_{x}\left[\left(\mathop{\mathbb{E}}_{Z_r|x}(e^{\varepsilon Z_r/r^2-\varepsilon^2/2r^2}-1)\right)\left(\mathop{\mathbb{E}}_{Z_r|x}\frac{Z_r}{r^2}\right)^2\right]. \tag{7}$$

We can get bound this as follows. The standard rearrangement inequality states that, if g, h are monotonically non-decreasing functions, $\mathbb{E}_z[g(z)]\,\mathbb{E}_z[h(z)] \le \mathbb{E}[g(z)h(z)]$. Therefore, for any x and parameter $\alpha \lesssim r/\varepsilon$,

$$\begin{aligned}
\left(\mathop{\mathbb{E}}_{Z_r|x}(e^{\varepsilon Z_r/r^2-\varepsilon^2/2r^2}-1)\right)\left(\mathop{\mathbb{E}}_{Z_r|x}\frac{Z_r}{r^2}\right)^2 &\le \left(O(\varepsilon\alpha/r) + \mathop{\mathbb{E}}_{Z_r|x}\mathbb{1}_{Z_r>\alpha r}(e^{\varepsilon Z_r/r^2}-1)\right)\left(\mathop{\mathbb{E}}_{Z_r|x}\frac{Z_r}{r^2}\right)^2 \\
&\le O(\varepsilon\alpha/r)\left(\mathop{\mathbb{E}}_{Z_r|x}\frac{Z_r}{r^2}\right)^2 + \left(\mathop{\mathbb{E}}_{Z_r|x}\mathbb{1}_{Z_r>\alpha r}\frac{Z_r}{r^2}(e^{\varepsilon Z_r/r^2}-1)\right)\left(\mathop{\mathbb{E}}_{Z_r|x}\frac{Z_r}{r^2}\right) \\
&\le O(\varepsilon\alpha/r)\left(\mathop{\mathbb{E}}_{Z_r|x}\frac{Z_r}{r^2}\right)^2 + \mathop{\mathbb{E}}_{Z_r|x}\left(\mathbb{1}_{Z_r>\alpha r}\frac{Z_r^2}{r^4}(e^{\varepsilon Z_r/r^2}-1)\right)
\end{aligned}$$

Thus

$$\mathop{\mathbb{E}}_{x\sim f_r}\left[s_r^2(x+\varepsilon)\right] - \mathcal{I}_r \le O(\varepsilon\alpha/r)\mathcal{I}_r + \mathop{\mathbb{E}}_{Z_r}\left[\mathbb{1}_{Z_r>\alpha r}\frac{Z_r^2}{r^4}(e^{\varepsilon Z_r/r^2}-1)\right].$$

The contribution to the second term from $|Z_r| \lesssim r^2/\varepsilon$ is

$$\varepsilon\mathop{\mathbb{E}}_{Z_r}[\mathbb{1}_{Z_r>\alpha r}\frac{Z_r^3}{r^6}] \lesssim \frac{\varepsilon\alpha^3}{r^3}e^{-\alpha^2/2}.$$

The contribution from $|Z_r| \gtrsim r^2/\varepsilon$ is negligible (as in the proof of Lemma B.1) because the chance of such Z_r is exponentially small:

$$\begin{aligned}
\mathop{\mathbb{E}}_{Z_r}\left[\mathbb{1}_{Z_r \gtrsim \max(r^2/\varepsilon, \alpha r)} \frac{Z_r^2}{r^4}(e^{\varepsilon Z_r/r^2} - 1)\right] &\le \int_{\Omega(r^2/\varepsilon+\alpha r)}^{\infty} \frac{1}{\sqrt{2\pi r^2}} \frac{z^2}{r^4} e^{\frac{\varepsilon z}{r^2}} e^{-\frac{z^2}{2r^2}} \,\mathrm{d}z \\
&= e^{\frac{\varepsilon^2}{2r^2}} \int_{\Omega(r^2/\varepsilon+\alpha r)}^{\infty} \frac{1}{\sqrt{2\pi r^2}} \frac{z^2}{r^4} e^{-\frac{(z-\varepsilon)^2}{2r^2}} \,\mathrm{d}z \\
&\lesssim \int_{\Omega(r^2/\varepsilon+\alpha r)}^{\infty} \frac{1}{\sqrt{2\pi r^2}} \frac{z^2}{r^4} e^{-\frac{z^2}{2r^2}} \,\mathrm{d}z \quad \text{since } \varepsilon \le r/2 \\
&= \frac{1}{r^2} \int_{\Omega(r/\varepsilon+\alpha)}^{\infty} \frac{1}{\sqrt{2\pi}} z^2 e^{-\frac{z^2}{2}} \,\mathrm{d}z \\
&\lesssim \frac{1}{r^2} O\left(e^{-\Omega(r^2/\varepsilon^2+\alpha^2)}\right) \lesssim \frac{1}{r^2}\frac{\varepsilon}{r} e^{-\Omega(\alpha^2)}
\end{aligned}$$

Thus, we have established that

$$\mathop{\mathbb{E}}_{x\sim f_r}\left[s_r^2(x+\varepsilon)\right] - \mathcal{I}_r \lesssim \frac{\varepsilon\alpha}{r}\left(\mathcal{I}_r + \frac{\alpha^2}{r^2} e^{-\alpha^2/2} + \frac{1}{r^2} e^{-\Omega(\alpha^2)}\right)$$

Set $\alpha = O(\sqrt{\log \frac{1}{r^2\mathcal{I}_r}})$ and recall from Lemma 3.1 that $\mathcal{I}_r \le 1/r^2$. Therefore, the first term dominates the right hand side, and the lemma follows.

□

With the above lemmas, we are now ready to prove Lemma 4.2, which we also restate here for the reader's convenience.

Lemma 4.2. *Suppose we have a known model f_r that is the result of r-smoothing with Fisher information $\mathcal{I}_r$, and a given parameter $\varepsilon_{\max}$. Let β and η be the hidden multiplicative constants in Lemmas B.2 and B.3. Further suppose that r satisfies $r \ge 2\varepsilon_{\max}$ and $r^2\sqrt{\mathcal{I}_r} \ge \gamma\varepsilon_{\max}$ for some parameter $\gamma \ge \beta$. Also define the notation $\tilde{\rho}$ (a "$o(1)$" term) by*

$$1 + \tilde{\rho} = \sqrt{1 + \frac{\eta}{\gamma}} + \frac{15}{2\sqrt{\gamma}}\left(\frac{2\log\frac{1}{\delta}}{n}\right)^{\frac{1}{4}}$$

Then, for every $\xi \ll 1$, with probability at least $1 - \delta \cdot \frac{2}{\xi r^2 \mathcal{I}_r (1-\frac{\beta}{\gamma})(1-\delta)}$, for all $\varepsilon \in \left((1+\xi)\frac{1+\tilde{\rho}}{1-\frac{\beta}{\gamma}}\sqrt{\frac{2\log\frac{1}{\delta}}{n\mathcal{I}_r}}, \varepsilon_{\max}\right]$, $\hat{s}(\lambda-\varepsilon)$ is strictly negative and $\hat{s}(\lambda+\varepsilon)$ is strictly positive.

Proof. Without loss of generality, we only show the $\lambda - \varepsilon$ case, and the $\lambda + \varepsilon$ case follows by doubling the failure probability.

Combining Corollary 3.3, Lemmas B.2 and B.3 as well as the assumption that $r^2\sqrt{\mathcal{I}_r} \geq \gamma\varepsilon_{\max}$, we have that, for all $0 < \varepsilon < \min(|r|, \varepsilon_{\max})$, with probability at most δ, we have

$$\begin{aligned}
\hat{s}(\lambda-\varepsilon) - \left(-\mathcal{I}_r\varepsilon + \beta\sqrt{\mathcal{I}_r}\frac{\varepsilon^2}{r^2}\right) &\leq \sqrt{\frac{2\log\frac{1}{\delta}}{n}\mathcal{I}_r}\sqrt{1+\eta\frac{\varepsilon}{r}\sqrt{\log\frac{1}{r^2\mathcal{I}_r}}} + \frac{15\log\frac{1}{\delta}}{nr} \\
&\leq \sqrt{\frac{2\log\frac{1}{\delta}}{n}\mathcal{I}_r}\sqrt{1+\eta\frac{\varepsilon_{\max}}{r}\sqrt{\log\frac{1}{r^2\mathcal{I}_r}}} + \frac{15\log\frac{1}{\delta}}{nr} \\
&\leq \sqrt{\frac{2\log\frac{1}{\delta}}{n}\mathcal{I}_r}\sqrt{1+\eta\frac{\varepsilon_{\max}}{r}\frac{1}{\sqrt{r^2\mathcal{I}_r}}} + \frac{15\log\frac{1}{\delta}}{nr} \\
&\leq \sqrt{1+\frac{\eta}{\gamma}}\sqrt{\frac{2\log\frac{1}{\delta}}{n}\mathcal{I}_r} + \frac{15\log\frac{1}{\delta}}{nr} \\
&\leq \sqrt{1+\frac{\eta}{\gamma}}\sqrt{\frac{2\log\frac{1}{\delta}}{n}\mathcal{I}_r} + \frac{15}{2\sqrt{\gamma}}\left(\frac{2\log\frac{1}{\delta}}{n}\right)^{\frac{1}{4}}\sqrt{\frac{2\log\frac{1}{\delta}}{n}\mathcal{I}_r} \\
&= \left(\sqrt{1+\frac{\eta}{\gamma}} + \frac{15}{2\sqrt{\gamma}}\left(\frac{2\log\frac{1}{\delta}}{n}\right)^{\frac{1}{4}}\right)\sqrt{\frac{2\log\frac{1}{\delta}}{n}\mathcal{I}_r}
\end{aligned}$$

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

For the rest of the proof, we denote the multiplicative term $\sqrt{1+\frac{\eta}{\gamma}} + \frac{15}{2\sqrt{\gamma}}\left(\frac{2\log\frac{1}{\delta}}{n}\right)^{\frac{1}{4}}$ simply by $1+\rho$, as defined in the theorem statement.

We note also that, since $r^2\sqrt{\mathcal{I}_r} \geq \gamma\varepsilon$, we have $-\mathcal{I}_r\varepsilon + \beta\sqrt{\mathcal{I}_r}\frac{\varepsilon^2}{r^2} \leq \left(-1+\frac{\beta}{\gamma}\right)\mathcal{I}_r\varepsilon$. Therefore, for any $0 < \varepsilon < \min(|r|, \varepsilon_{\max})$, we have that with probability at least $1-\delta$,

$$\hat{s}(\lambda-\varepsilon) \leq \left(-1+\frac{\beta}{\gamma}\right)\mathcal{I}_r\varepsilon + (1+\rho_r)\sqrt{\frac{2\log\frac{1}{\delta}}{n}\mathcal{I}_r}$$

By Lemma A.2, we also have that for *any* x

$$\hat{s}'(x) \leq \frac{1}{r^2}$$

Let $\xi \ll 1$ be a parameter that we choose at the end of the proof. We will show that with probability at least $1 - \delta\cdot\frac{1}{\xi r^2\mathcal{I}_r(1-\frac{\beta}{\gamma})(1-\delta)}$, we have for all $\varepsilon \in \left((1+\xi)\frac{1+\rho_r}{1-\frac{\beta}{\gamma}}\sqrt{\frac{2\log\frac{1}{\delta}}{n\mathcal{I}_r}}, \varepsilon_{\max}\right]$, $\hat{s}(\lambda-\varepsilon) < 0$.

Consider a net N of spacing $\xi r^2(1+\rho_r)\sqrt{\frac{2\log\frac{1}{\delta}}{n}\mathcal{I}_r}$ over the interval $\left((1+\xi)\frac{1+\rho_r}{1-\frac{\beta}{\gamma}}\sqrt{\frac{2\log\frac{1}{\delta}}{n\mathcal{I}_r}}, \varepsilon_{\max}\right]$ in the theorem statement. We can check that, if for all points $\varepsilon \in N$, we have

$$\hat{s}(\lambda-\varepsilon) \leq -\xi(1+\rho_r)\sqrt{\frac{2\log\frac{1}{\delta}}{n}\mathcal{I}_r} \tag{8}$$

then, because $\hat{s}' \leq 1/r^2$, we have $\hat{s}(x) < 0$ for all $x \in \lambda + \left((1+\xi)\frac{1+\rho_r}{1-\frac{\beta}{\gamma}}\sqrt{\frac{2\log\frac{1}{\delta}}{n\mathcal{I}_r}}, \varepsilon_{\max}\right]$. This is done by considering two consecutive net points $0 < \varepsilon_1 < \varepsilon_2$, and observing that $\hat{s}(\lambda-\varepsilon) \leq \hat{s}(\lambda-\varepsilon_1) - \frac{\varepsilon_1-\varepsilon}{r^2}$ for $\varepsilon \in (\varepsilon_1, \varepsilon_2]$, which is in turn strictly negative. (For the essentially symmetric case of $\lambda+\varepsilon$, we would instead use the inequality that $\hat{s}(\lambda+\varepsilon) \geq \hat{s}(\lambda+\varepsilon_1) + \frac{\varepsilon_1-\varepsilon}{r^2} > 0$.)

Thus, it suffices to bound the probability that the above inequality holds for all points in N. For a natural number $i \geq 1$, consider the subset N_i of N that intersects with $([i, i+1]+\xi)\frac{1+\rho_r}{1-\frac{\beta}{\gamma}}\sqrt{\frac{2\log\frac{1}{\delta}}{n\mathcal{I}_r}}$,

where here we interpret addition and multiplication as scalar operations on every point in the interval. For each $\varepsilon \in N_i$, Equation 8 holds except for probability δ^{i^2}. Furthermore, each N_i consists of $\frac{1+\rho_r}{1-\frac{\beta}{\gamma}}\sqrt{\frac{2\log\frac{1}{\delta}}{n\mathcal{I}_r}}$ divided by $\xi r^2(1+\rho_r)\sqrt{\frac{2\log\frac{1}{\delta}}{n}\mathcal{I}_r}$ many points, which equals to $\frac{1}{\xi r^2\mathcal{I}_r(1-\frac{\beta}{\gamma})}$ many points. Therefore, the total failure probability is at most $\frac{1}{\xi r^2\mathcal{I}_r(1-\frac{\beta}{\gamma})}\cdot\sum_{i\geq 1}\delta^{i^2} \leq \delta\cdot\frac{1}{\xi r^2\mathcal{I}_r(1-\frac{\beta}{\gamma})(1-\delta)}$. An extra factor of 2 in the failure probability in the theorem statement accounts for the symmetric case of $\hat{s}(\lambda+\varepsilon) > 0$. □

C Proof of Theorem 1.3 in Section 6

The goal of this appendix is to prove Theorem 1.3, which we restate here for the reader's convenience.

Theorem 1.3. *Suppose f_r is an r-smoothed distribution with Fisher information $\mathcal{I}_r$. Given failure probability δ and sample size n, no algorithm can distinguish f_r and $f_r^{2\varepsilon}$ with probability $1-\delta$, where $\varepsilon = (1-o(1))\sqrt{2\log\frac{1}{\delta}/(n\mathcal{I}_r)}$. Here, the $o(1)$ term tends to 0 as $\delta \to 0$ and $\log\frac{1}{\delta}/n \to 0$, for a fixed $r^2\mathcal{I}_r$.*

We use the standard proof technique of reducing distinguishing two "close" distributions to estimation. In particular, we show that it is statistically impossible to distinguish between f_r and $f_r^{2\varepsilon}$ with probability $1-\delta$ using n samples. In order to show such an indistinguishability result, we need the following standard fact (essentially the Neyman-Pearson lemma):

Fact C.1. *Consider a game, where an adversary picks arbitrarily either distribution p or distribution q, and we want an algorithm which, on input n independent samples from the chosen distribution, decide whether the samples came from p or q, succeeding with probability at least $1-\delta$. Then, there is no algorithm $\mathcal{A}$ such that:*

$$\mathbb{P}(\mathcal{A} \text{ returns } p \mid \text{adversary picked } p) - \mathbb{P}(\mathcal{A} \text{ returns } p \mid \text{adversary picked } q) > d_{\mathrm{TV}}(p^{\otimes n}, q^{\otimes n})$$

where $p^{\otimes n}$ denotes the n-fold product distribution of p. In particular, this implies that there is no algorithm $\mathcal{A}$ such that both of the following hold:

- $\mathbb{P}(\mathcal{A} \text{ returns } p \mid \text{adversary picked } p) > \frac{1}{2} + \frac{1}{2}d_{\mathrm{TV}}(p^{\otimes n}, q^{\otimes n})$
- $\mathbb{P}(\mathcal{A} \text{ returns } q \mid \text{adversary picked } q) > \frac{1}{2} + \frac{1}{2}d_{\mathrm{TV}}(p^{\otimes n}, q^{\otimes n})$

So if $d_{\mathrm{TV}}(p^{\otimes n}, q^{\otimes n}) < 1-2\delta$, there is no algorithm that will succeed in distinguishing between two distributions with probability $\geq 1-\delta$ using only n samples.

Thus, we need to upper bound the n-sample total variation distance between f_r and $f_r^{2\varepsilon}$. Standard inequalities for doing so involve calculating and plugging-in the single-sample KL-divergence $D_{\mathrm{KL}}(f_r \,\|\, f_r^{2\varepsilon})$ or squared Hellinger distance $d^2_{\mathrm{H}}(f_r, f_r^{2\varepsilon})$, however, they yield only constant factor tightness in the exponent of $1 - d_{\mathrm{TV}}(p^{\otimes n}, q^{\otimes n})$, and hence only constant factor tightness in sample complexity or estimation error lower bounds. As such, in this paper, we prove a new lemma (Lemma C.2) that involves both the KL-divergence and squared Hellinger distance, as well as assumptions on the concentration of the log-likelihood ratio between f_r and $f_r^{2\varepsilon}$ (which will be satisfied by r-smoothed distributions), which allows us to bound the n-sample total variation distance tightly. After that, we calculate the KL divergence and squared Hellinger distance of f_r and $f_r^{2\varepsilon}$ as well as show the concentration of their log likelihood ratio (Appendix C.1), which when applied to the lemma yields the lower bound result (Appendix C.2).

Lemma C.2. *Consider two arbitrary distributions p, q. Let the log-likelihood ratio be defined as $\gamma = \log\frac{q}{p}$. Suppose there is a parameter $\kappa > 0$ such that we have the following conditions:*

1. $d^2_{\mathrm{H}}(p,q) \leq (1+\kappa)\frac{1}{4}D_{\mathrm{KL}}(p \,\|\, q)$
2. $\frac{D_{\mathrm{KL}}(p\,\|\,q)}{D_{\mathrm{KL}}(q\,\|\,p)} \in [(1+\kappa)^{-1}, 1+\kappa]$
3. $\mathbb{E}_p[\gamma^2] \leq (1+\kappa)2D_{\mathrm{KL}}(p \,\|\, q)$
4. $\mathbb{E}_q[\gamma^2] \leq (1+\kappa)2D_{\mathrm{KL}}(p \,\|\, q)$

5. $\mathbb{E}_p[|\gamma|^k] \le (1+\kappa)\frac{k!}{2} 2D_{\mathrm{KL}}(p \,\|\, q)\kappa^{k-2}$

6. $\mathbb{E}_q[|\gamma|^k] \le (1+\kappa)\frac{k!}{2} 2D_{\mathrm{KL}}(p \,\|\, q)\kappa^{k-2}$

Then, for $\kappa \le 0.01$, $nD_{\mathrm{KL}}(p \,\|\, q) \gg 1$ *and* $D_{\mathrm{KL}}(p \,\|\, q) \ll 1$,

$$1 - d_{\mathrm{TV}}(p^{\otimes n}, q^{\otimes n}) \ge 2e^{-(1+O(\kappa)+O(1/\sqrt{nD_{\mathrm{KL}}(p \,\|\, q)})+O(D_{\mathrm{KL}}(p \,\|\, q)))nD_{\mathrm{KL}}(p \,\|\, q)/4}$$

Proof. Define

$$BC_S(p,q) = \int_S \sqrt{pq} \le \sqrt{p(S)q(S)}$$

to be the restriction of the Bhattacharyya coefficient $BC(p,q) = 1 - d_{\mathrm{H}}^2(p,q)$ to a subset S of the domain. For any S, we have

$$1 - d_{\mathrm{TV}}(p,q) = \int \min(p,q) \ge \int_S \min(p,q) \ge \frac{(\int_S \sqrt{\min(p,q)\max(p,q)})^2}{\int_S \max(p,q)} = \frac{BC_S(p,q)^2}{p(S)+q(S)}.$$

We apply this to $p^{\otimes n}$ and $q^{\otimes n}$, getting for any S:

$$1 - d_{\mathrm{TV}}(p^{\otimes n}, q^{\otimes n}) \ge \frac{BC_S(p^{\otimes n}, q^{\otimes n})^2}{p^{\otimes n}(S) + q^{\otimes n}(S)}. \tag{9}$$

Thus, the goal now is to find an event S such that $BC_S(p^{\otimes n}, q^{\otimes n})$ is big relative to $p^{\otimes n}(S)+q^{\otimes n}(S)$.

For the rest of the proof, we use the notation $\bar{\gamma}$ to denote the n-sample empirical log-likelihood ratio, namely $\frac{1}{n}\sum_i \gamma_i = \frac{1}{n}\sum_i \log \frac{q(x_i)}{p(x_i)}$.

We now define S_k, for $k \in \mathbb{Z}$, to be the event $\{\overline{\gamma} \in [k-\frac{1}{2}, k+\frac{1}{2}] \cdot \alpha D_{\mathrm{KL}}(p \,\|\, q)\}$ for some parameter $\alpha = \Theta(\max(\kappa, 1/\sqrt{nD_{\mathrm{KL}}(p \,\|\, q)}, D_{\mathrm{KL}}(p \,\|\, q)))$, and set $S = S_0$. We have that

$$BC_S(p^{\otimes n}, q^{\otimes n}) = BC(p^{\otimes n}, q^{\otimes n}) - \sum_{k \ne 0} BC_{S_k}(p^{\otimes n}, q^{\otimes n}) \ge BC(p,q)^n - \sum_{k \ne 0} \sqrt{p^{\otimes n}(S_k) q^{\otimes n}(S_k)} \tag{10}$$

Now, define

$$\delta := e^{-n\min(D_{\mathrm{KL}}(p \,\|\, q), D_{\mathrm{KL}}(q \,\|\, p))/4}$$

We note that $\delta \le (BC(p,q)^n)^{(1-O(\kappa)-O(D_{\mathrm{KL}}(p \,||\, q)))}$, as follows:

$$\begin{aligned} BC(p,q) &= 1 - d_{\mathrm{H}}^2(p,q) \\ &\ge 1 - (1+O(\kappa))\frac{1}{4}\min(D_{\mathrm{KL}}(p \,\|\, q), D_{\mathrm{KL}}(q \,\|\, p)) \\ &\ge \exp\left(-(1+O(\kappa)+O(D_{\mathrm{KL}}(p \,\|\, q)))\frac{1}{4}\min(D_{\mathrm{KL}}(p \,\|\, q), D_{\mathrm{KL}}(q \,\|\, p))\right) \end{aligned}$$

where the first inequality follows from conditions 1 and 2 in the lemma statement, and the second inequality follows from the fact that $1 - x = \exp(-(1+\Theta(x))x)$. The above claim follows from raising both sides to the power of n.

We shall now bound $p^{\otimes n}(S_k)$ and $q^{\otimes n}(S_k)$ in terms of δ. By standard sub-Gamma concentration bounds, conditions 3–6 imply that

$$\Pr_p\left[\overline{\gamma} > -D_{\mathrm{KL}}(p \,\|\, q) + t\sqrt{2(1+\kappa)D_{\mathrm{KL}}(p \,\|\, q)} + \frac{\kappa}{2}t^2\right] < e^{-nt^2/2} \tag{11}$$

and

$$\Pr_q\left[\overline{\gamma} < D_{\mathrm{KL}}(q \,\|\, p) - t\sqrt{2(1+\kappa)D_{\mathrm{KL}}(p \,\|\, q)} - \frac{\kappa}{2}t^2\right] < e^{-nt^2/2} \tag{12}$$

We now bound $p^{\otimes n}(S_k)$ by

$$p^{\otimes n}(S_k) \le \Pr_p\left[\overline{\gamma} \ge \left(k - \frac{1}{2}\right)\alpha D_{\mathrm{KL}}(p \,\|\, q)\right]$$

Solving the equation

$$\frac{\kappa}{2}t_k^2 + \sqrt{2(1+\kappa)D_{\mathrm{KL}}(p\,\|\,q)}\,t_k - D_{\mathrm{KL}}(p\,\|\,q) = \left(k-\frac{1}{2}\right)\alpha D_{\mathrm{KL}}(p\,\|\,q)$$

yields

$$t_k = \frac{\sqrt{1+\kappa}}{\kappa}\sqrt{2D_{\mathrm{KL}}(p\,\|\,q)}\left(\sqrt{1+\frac{\kappa}{1+\kappa}\left(1+\left(k-\frac{1}{2}\right)\alpha\right)}-1\right)$$

By Equation 11, $p^{\otimes n}(S_k) \le e^{-nt_k^2/2}$ whenever $(k-\frac{1}{2})\alpha > -1$.

Also observe that, when $\kappa \le 0.01$, the function $\frac{1}{\kappa}(\sqrt{1+\frac{\kappa}{1+\kappa}(1+x)}-1)$ within the range $x \in [-1, 1.01]$ can be lower bounded by simply $\frac{1}{2}(1-2\kappa)(1+x)$. For the range $x \ge 1$, we can lower bound the function by $(1-2\kappa)\sqrt{x}$. This implies that $p^{\otimes n}(S_k)$ can be upper bounded by $e^{-\frac{n}{4}(1-O(\kappa))D_{\mathrm{KL}}(p\,\|\,q)(1+(k-\frac{1}{2})\alpha)^2} \le \delta^{(1-O(\kappa))(1+(k-\frac{1}{2})\alpha)^2}$ when $(k-\frac{1}{2})\alpha \in [-1, 1.01]$, and similarly, upper bounded by $e^{-n(1-O(\kappa))D_{\mathrm{KL}}(p\,\|\,q)(k-\frac{1}{2})\alpha} \le \delta^{4(1-O(\kappa))(k-\frac{1}{2})\alpha}$ when $(k-\frac{1}{2})\alpha \ge 1$. Finally, observe that for $(k-\frac{1}{2})\alpha \le -1$, we can trivially bound $p^{\otimes n}(S_k)$ by 1.

We now bound $q^{\otimes n}(S_k)$ by

$$q^{\otimes n}(S_k) \le \Pr_q\left[\overline{\gamma} \le \left(k+\frac{1}{2}\right)\alpha D_{\mathrm{KL}}(p\,\|\,q)\right]$$

Solving the equation

$$-\frac{\kappa}{2}(t_k')^2 - \sqrt{2(1+\kappa)D_{\mathrm{KL}}(p\,\|\,q)}\,t_k' + D_{\mathrm{KL}}(q\,\|\,p) = \left(k+\frac{1}{2}\right)\alpha D_{\mathrm{KL}}(p\,\|\,q)$$

yields (by condition 2)

$$t_k' \ge \frac{\sqrt{1+\kappa}}{\kappa}\sqrt{2D_{\mathrm{KL}}(p\,\|\,q)}\left(\sqrt{1+\frac{\kappa}{1+\kappa}\left(1-\kappa-\left(k+\frac{1}{2}\right)\alpha\right)}-1\right)$$

By Equation 12, $q^{\otimes n}(S_k) \le e^{-n(t_k')^2/2}$ whenever $(k+\frac{1}{2})\alpha < 1-\kappa$.

We now bound $q^{\otimes n}(S_k)$ similar to how we bounded $p^{\otimes n}(S_k)$. When $\kappa \le 0.01$, the function $\frac{1}{\kappa}(\sqrt{1+\frac{\kappa}{1+\kappa}(1-\kappa+x)}-1)$ within the range $x \in [-1.01, 1-\kappa]$ can be lower bounded by simply $\frac{1}{2}(1-2\kappa)(1-\kappa-x)$. For the range $x \le -1$, we can lower bound the function by $(1-2\kappa)\sqrt{x}$. This implies that $q^{\otimes n}(S_k)$ can be upper bounded by $e^{-\frac{n}{4}(1-O(\kappa))D_{\mathrm{KL}}(p\,\|\,q)(1-\kappa-(k+\frac{1}{2})\alpha)^2} \le \delta^{(1-O(\kappa))(1-\kappa-(k+\frac{1}{2})\alpha)^2}$ when $(k+\frac{1}{2})\alpha \in [-1.01, 1-\kappa]$, and similarly, upper bounded by $e^{n(1-O(\kappa))D_{\mathrm{KL}}(p\,\|\,q)(k+\frac{1}{2})\alpha} \le \delta^{-4(1-O(\kappa))(k+\frac{1}{2})\alpha}$ when $(k+\frac{1}{2})\alpha \le -1$. Finally, observe that for $(k+\frac{1}{2})\alpha \ge 1-\kappa$, we can trivially bound $q^{\otimes n}(S_k)$ by 1.

We are now ready to upper bound $\sum_{k\neq 0}\sqrt{p^{\otimes n}(S_k)q^{\otimes n}(S_k)}$. We decompose this sum into three regions of non-zero k.

The main region K_1 is where $(k-\frac{1}{2})\alpha \ge -1$ and $(k+\frac{1}{2})\alpha \le 1-\kappa$. In this region, $\sqrt{p^{\otimes n}(S_k)q^{\otimes n}(S_k)}$ can be upper bounded by $\delta^{\frac{1-O(\kappa)}{2}\left[(1+(k-\frac{1}{2})\alpha)^2+(1-\kappa-(k+\frac{1}{2})\alpha)^2\right]} \le \delta^{(1-O(\kappa))(1+(|k|-\frac{1}{2})^2\alpha^2))}$. Now observe that $\sum_{k\in K_1}\delta^{(1-O(\kappa))(1+(|k|-\frac{1}{2})^2\alpha^2))} \lesssim \delta^{1-O(\kappa)+\Omega(\alpha)} \le \delta^{1+\Omega(\alpha)}$ as long as $\delta^{O(\alpha^2)} \ll 1$ and $\alpha = \Omega(\kappa)$. These two conditions are satisfied by the choice of $\alpha = \Omega(\max(\kappa, 1/\sqrt{nD_{\mathrm{KL}}(p\,\|\,q)}))$.

The second region K_2 is where $(k-\frac{1}{2})\alpha \le -1$, which also means that $(k+\frac{1}{2})\alpha \le -(1+\Omega(\alpha))$. In this case, we use the bound $p^{\otimes n}(S_k) \le 1$ and $q^{\otimes n}(S_k) \le \delta^{-4(1-O(\kappa))(k+\frac{1}{2})\alpha}$. Thus, in this region, $\sqrt{p^{\otimes n}(S_k)q^{\otimes n}(S_k)}$ is upper bounded $\delta^{-2(1-O(\kappa))(k+\frac{1}{2})\alpha}$. This means that $\sum_{k\in K_2}\sqrt{p^{\otimes n}(S_k)q^{\otimes n}(S_k)} \lesssim \delta^{2(1-O(\kappa))(1-O(\alpha))} \ll \delta^{1+\Omega(\alpha)}$, as long as $\delta^{\alpha} \ll 1$ and as long as $\kappa \ll 1$ and $\alpha \ll 1$. This is again satisfied by our choice of α.

The last region K_3 is where $(k+\frac{1}{2})\alpha \geq 1-\kappa$, which also means that $(k-\frac{1}{2})\alpha \geq (1-O(\kappa)-O(\alpha))$. In this case, we use the bound $p^{\otimes n}(S_k) \leq \delta^{4(1-O(\kappa))(k-\frac{1}{2})\alpha}$ and $q^{\otimes n}(S_k) \leq 1$. Thus, in this region, $\sqrt{p^{\otimes n}(S_k)q^{\otimes n}(S_k)}$ is upper bounded $\delta^{2(1-O(\kappa))(k-\frac{1}{2})\alpha}$. This means that $\sum_{k\in K_3}\sqrt{p^{\otimes n}(S_k)q^{\otimes n}(S_k)} \lesssim \delta^{2(1-O(\kappa))(1-O(\alpha))} \ll \delta^{1+\Omega(\alpha)}$, as long as $\delta^\alpha \ll 1$ and as long as $\kappa \ll 1$ and $\alpha \ll 1$. This is again satisfied by our choice of α.

Summarizing, we have shown that $\sum_{k\neq 0}\sqrt{p^{\otimes n}(S_k)q^{\otimes n}(S_k)} \lesssim \delta^{1+\Omega(\alpha)}$ for $\alpha = \Omega(\max(\kappa, 1/\sqrt{nD_{\mathrm{KL}}(p\,\|\,q)}))$. Furthermore, as $\alpha \geq \Omega(D_{\mathrm{KL}}(p\,\|\,q))$ by construction, the above bound is much less than $(BC(p,q))^n$, since $(BC(p,q))^n \geq \delta^{1+O(\kappa)+O(D_{\mathrm{KL}}(p\,\|\,q))}$ from earlier in this proof. This yields that $BC_S(p^{\otimes n}, q^{\otimes n}) \geq BC(p,q)^n - \sum_{k\neq 0}\sqrt{p^{\otimes n}(S_k)q^{\otimes n}(S_k)} \geq \delta^{1+O(\alpha)}$ since $\alpha = \Omega(D_{\mathrm{KL}}(p\,\|\,q))$.

The last quantities we have to bound are $p^{\otimes n}(S_0)$ and $q^{\otimes n}(S_0)$. These were already bounded in the respective paragraphs bounding $p^{\otimes n}(S_k)$ and $q^{\otimes n}(S_k)$ for general k. When $k=0$, the bounds are at most $\delta^{1+\Omega(\alpha)}$ (again, when $\alpha = \Omega(\kappa)$). Finally, we get that

$$1 - d_{\mathrm{TV}}(p^{\otimes n}, q^{\otimes n}) \geq \frac{(BC_S(p^{\otimes n}, q^{\otimes n}))^2}{p^{\otimes n}(S) + q^{\otimes n}(S)} \gtrsim \frac{\delta^{2+O(\alpha)}}{\delta^{1+\Omega(\alpha)}} = \delta^{1+O(\alpha)}$$

Expanding the definition of δ as well as the choice of $\alpha = \Theta(\max(\kappa, 1/\sqrt{nD_{\mathrm{KL}}(p\,\|\,q)}, D_{\mathrm{KL}}(p\,\|\,q)))$ gives the lemma statement. □

C.1 Showing the conditions for Lemma C.2

In this subsection, we calculate the KL-divergence, squared Hellinger distance, as well as moment bounds for the log-likelihood ratio for f_r and $f_r^{2\varepsilon}$ for a generic r-smoothed distribution f_r.

Lemma C.3. *Consider the parametric family $f_r^\lambda(x) = f_r(x-\lambda)$ for some r-smoothed distribution f_r with Fisher information $\mathcal{I}_r$. Then for $\varepsilon < \frac{r}{4}$,*

$$D_{\mathrm{KL}}(f_r\,\|\,f_r^{2\varepsilon}) = 2\varepsilon^2\mathcal{I}_r\left(1+\Theta\left(\frac{\varepsilon}{r^2\sqrt{\mathcal{I}_r}}\right)\right)$$

Proof. Let $\ell(x) = \log f_r(x)$

$$\begin{aligned}
D_{\mathrm{KL}}(f_r\,\|\,f_r^{2\varepsilon}) &= -\int_{-\infty}^{\infty} f_r(x)\log\frac{f_r^{2\varepsilon}(x)}{f_r(x)}\,\mathrm{d}x \\
&= -\int_{-\infty}^{\infty} f_0(x)\int_x^{x+2\varepsilon}\ell'(y)\,\mathrm{d}y \\
&= -\int_0^{2\varepsilon}\mathop{\mathbb{E}}_{z\leftarrow f_r}\left[s_r(z+y)\right]\,\mathrm{d}y \\
&= \int_0^{2\varepsilon}\mathcal{I}_r y + \Theta\left(\sqrt{\mathcal{I}_r}\frac{y^2}{r^2}\right)\,\mathrm{d}y \\
&= 2\varepsilon^2\mathcal{I}_r + \Theta\left(\sqrt{\mathcal{I}_r}\frac{\varepsilon^3}{r^2}\right) = 2\varepsilon^2\mathcal{I}_r\left(1+\Theta\left(\frac{\varepsilon}{r^2\sqrt{\mathcal{I}_r}}\right)\right)
\end{aligned}$$

where the Θ result is from Lemma B.2.

□

Lemma C.4. *Consider the parametric family $f_r^\lambda(x) = f_r(x-\lambda)$ for some r-smoothed distribution f_r with Fisher information $\mathcal{I}_r$. Then for $\varepsilon \leq r$,*

$$d_{\mathrm{H}}^2(f_r, f_r^{2\varepsilon}) \leq \frac{1}{4}D_{\mathrm{KL}}(f_r\,\|\,f_r^{2\varepsilon}) + O(\frac{\varepsilon^3}{r^3})$$

Proof. Observe that the squared Hellinger distance is

$$d_{\mathrm{H}}^2(f_r, f_r^{2\varepsilon}) = \frac{1}{2}\int_{-\infty}^{\infty}(\sqrt{f_r(x)} - \sqrt{f_r^{2\varepsilon}(x)})^2\,\mathrm{d}x$$
$$= \frac{1}{2}\int_{-\infty}^{\infty} f_r(x)(1 - \sqrt{\frac{f_r^{2\varepsilon}(x)}{f_r(x)}})^2\,\mathrm{d}x$$

Since for $y > 0$ we have $(1-\sqrt{y})^2 \le -\frac{1}{2}(\log y) + \frac{y-1}{2} + \frac{1}{24}(y-1)_+^3$, substituting $y = \frac{f_r^{2\varepsilon}(x)}{f_r(x)}$, the previous expression is at most

$$\frac{1}{2}\int_{-\infty}^{\infty} f_r(x)\left(-\frac{1}{2}\log\frac{f_r^{2\varepsilon}(x)}{f_r(x)} + \frac{1}{2}\frac{f_r^{2\varepsilon}(x)}{f_r(x)} - \frac{1}{2} + \frac{1}{24}\left(\frac{f_r^{2\varepsilon}(x)}{f_r(x)} - 1\right)_+^3\right)dx$$

The first term inside the big parentheses equals $\frac{1}{4}D_{\mathrm{KL}}(f_r \,\|\, f_r^{2\varepsilon})$; the next two terms cancel out (since they each integrate all the probability mass of f); the final, cubic, term we bound now.

We start by bounding the cubic term for a Gaussian g of standard deviation r:

$$\int_{-\infty}^{\infty} g(x)\left(\frac{g(x+2\varepsilon)}{g(x)} - 1\right)_+^3 \mathrm{d}x = O(\frac{\varepsilon^3}{r^3})$$

where the bound is easily computed from the closed form evaluation of the integral, valid while ε is bounded by some fixed multiple of r.

Now the r-smoothed distribution f is just a convex combination of Gaussians of width r, and the desired inequality follows from the observation that the expression $f_r(x)\left(\frac{f_r^{2\varepsilon}(x)}{f_r(x)} - 1\right)_+^3$ is convex in the sense that, in terms of $y, z > 0$, the function $y\left(\frac{z}{y} - 1\right)_+^3$ is a convex 2-variable function. Thus the total contribution of the cubic term is bounded by its value for the Gaussian, namely $O(\frac{\varepsilon^3}{r^3})$.

□

Lemma C.5. *Let $k \ge 3$. For an r-smoothed distribution f_r, let $\gamma = \log\frac{f_r^{2\varepsilon}}{f_r}$. For $\varepsilon \le r$, we have*

$$\mathop{\mathbb{E}}_{p}[|\gamma|^k] \le \frac{k!}{2}(30\varepsilon/r)^{k-2}4\varepsilon^2\mathcal{I}\left(1 + O\left(\frac{\varepsilon}{r}\sqrt{\log\frac{1}{r^2\mathcal{I}}}\right)\right)$$

Proof. Let $\ell(x) = \log f_r(x)$. We have

$$\begin{aligned}
\int_{-\infty}^{\infty} f_r(x)\,|\log f_r(x) - \log f_r(x+2\varepsilon)|^k &= \int_{-\infty}^{\infty} p(x)\left|\int_x^{x+2\varepsilon}\ell'(y)\,\mathrm{d}y\right|^k \mathrm{d}x \\
&\le \varepsilon^{k-1}\int_{-\infty}^{\infty} p(x)\int_x^{x+2\varepsilon}|\ell'(y)|^k\,\mathrm{d}y\,\mathrm{d}x \\
&= (2\varepsilon)^{k-1}\int_0^{\varepsilon}\mathop{\mathbb{E}}_{x}[|s_r(x+y)|^k]\,\mathrm{d}y \\
&\le (2\varepsilon)^{k-1}\frac{k!}{2}(15/r)^{k-2}\int_0^{2\varepsilon}\mathop{\mathbb{E}}_{x}[s_r(x+y)^2]\,\mathrm{d}y \\
&\le (2\varepsilon)^{k-1}\frac{k!}{2}(15/r)^{k-2}\int_0^{2\varepsilon}\mathcal{I} + O\left(\frac{y}{r}\mathcal{I}\sqrt{\log\frac{1}{r^2\mathcal{I}}}\right)\mathrm{d}y \\
&= (2\varepsilon)^{k}\frac{k!}{2}(15/r)^{k-2}\mathcal{I}\left(1 + O\left(\frac{\varepsilon}{r}\sqrt{\log\frac{1}{r^2\mathcal{I}}}\right)\right) \\
&= \frac{k!}{2}(30\varepsilon/r)^{k-2}4\varepsilon^2\mathcal{I}\left(1 + O\left(\frac{\varepsilon}{r}\sqrt{\log\frac{1}{r^2\mathcal{I}}}\right)\right)
\end{aligned}$$

where the first three inequalities are by convexity, by Lemma A.6, and by Lemma B.3. □

Lemma C.6. *For an r-smoothed distribution f_r, let $\gamma = \log \frac{f_r^{2\varepsilon}}{f_r}$. For $\varepsilon \le r$, we have*

$$\mathop{\mathbb{E}}_{f_r}[|\gamma|^2] \le 4\varepsilon^2\mathcal{I}\left(1 + O\left(\frac{\varepsilon}{r}\sqrt{\log\frac{1}{r^2\mathcal{I}}}\right)\right)$$

Proof. Let $\ell(x) = \log f_r(x)$. We have

$$\begin{aligned}
\int_{-\infty}^{\infty} f_r(x)\left|\log f_r(x) - \log f_r(x+2\varepsilon)\right|^2 &= \int_{-\infty}^{\infty} f_r(x)\left|\int_x^{x+2\varepsilon} \ell'(y)\,\mathrm{d}y\right|^2 \mathrm{d}x \\
&\le \varepsilon\int_{-\infty}^{\infty} f_r(x)\int_x^{x+\varepsilon} |\ell'(y)|^2\,\mathrm{d}y\,\mathrm{d}x \\
&= 2\varepsilon\int_0^{2\varepsilon} \mathop{\mathbb{E}}_x[s_r(x+y)^2]\,\mathrm{d}y \\
&\le 2\varepsilon\int_0^{2\varepsilon} \mathcal{I} + O\left(\frac{y}{r}\mathcal{I}\sqrt{\log\frac{1}{r^2\mathcal{I}}}\right)\mathrm{d}y \\
&= 4\varepsilon^2\mathcal{I}\left(1 + O\left(\frac{\varepsilon}{r}\sqrt{\log\frac{1}{r^2\mathcal{I}}}\right)\right)
\end{aligned}$$

where the first two inequalities are by convexity, and by Lemma B.3. □

C.2 Proving Theorem 1.3

We are ready to prove Theorem 1.3.

Proof of Theorem 1.3. We will be applying Lemma C.2 on the distributions f_r and $f_r^{2\varepsilon}$ for an appropriately chosen ε, with $\kappa = O(\frac{\varepsilon}{r}\frac{1}{r^2\mathcal{I}_r})$ (note that by Lemma 3.1, $\mathcal{I}_r \le 1/r^2$ so $r^2\mathcal{I}_r \le 1$).

Lemma C.4 combined with Lemma C.3 show condition 1 on Lemma C.2. Lemma C.3 shows condition 2. Lemma C.6 shows condition 3, and an essentially identical calculation shows condition 4. Lemma C.5 shows condition 5, and again an essentially identical calculation shows condition 6.

Thus, applying Lemma C.2 and Fact C.1, the failure probability of distinguishing $p = f_r$ and $q = f_r^{2\varepsilon}$ is at least

$$e^{-(1+O(\frac{\varepsilon}{r}\frac{1}{r^2\mathcal{I}_r})+O(1/\sqrt{n\varepsilon^2\mathcal{I}_r})+O(\varepsilon^2\mathcal{I}_r))n\varepsilon^2\mathcal{I}_r/2}$$

Picking

$$\varepsilon = \left(1 - O\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}}\frac{1}{r^3\mathcal{I}_r^{1.5}}\right) - O\left(\frac{1}{\log\frac{1}{\delta}}\right) - O\left(\frac{\log\frac{1}{\delta}}{n}\right)\right)\sqrt{\frac{2\log\frac{1}{\delta}}{n\mathcal{I}_r}}$$

yields a failure probability lower bound of δ, thus showing the theorem statement. □