Proceedings of Machine Learning Research vol 195:1–53, 2023

# Finite-Sample Symmetric Mean Estimation with Fisher Information Rate

**Shivam Gupta**

**Jasper C.H. Lee**

**Eric Price**

## Abstract

The mean of an unknown variance-$\sigma^2$ distribution $f$ can be estimated from $n$ samples with variance $\frac{\sigma^2}{n}$ and nearly corresponding subgaussian rate. When $f$ is known up to translation, this can be improved asymptotically to $\frac{1}{n\mathcal{I}}$, where $\mathcal{I}$ is the Fisher information of the distribution. Such an improvement is not possible for general unknown $f$, but Stone (1975) showed that this asymptotic convergence *is* possible if $f$ is *symmetric* about its mean. Stone's bound is asymptotic, however: the $n$ required for convergence depends in an unspecified way on the distribution $f$ and failure probability $\delta$. In this paper we give finite-sample guarantees for symmetric mean estimation in terms of Fisher information. For every $f, n, \delta$ with $n > \log \frac{1}{\delta}$, we get convergence close to a subgaussian with variance $\frac{1}{n\mathcal{I}_r}$, where $\mathcal{I}_r$ is the $r$-*smoothed* Fisher information with smoothing radius $r$ that decays polynomially in $n$. Such a bound essentially matches the finite-sample guarantees in the known-$f$ setting.

**Keywords:** Cramér-Rao; Fisher Information; Kernel Density Estimation

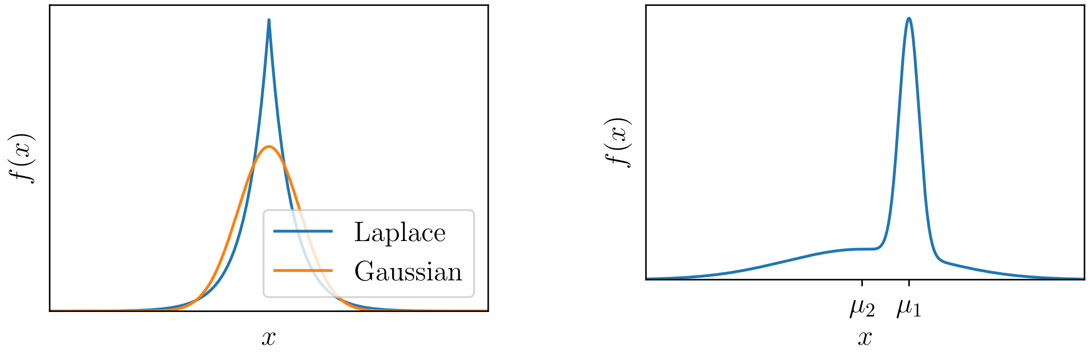
## 1. Introduction

Mean estimation is a fundamental problem in statistics. For a distribution with variance $\sigma^2$, the empirical mean over $n$ samples has variance $\frac{\sigma^2}{n}$ and enjoys central limit behavior, asymptotically yielding error $\sigma\sqrt{2\log\frac{1}{\delta}/n}$ with failure probability $\delta$. Substantial work Catoni (2012); Devroye et al. (2016); Lee and Valiant (2022b) has led to an estimator with a corresponding *finite-sample* guarantee, achieving the same error up to a $1 + o(1)$ factor.

On the other hand, consider the related problem of location estimation: if we know the exact shape of the distribution, except for an unknown translation parameter, the (asymptotic) estimation accuracy is characterized by the Fisher information. More formally, suppose $x \sim f^\lambda(x) = f(x-\lambda)$ for some known $f$ but some unknown parameter $\lambda$. The Fisher information of $f$ is defined as $\mathcal{I} := \mathbb{E}_{x\sim f}[s(x)^2]$ where $s(x)$ is the "score" $s(x) := f'(x)/f(x)$. The *maximum likelihood estimate* (MLE) is asymptotically normal with variance $\frac{1}{n\mathcal{I}}$, which is at most $\frac{\sigma^2}{n}$; and asymptotically, the standard Cramér-Rao bound Cramér et al. (1946); Rao (1945) shows that this is optimal.

For example, the Laplace distribution has Fisher information $\frac{2}{\sigma^2}$, and the MLE for the Laplace is the empirical median. Thus, for the Laplace, the empirical median has half the asymptotic variance of the empirical mean, so it needs half as many samples to achieve the same accuracy. The Fisher information can sometimes be *much* larger than $1/\sigma^2$: consider Figure 2(*b*), a 50-50 mixture of two Gaussians $\frac{1}{2}N(\mu_1, \sigma_1^2) + \frac{1}{2}N(\mu_2, \sigma_2^2)$ with means $\mu_1, \mu_2 \in [-1, 1]$ and variances $\sigma_2^2 \gg 1 \gg \sigma_1^2$.

Figure 1: Example distributions



(*a*) The Laplace distribution has twice the Fisher information of a Gaussian with the same variance, so it can be estimated with asymptotically half the variance.

(*b*) When estimating a mixture of a wide and narrow Gaussian, it is easier to estimate the mean of the narrow Gaussian. When the distribution is known up to location, estimating this mean suffices.

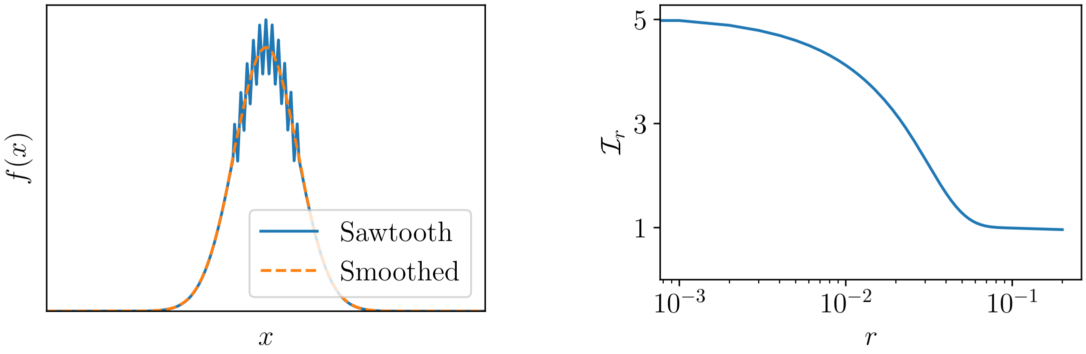
This has variance $\Theta(\sigma_2^2)$ and Fisher information $\Theta(\frac{1}{\sigma_1^2})$. Thus, the empirical mean has accuracy proportional to the *larger* standard deviation, while the MLE has accuracy proportional to the *smaller* standard deviation. In summary, for a *known* $f$ at an unknown offset $\mu$, one can achieve an accuracy based on Fisher information, which is never worse than the generic $\sigma^2$-dependence but can be much better.

This poses a natural question: can we get Fisher-information–style improvements for *unknown* $f$? Unfortunately, the answer is no. In the mixture of Gaussians example of Figure 2(*b*), in the known-distribution case we are given $\mu_2 - \mu_1$ so it suffices to estimate $\mu_1$. This can be done with variance $\Theta(\frac{\sigma_1^2}{n})$. In the unknown-distribution case we need both $\mu_1$ and $\mu_2$, and estimating $\mu_2$ induces variance $\Theta(\frac{\sigma_2^2}{n})$. In fact, recent work has shown (Anonymous, 2023) that the variance-based subgaussian error bounds are essentially instance-optimal: for *every* distribution $p$ of variance $\sigma^2$, and any $n, \delta$ with $n \gg \log \frac{1}{\delta}$, there exists a distribution $q$ of variance $\Theta(\sigma^2)$ where $|\mu_p - \mu_q| = \Omega(\sigma\sqrt{\log \frac{1}{\delta}/n})$, yet $p$ and $q$ are not distinguishable using $n$ samples with probability $1 - \delta$.

In this paper, we consider a restriction that allows for the Fisher information benefit in mean estimation: *symmetry*. We give an estimator that, for every *symmetric* distribution $f$, estimates its mean with an accuracy related to Fisher information.

**Smoothed Fisher information.** To state our results, we need the notion of *smoothed* Fisher information. One issue with the aforementioned Fisher information results is that they are asymptotic: the $n$ required for convergence depends on the distribution in a possibly arbitrary way. As one simple example, if $f(x) = (1 - \varepsilon)N(0, 1) + \varepsilon\delta_0$, the Fisher information is infinite (if we see the same real-valued sample twice, that is the exact mean) but with fewer than $1/\varepsilon$ samples we probably only see the $N(0, 1)$ samples; here the best estimator is the empirical mean, with error $N(0, \frac{1}{n})$. Thus, for finite $n$, one cannot hope for accuracy approaching the true inverse Fisher information of a general distribution.

Figure 2: Gaussian + Sawtooth



(*a*) In the "Gaussian+sawtooth" example, we add tall but narrow "teeth" to a standard Gaussian. Smoothing by radius larger than the width returns the distribution to nearly Gaussian.

(*b*) The smoothed Fisher information has a phase transition, from a large value when $r$ is small, to the standard Gaussian's 1 when $r$ is larger than the tooth width.

Recent work by Gupta et al. (2022, 2023) has given finite-$n$ bounds for the known-distribution case in terms of the "smoothed Fisher information." For a distribution $f$, the $r$-smoothed Fisher information $\mathcal{I}_r$ is the Fisher information of $f$ convolved with a Gaussian of variance $r^2$. In these results, $r \to 0$ as $n \to \infty$, capturing the asymptotic behavior but giving bounds that still apply when $f$ and $n$ vary together.

Figure 2 shows an example based on adding tall but narrow "teeth" to a standard Gaussian. These teeth are useful for alignment *within* the the correct tooth, but not very useful for alignment errors that are integer multiples of the tooth width. As a result, if the tooth width is $w$, the optimal estimator exhibits a phase transition in its variance, with about $\frac{1}{n}$ variance for $n \ll \frac{1}{w^2}$ and $\frac{1}{n\mathcal{I}}$ variance for $n \gg \frac{1}{w^2}$ (see Gupta et al. (2022)). Such a phase transition is captured by the smoothed Fisher information, which transitions at $r \approx w$.

**Our result.** Our main theorem is the following:

**Theorem 1** *Let* $\eta = (\frac{\log \frac{1}{\delta}}{n})^{\frac{1}{13}} < 1$, *and let* $\log \frac{1}{\delta} \leq n/C$ *for sufficiently large constant* $C > 1$. *Let* $f^*$ *be an arbitrary symmetric distribution with variance* $\sigma^2$ *and mean* $\mu$. *For* $\eta\sigma \leq r \leq \sigma$, *we have*

$$|\widehat{\mu} - \mu| \leq (1 + \eta)\sqrt{\frac{2 \log \frac{2}{\delta}}{n\mathcal{I}_r}}$$

*with probability* $1 - \delta$.

For "nice" distributions like the Laplace, $1/\mathcal{I}_r \approx 1/\mathcal{I} + O(r^2)$, so Theorem 1 gives an error bound within $(1 + O((\frac{\log \frac{1}{\delta}}{n})^{1/13}))$ of the instance-optimal Cramér-Rao bound. For other distributions, like the Gaussian+sawtooth example of Figure 2, $\mathcal{I}_r$ exhibits a phase transition and the error does not approach $\mathcal{I}$ until $n$ grows larger than some distribution-dependent quantity; in the sawtooth example, $n/\log\frac{1}{\delta}$ must be at least $O(1/w^{13})$. As discussed above, this is qualitatively correct but with a suboptimal polynomial.

This theorem has the same form as Gupta et al. (2023), except our theorem applies to unknown symmetric $f^*$ while theirs applies to known, possibly asymmetric $f^*$. The differences are (I) their $\varepsilon$ is a better polynomial, $C(\frac{\log\frac{1}{\delta}}{n})^{1/10}$; and (II) their theorem sets $r$ in terms of the interquartile range instead of standard deviation, and so applies to infinite-variance distributions.

Since $\frac{1}{\mathcal{I}_r} \leq \sigma^2 + r^2$, for appropriately chosen $r$ our bound is never more than a $(1+O(\varepsilon))$-factor worse than the subgaussian tail with variance $\frac{\sigma^2}{n}$. This is comparable to the results of Catoni (2012), although with a (slightly) weaker convergence rate (Catoni has rate $\varepsilon = \frac{\log\frac{1}{\delta}}{n}$). However, since Theorem 1 depends on the Fisher information, it can be much better: for example, it gives a factor of $2 - O(\varepsilon)$ improvement when estimating a Laplace distribution, and variance $\Theta\left(\frac{\min(\sigma_1^2,\sigma_2^2)}{n}\right)$ rather than $\Theta\left(\frac{\max(\sigma_1^2,\sigma_2^2)}{n}\right)$ when $f$ is a Gaussian mixture $\frac{1}{2}(N(\mu,\sigma_1^2) + N(\mu,\sigma_2^2))$.

Theorem 1 assumes that we are given $r$; to choose $r$ in general, we would want a (constant-factor) estimate of $\sigma$, which can be done if $f^*$ has bounded kurtosis. Avoiding this dependence is an interesting open question.

Our estimator is based on using a small fraction of samples to construct a kernel density estimate (KDE) of $f$, then finding a variant of the maximum likelihood estimate (MLE). A similar approach was used in Stone (1975) to get an asymptotic bound in terms of $\mathcal{I}$; our contribution is an effective bound for finite $n$ that applies to any distribution,as well as high-probability bounds.

## 2. Related Work

One dimensional mean estimation is one of the most fundamental problems in statistics. Under the assumption of finite variance, the celebrated Central Limit Theorem states that the distribution of the sample mean asymptotically convergences to a Gaussian with variance $\sigma^2/n$. For finite-sample performance, Nemirovsky and Yudin (1983); Jerrum et al. (1986); Alon et al. (1999) independently invented the Median-of-Means estimator, which achieves the same subgaussian concentration up to a constant factor. A decade ago, the seminal work of Catoni (2012) initiated the search for a finite-sample subgaussian estimator with a tight multiplicative constant. Subsequent improvements by Devroye et al. (2016) and Lee and Valiant (2022b) showed how to construct a subgaussian estimator tight up to a $1 + o(1)$ multiplicative factor.

This work, by contrast, assumes symmetry of the distribution about its mean. Stone (1975) showed that asymptotically, the performance of mean estimation for symmetric distributions is controlled by the Fisher information instead of the variance. Our approach is inspired by that of Stone: construct a kernel density estimate (KDE) of the underlying distribution, and perform maximum likelihood estimation (MLE) based on the KDE. On the other hand, our bounds are explicit finite-sample bounds, and characterize the performance in terms of *smoothed* Fisher information, with a smoothing radius $r$ that vanishes as $n/\log\frac{1}{\delta} \to \infty$.

Fisher information also characterizes the asymptotic error in the closely-related problem of location estimation—a parametric variant of mean estimation—under the much stronger assumption that we know the shape of the entire distribution up to some unknown translation van der Vaart (2000). The recent works by Gupta et al. (2022; 2023) developed a finite-sample theory of location estimation with error in terms of the smoothed Fisher information, up to a $1 + o(1)$ factor. Our algorithm also draws from the techniques in this line of work. In particular, instead of finding the maximum of the empirical log-likelihood function, they perform a single step of Newton's method to approximate a root of the derivative. This modification both simplifies the algorithmic implementation and yields analysis advantages. Our algorithm and analysis crucially leverage the same simplified view of the MLE.

The statistics and computer science communities have also been actively studying the high-dimensional mean estimation problem. Lugosi and Mendelson (2019) proposed the first subgaussian high-dimensional mean estimator up to a multiplicative constant, but with exponential time. Hopkins (2020) and Cherapanamjeri et al. (2019) later improved the result to take quadratic time. A tight constant factor was achieved by Lee and Valiant (2022a) in the "very high-dimensional" regime, but it remains an open problem to achieve a subgaussian estimator with tight constants in general.

Recent years have seen a surge of interest in using maximum likelihood in theoretical computer science, as a generic algorithm that can give efficient guarantees. Such papers include, for example, profile maximum likelihood for distribution testing and functional estimation Acharya et al. (2011, 2017); Hao and Orlitsky (2019); Pavlichin et al. (2019); Charikar et al. (2019); Anari et al. (2020); space-efficient streaming algorithms Pettie and Wang (2021); and other statistical estimation problems Daskalakis et al. (2018); Vinayak et al. (2019); Awasthi et al. (2022).

The result of this work has an "instance optimal" flavor: for each distribution, the error bounds are phrased in terms of the (smoothed) Fisher information. The Cramér-Rao bound shows that, even if we knew the distribution shape, we cannot hope to do better than the Fisher information bounds. Instance optimality and related notions have also been studied in the context of other statistical problems, for example, identity testing Valiant and Valiant (2017), learning discrete distributions Valiant and Valiant (2016), mean estimation without symmetry Anonymous (2023) and differentially-private mean estimation Asi and Duchi (2020a,b); Huang et al. (2021).

## 3. Proof Sketch

In this section we give a very high-level overview of our proof approach; for a more detailed quantitative overview, see Section 4. Here, we will describe how to use $(1 + O(\eta))n$ samples to get accuracy $(1 + O(\eta))\sqrt{\frac{2 \log \frac{2}{\delta}}{n}}$ with probability $1 - \delta$, for $\eta = (\log \frac{1}{\delta}/n)^{1/13}$; rescaling parameters gives the result.

Our algorithm proceeds in two phases. In the first phase, we use a small number of samples (namely $\eta n$) to produce an initial estimate $\mu_1$ of $\mu$, and an approximation $\widehat{f_r}$ to $f_r$. Since $f$ is symmetric, we can use the median of pairwise means estimator Minton and Price (2014): $\mu_1 = \text{median}_{i \in [\eta n/2]} \frac{x_{2i-1}+x_{2i}}{2}$. This has subgaussian tails corresponding to the variance of $f$:

$$\varepsilon := \mu_1 - \mu \quad \text{satisfies} \quad |\varepsilon| \lesssim \sqrt{\frac{\sigma^2 \log \frac{2}{\delta}}{\eta n}}$$

with probability $1 - \delta$, for every $\delta > 0$. In the second stage, we want to refine this estimate to $(1 + O(\eta))\sqrt{\frac{\log \frac{2}{\delta}}{n \mathcal{I}_r}}$ error, which is a small polynomial factor better (by at least $\sqrt{\eta}$, but perhaps even better, like $(\frac{n}{\log \frac{1}{\delta}})^{0.1}$). We do so with, essentially, one step of Newton's method.

**Background: known distribution case.** Suppose we knew the distribution of $f_r$, except for the location shift. We consider centering the distribution at $\mu_1 = \mu + \varepsilon$, i.e. define $\widetilde{f}_r(x) = f_r(x - \varepsilon)$, in order to estimate $\varepsilon$. To do so, take the score function

$$\widetilde{s}_r(x) := \frac{\widetilde{f}_r'(x)}{\widetilde{f}_r(x)}$$

which satisfies $\widetilde{s}_r(x + \varepsilon) = s_r(x)$, where $s_r$ is the score function of $f$. Therefore, by standard properties of the score function,

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

Since we know $\widetilde{s}_r$, we can take our $n$ samples $x_i$, add fresh independent noise $w_i \sim N(0, r^2)$ to get $x_i + w_i \sim f_r$, and compute the empirical average

$$\hat{\mathbb{E}}[\widetilde{s}_r(x_i + w_i)] := \frac{1}{n}\sum_{i=1}^{n} \widetilde{s}_r(x_i + w_i)$$

One can show that this concentrates, so by a Taylor approximation

$$\hat{\mathbb{E}}[\widetilde{s}_r(x_i + w_i)] \approx \mathop{\mathbb{E}}_{x\sim f_r}[\widetilde{s}_r(x)] \approx \mathop{\mathbb{E}}_{x\sim f_r}[\widetilde{s}_r(x + \varepsilon) - \varepsilon\tilde{s}_r'(x + \varepsilon)] = -\varepsilon \mathop{\mathbb{E}}_{x\sim f_r}[\tilde{s}_r'(x + \varepsilon)] = \varepsilon\mathcal{I}_r \quad (1)$$

Thus we can estimate $\mu$ as

$$\widehat{\mu} := \mu_1 - \mathcal{I}_r^{-1}\hat{\mathbb{E}}[\widetilde{s}_r(x_i + w_i)] \approx \mu_1 - \varepsilon = \mu.$$

The new estimate $\widehat{\mu}$ has error only from the two approximations in (1): (I) how well the empirical average score concentrates to the true average, and (II) the Taylor approximation.

At $\varepsilon = 0$, error (II) is zero and error (I) has variance $\frac{1}{n}\text{Var}(s_r(x)) = \frac{\mathcal{I}_r}{n}$. Since $\widehat{\mu}$ rescales by $\mathcal{I}_r^{-1}$, this means $\widehat{\mu}$ has variance $\frac{1}{n\mathcal{I}_r}$ at $\varepsilon = 0$—precisely the Cramér-Rao bound we want to achieve. It was shown by Gupta et al. (2022) that the same bound holds to within a $1 + o(1)$ factor as long as $\varepsilon$ is small relative to $r$ (namely, $|\varepsilon| \ll r^2\sqrt{\mathcal{I}_r}$), and that the error satisfies a subgaussian tail bound matching this variance.

**Our setting: unknown distribution case.** The above algorithm for the known-distribution case uses knowledge of the distribution in two ways: to compute $\widetilde{s}$, and to estimate $\mathcal{I}_r$ to rescale it. But what happens if we use some function $g(x)$ other than the score? If $g$ is antisymmetric about $\mu_1$, we still have $\mathbb{E}_{x\sim f_r}[g(x + \varepsilon)] = 0$, and so

$$\hat{\mathbb{E}}[g(x_i + w_i)] \approx \mathop{\mathbb{E}}_{x\sim f_r}[g(x)] \approx \mathop{\mathbb{E}}_{x\sim f_r}[g(x + \varepsilon) - \varepsilon g'(x)] = -\varepsilon \mathop{\mathbb{E}}_{x\sim f_r}[g'(x)] \quad (2)$$

If $g$ is reasonably smooth and $\varepsilon$ is small relative to $r$, the Taylor approximation will be quite good, in which case

$$\widehat{\varepsilon} := -\frac{\hat{\mathbb{E}}[g(x_i + w_i)]}{\mathbb{E}_{x\sim f_r}[g'(x)]} \quad (3)$$

is a low-bias estimator of $\varepsilon$. For small $\varepsilon$, we expect this estimator to have variance close to the variance at $\varepsilon = 0$, which is

$$\mathrm{Var}(\widehat{\mu}) \approx \frac{1}{n} \frac{\mathbb{E}_{f_r}[g^2(y)]}{\mathbb{E}_{f_r}[g'(y)]^2}. \tag{4}$$

One can show that this variance is at least $\frac{1}{n\mathcal{I}_r}$, with minimum achieved when $g$ is the score function $s_r(x) = \frac{f_r'(x)}{f_r(x)}$, matching the Cramér-Rao bound; see Proposition 2 at the end of the section. But this argument is fairly robust: we just need $g(x)$ to be an antisymmetric function that approximates $s_r$ well under these two expectations, and that is robust to perturbations $\varepsilon \ll r$.

Our algorithm then is: using our initial set of $\eta n$ samples in the first stage, we compute the kernel density estimate (KDE)

$$\widehat{f}_r(x) := \frac{1}{n} \sum_{i=1}^{n} \phi\left(\frac{x - x_i}{r}\right)$$

where $\phi(t)$ is the Gaussian density $\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$. This has corresponding score function $\widehat{s}_r(x) = \frac{\widehat{f}_r'(x)}{\widehat{f}_r(x)}$. We first clip the score function to have magnitude at most $T \approx \frac{\sqrt{\log n}}{r}$, and then antisymmetrize this score function about $\mu_1$ by just copying the right side over: setting $\widehat{s}_r^{sym}(x) = \widehat{s}_r^{clip}(2\mu_1 - x)$ for $x \le \mu_1$. This produces the antisymmetric function $\widehat{s}_r^{sym}$ we use as $g$ in the above proof outline.

The final step in our algorithm is that, in order to estimate our target via (3), we need to approximate $\mathbb{E}_{x\sim f_r}[\widehat{s}^{sym\prime}(x)]$. Since $\widehat{s}^{sym}$ is close to the true score $s_r(x)$, this value is close (within $1 + O(\eta)$) to $\mathcal{I}_r$. Thus, we can just make an estimate $\widehat{\mathcal{I}}_r$ of $\mathcal{I}_r$ using the distribution $\widehat{f}_r$.

Our sources of error are the following: (I) the empirical concentration to the expectation of $\widehat{s}^{sym}(x)$; (II) the Taylor approximation in (2); (III) the increase in variance (4) due to $\widehat{s}^{sym}$ not being the exact score; and (IV) error from approximating $\mathbb{E}_{x\sim f_r}[\widehat{s}^{sym\prime}(x)]$ by $\widehat{\mathcal{I}}_r$ in (3).

Unlike in the known-distribution case, clipping is necessary for bounding error (I). The true score concentrates in expectation because $s_r(x)$ is subgamma over $x \sim f_r$. However, $\widehat{s}_r(x)$ may not be so concentrated. Consider the example $f(x) = (1 - \frac{2}{n})N(0,1) + \frac{1}{n}\delta_{-\sqrt{n}} + \frac{1}{n}\delta_{\sqrt{n}}$. In this example, $x \sim f$ is usually constant but has a $\Theta(\frac{1}{n})$ chance of being quite large; in this case, it is likely that the large points will not appear for the KDE but will appear exactly once for the second stage. The KDE then gives them large scores (about $\sqrt{n}/r$), leading to excessive final error ($\Theta(\frac{1}{r\sqrt{n}})$ not $\Theta(\frac{1}{\sqrt{n}})$). Once the scores are clipped, however, we can bound the error (I) with high probability via Bernstein's inequality. The clipping threshold $T$ is large enough to have negligible effect on the expectations (III-IV); since the true score is subgaussian, with high probability it is not clipped. Specifically, in the Gaussian + Symmetric Dirac Deltas example above , the "excess error" in the above constant probability event is now $O(\sqrt{\log \eta n}/(rn))$. Recalling that $r \approx 1/n^{1/13}$ in Theorem 1, the excess error after clipping is $\ll O(1/\sqrt{n})$.

Error (II) is bounded when $\varepsilon$ is small in a similar manner to previous work in the known distribution case. For errors (III) and (IV), we just need to show that, with high probability, our KDE $\widehat{f}_r \approx f_r$ and $\widehat{s}_r^{sym} \approx s_r$, in different metrics but all in expectation over $\widehat{f}$.

**Comparison to Stone (1975).** Our approach, of taking an initial estimate and KDE and refining it with one Newton step, is similar to Stone (1975). The main difference is that our work needs more careful bounds: Stone (1975) shows convergence in probability to $N(0, \frac{1}{n\mathcal{I}})$, which requires fixing the distribution $f$ and failure probability $\delta$ before sending $n \to \infty$. By separating the distribution dependence into $\mathcal{I}_r$, we can express and prove bounds for any $f, n, \delta$.

We end this section with a short proof relating (4) to the score and Fisher information.

**Proposition 2** *For every antisymmetric function $g$ that is continuously differentiable and whose derivative $g'$ is integrable under $f_r$, we have*

$$\frac{\mathbb{E}_{f_r}[g^2(y)]}{\mathbb{E}_{f_r}[g'(y)]^2} \geq \frac{1}{\mathcal{I}_r}$$

*with equality achieved when $g(y) = s_r(y) = f_r'(y)/f_r(y)$.*

**Proof** First, observe that by integration by parts, we have

$$\mathop{\mathbb{E}}_{f_r}[g'(y)] = \int_{\mathbb{R}} f_r(y)g'(y)\,dy = [f_r(y)g(y)]_{-\infty}^{\infty} - \int_{\mathbb{R}} f_r'(y)g(y)\,dy = -\int_{\mathbb{R}} f_r'(y)g(y)\,dy$$

where the last equality is by the symmetry of $f$ and antisymmetry of $g$. Furthermore,

$$\int_{\mathbb{R}} f_r'(y)g(y)\,dy = \int_{\mathbb{R}} \frac{f_r'(y)}{f_r(y)}g(y)f_r(y)\,dy = \mathop{\mathbb{E}}_{f_r}[s_r(y)g(y)]$$

which means

$$\mathop{\mathbb{E}}_{f_r}[g'(y)]^2 = \mathop{\mathbb{E}}_{f_r}[s_r(y)g(y)]^2 \leq \mathop{\mathbb{E}}_{f_r}[s_r(y)^2]\mathop{\mathbb{E}}_{f_r}[g(y)]^2 = \mathcal{I}_r \mathop{\mathbb{E}}_{f_r}[g(y)]^2$$

by Cauchy-Schwarz, with equality achieved when $g(y) = s_r(y)$. ■

## 4. Key Steps in Proof

Here, we highlight the key steps of our proof. For the full proofs, see the Appendix.

**Notation.** Let $f^*$ be an arbitrary symmetric distribution with mean $\mu$ and variance $\sigma^2$, and let $f_r$ be the $r$-smoothed version of $f^*$. Let $s_r$ be the score function of $f_r$, so that $s_r(x) = \frac{f_r'(x)}{f_r(x)}$. Let $\mathcal{I}_r = \mathbb{E}_{x\sim f_r}\left[s_r(x)^2\right] = -\mathbb{E}_{x\sim f_r}\left[s_r'(x)\right]$ be the Fisher information of $f_r$.

Let $w_r$ be the density function of $\mathcal{N}(0, r^2)$. Then, the Kernel Density Estimate (KDE) $\widehat{f}_r$ from $N$ samples $Y_1, \ldots Y_N \sim f^*$ is given by

$$\widehat{f}_r(x) = \frac{1}{N}\sum_{i=1}^{N} w_r(x - Y_i) \tag{5}$$

It has score function $\widehat{s}_r$ with $\widehat{s}_r(x) = \frac{\widehat{f}_r'(x)}{\widehat{f}_r(x)}$. Let $\widehat{s}_r^{\text{clip}}$ be the clipped KDE score from $N$ samples with associated failure probability $\delta$, given by

$$\widehat{s}_r^{\text{clip}}(x) = \text{sign}(\widehat{s}_r(x)) \cdot \min\left(|\widehat{s}_r(x)|, \frac{2}{r}\sqrt{\log\frac{N}{\log\frac{1}{\delta}}}\right) \tag{6}$$

Define the symmetrized clipped KDE score $\widehat{s}_r^{\text{sym}}$ from $N$ samples, symmetrized around $y$, as

$$\widehat{s}_r^{\text{sym}}(x) = \begin{cases} \widehat{s}_r^{\text{clip}}(x) & x \geq y \\ -\widehat{s}_r^{\text{clip}}(2y - x) & x < y \end{cases} \tag{7}$$

In what follows, we first analyze the clipped KDE score (Section 4.1), before showing that symmetrizing it at a $\mu + \varepsilon$ for small $|\varepsilon|$ does not add too much error (Section 4.2). Using similar techniques, we prove that $\mathcal{I}_r$ can be computed directly from the KDE (Section 4.3). Section 4.4 then analyzes the Newton step of the estimation, and finally Section 4.5 assembles all the guarantees into Lemma 10, from which our main result Theorem 1 follows as a corollary.

### 4.1. Clipped KDE Score

We first show that the clipped KDE score $\widehat{s}_r^{\text{clip}}$ approximates the true score $s_r$ in an $\ell_2$ sense:

**Lemma 3 (Clipped KDE score error)** *Let $\widehat{s}_r^{clip}$ be the clipped Kernel Density estimate from $N$ samples, defined in (9). Let $\gamma > C$ be a parameter, for large enough constant $C \geq 1$. Then for any $r \leq \sigma$ and $\frac{N}{\log \frac{1}{\delta}} \geq \left(\frac{\gamma^{5/12}\sigma}{r}\right)^{6+\beta}$ for $\beta > 0$, with probability $1 - \delta$, we have that,*

$$\mathop{\mathbb{E}}_{x \sim f_r}\left[(\widehat{s}_r^{clip}(x) - s_r(x))^2\right] \lesssim \frac{\mathcal{I}_r}{\gamma}$$

*This holds even for asymmetric $f^*$ and $f_r$.*

**Proof Sketch** We refer to the radius $t\sigma_r$ region around the true mean of $f^*$ as the "typical region", and to the region with density at least $\alpha = \frac{1}{t^3\sigma_r}$ as the "large density region". We break up the expectation above into 3 parts: (I) the typical, large density region, (II) the typical, small density region, and (III) the atypical region. We then bound the expectation in each of these regions individually.

To bound the expectation in regions (II) and (III), observe that both regions (II) and (III) have total probability at most $O\left(\frac{1}{t^2}\right)$. For our clipping threshold, we show that the expectation of $\widehat{s}_r^{\text{clip}}(x)^2$ and $s_r$ on a region with this probability is bounded by $O\left(\frac{\mathcal{I}_r}{\gamma}\right)$.

For region (I), we employ a binning argument. We first show that if we fix $x$ with $f_r(x) \geq \alpha$, then, for small enough $\varepsilon$ and for all $|\zeta| \leq |\varepsilon|$, with probability $1-\delta$, $\widehat{s}_r(x+\zeta)$ approximates $s_r(x+\zeta)$ up to error depending on $\varepsilon, \delta, \alpha$ and $N$. That is, our KDE score approximates the true score well within bins of size $\varepsilon$ with probability $1 - \delta$. Then, by union bounding over $O\left(\frac{t\sigma_r}{\varepsilon}\right)$ bins, for appropriately chosen $\varepsilon$, we show that, with probability $1-\delta$ for all $x$ in region (I), $|\widehat{s}_r^{\text{clip}}(x) - s_r(x)| \lesssim \sqrt{\frac{\mathcal{I}_r}{\gamma}}$ so that the expectation of $(\widehat{s}_r^{\text{clip}}(x) - s_r(x))^2$ in region (I) is bounded by $O\left(\frac{\mathcal{I}_r}{\gamma}\right)$.

Putting our bounds together then gives the claim. ■

### 4.2. Symmetrization

This section shows that $\widehat{s}_r^{\text{sym}}$ symmetrized at $\mu + \varepsilon$ for small $\varepsilon$ has mean $\approx \varepsilon\mathcal{I}_r$ and variance $\approx \mathcal{I}_r$.

**Lemma 4 (Symmetrized Clipped KDE score variance)** *Let $\widehat{s}_r^{sym}$ be the symmetrized clipped Kernel Density Estimate score from $N$ samples, symmetrized around $\mu + \varepsilon$ for $|\varepsilon| \leq r/60$, as defined* *in* (12). *Let $\gamma > C$ be a parameter for large enough constant $C$. Then for any $r \leq \sigma$ and $\frac{N}{\log \frac{1}{\delta}} \geq \left(\frac{\gamma^{5/12}\sigma}{r}\right)^{6+\beta}$ for $\beta > 0$, if $|\varepsilon| \leq r^2\sqrt{\frac{\mathcal{I}_r}{\gamma}}$, with probability $1-\delta$,*

$$|\mathop{\mathbb{E}}_{x\sim f_r}\left[\widehat{s}_r^{sym}(x)^2\right] - \mathcal{I}_r| \lesssim \frac{\mathcal{I}_r}{\sqrt{\gamma}}$$

**Proof Sketch** First, we show that

$$\mathop{\mathbb{E}}_{x\sim f_r}\left[(\widehat{s}_r^{\text{sym}}(x) - s_r(x))^2\right] \lesssim \mathop{\mathbb{E}}_{x\sim f_r}\left[(\widehat{s}_r^{\text{clip}}(x) - s_r(x))^2\right]$$

so that by the previous Lemma 3, it's bounded by $O\left(\frac{\mathcal{I}_r}{\gamma}\right)$. Then, we can show the claim using the triangle inequality in $\ell_2$. ■

The Taylor approximation (2) leads to:

**Lemma 5 (Symmetrized Clipped KDE score mean)** *Let $\widehat{s}_r^{sym}$ be the symmetrized clipped Kernel Density Estimate score from $N$ samples, symmetrized around $\mu + \varepsilon$ for $|\varepsilon| \leq r/60$, as defined in* (12). *Let $\gamma > C$ be a parameter for large enough constant $C$. Then for any $r \leq \sigma$ and $\frac{N}{\log \frac{1}{\delta}} \geq \left(\frac{\gamma^{5/12}\sigma}{r}\right)^{6+\beta}$ for $\beta > 0$, if $|\varepsilon| \leq r^2\sqrt{\frac{\mathcal{I}_r}{\gamma}}$, with probability $1-\delta$,*

$$\left|\mathop{\mathbb{E}}_{x\sim f_r}\left[\widehat{s}_r^{sym}(x)\right] - \varepsilon\mathcal{I}_r\right| \lesssim \frac{\varepsilon\mathcal{I}_r}{\sqrt{\gamma}}$$

### 4.3. Estimating $\mathcal{I}_r$

To perform a step of Newton's method, we need an estimate of the Fisher information $\mathcal{I}_r$. We show that $\widehat{\mathcal{I}}_r = \mathbb{E}_{x\sim \widehat{f}_r}\left[\widehat{s}_r^{\text{sym}}(x)^2\right]$ is a good estimate whenever $\widehat{s}_r^{\text{sym}}$ satisfies the conditions above.

**Lemma 6 (Smoothed Fisher information Estimation)** *Let $\gamma \geq C$ for large constant $C \geq 1$ be a parameter. Suppose we have a function $\tilde{s}_r$ that satisfies for $r \leq \sigma$*

$$\left|\mathop{\mathbb{E}}_{x\sim f_r}\left[\tilde{s}_r(x)^2\right] - \mathcal{I}_r\right| \lesssim \frac{\mathcal{I}_r}{\sqrt{\gamma}}$$

*and that $|\tilde{s}_r(x)| \leq \frac{2}{r}\sqrt{\log \frac{N}{\log \frac{1}{\delta}}}$ for all $x$. Let $\widehat{f}_r$ be the kernel density estimate for $f_r$ from $N$ samples, as defined in* (8). *Then, for $\frac{N}{\log \frac{1}{\delta}} \geq \left(\gamma^{5/12}\frac{\sigma}{r}\right)^{6+\beta}$ for some small constant $\beta > 0$, with probability $1-\delta$, we have*

$$\left|\mathop{\mathbb{E}}_{x\sim \widehat{f}_r}\left[\tilde{s}_r(x)^2\right] - \mathcal{I}_r\right| \lesssim \frac{\mathcal{I}_r}{\sqrt{\gamma}}$$

**Proof Sketch** We have

$$\mathop{\mathbb{E}}_{x\sim \widehat{f}_r}\left[\tilde{s}_r(x)^2\right] = \mathop{\mathbb{E}}_{x\sim f_r}\left[\tilde{s}_r(x)^2\right] + \int_{-\infty}^{\infty}\left(\widehat{f}_r(x) - f_r(x)\right)\tilde{s}_r(x)^2 dx$$

As in the proof of Lemma 3, we again break up the integral above into 3 parts and bound each part separately by $O\left(\frac{\mathcal{I}_r}{\sqrt{\gamma}}\right)$. Finally, we make use of our assumption that $\mathbb{E}_{x\sim f_r}[\tilde{s}_r(x)^2] \approx \mathcal{I}_r$ along with our bound on the integral to show the claim. ■

To conclude, we have with high probability $(1-\frac{\delta}{\xi})$ that our KDE satisfies the following:

**Property 7 (KDE Estimation Properties)** *Let $f_r$ be the $r$-smoothed version of symmetric distribution $f^*$ in Algorithm 1, with Fisher information $\mathcal{I}_r$. For parameters $\gamma > C$ for some sufficiently large constant $C$ and $\xi$, $\widehat{s}_r^{sym}$ satisfies that for symmetrization point $\mu_1 = \mu + \varepsilon$,*

$$\left|\mathop{\mathbb{E}}_{x\sim f_r}\left[\widehat{s}_r^{sym}(x)\right] - \varepsilon\mathcal{I}_r\right| \lesssim \frac{\varepsilon\mathcal{I}_r}{\sqrt{\gamma}} \quad \textit{and} \quad \left|\mathop{\mathbb{E}}_{x\sim f_r}\left[\widehat{s}_r^{sym}(x)^2\right] - \mathcal{I}_r\right| \lesssim \frac{\mathcal{I}_r}{\sqrt{\gamma}}$$

*and $|\widehat{s}_r^{sym}(x)| \leq \frac{2}{r}\sqrt{\log\frac{n}{\xi\log\frac{\xi}{\delta}}}$ for all $x$. Furthermore, the Fisher information estimate $\widehat{\mathcal{I}}_r$ satisfies*

$$\left|\widehat{\mathcal{I}}_r - \mathcal{I}_r\right| \lesssim \frac{\mathcal{I}_r}{\sqrt{\gamma}}$$

### 4.4. Local Estimation

We then show that Property 7 implies that Algorithm 1, which does one approximate Newton step, gets high accuracy.

**Algorithm 1** Local Estimation

**Input Parameters:**

- $n$ samples $x_1, \ldots, x_n \sim f^*$, the symmetrized and clipped KDE score function $\widehat{s}_r^{\text{sym}}$, symmetrization point $\mu_1$, Fisher information estimate $\widehat{\mathcal{I}}_r$

1. 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.
2. Compute $\hat{\varepsilon} = \frac{1}{\widehat{\mathcal{I}}_r n}\sum_{i=1}^n \widehat{s}_r^{\text{sym}}(x_i')$. Return $\hat{\mu} = \mu_1 - \hat{\varepsilon}$.

**Lemma 8 (Local Estimation)** *In Algorithm 1, let $f_r$ be the $r$-smoothed version of symmetric distribution $f^*$, with score function $s_r$ and Fisher information $\mathcal{I}_r$. Suppose for parameters $\gamma, \xi$, and symmetrized clipped KDE score $\widehat{s}^{sym}$ symmetrized around $\mu_1$, Property 7 is satisfied. Then, with probability $1-\delta$, the output $\hat{\mu}$ of Algorithm 1 satisfies*

$$|\hat{\mu} - \mu| \leq \left(1 + O\left(\frac{1}{\sqrt{\gamma}}\right)\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n\mathcal{I}_r}} + O\left(\frac{\sqrt{\log\frac{n}{\xi\log\frac{\xi}{\delta}}}}{r\mathcal{I}_r}\cdot\frac{\log\frac{2}{\delta}}{n}\right) + O\left(\frac{\varepsilon}{\sqrt{\gamma}}\right)$$

**Proof Sketch** We bound $\widehat{\mu} - \mu$ using (2) for $g(x) = \widehat{s}_r^{\text{sym}}(x)$. We apply Bernstein's inequality to concentrate $\widehat{\mathbb{E}}_{x\sim f_r}[\widehat{s}_r^{\text{sym}}(x)]$. The first term in our bound is the variance term and the second is the exponential term. The final term in our error bound comes from the difference between $\mathbb{E}[\widehat{s}_r^{\text{sym}}(x)]$ and $\varepsilon\mathcal{I}_r$ bounded by Property 7. ■

### 4.5. Global Estimation

In order to perform our final estimation, we compute an initial estimate $\mu_1$ of $\mu$, and our KDE $\widehat{f_r}$ with associated symmetrized clipped score $\widehat{s}_r^{\text{sym}}$ around $\mu_1$, along with our Fisher information estimate $\widehat{\mathcal{I}}_r$. We combine these with our Local Estimation algorithm to obtain our final estimate $\widehat{\mu}$. Lemma 10 shows our formal guarantee on the performance of our final estimate $\hat{\mu}$. Then,

For our initial estimate $\mu_1$, we make use of the median of pairwise means estimator for symmetric distributions, implied by Minton and Price (2014).

**Lemma 9 (Median of pairwise means estimator)** *Let $X_1, X_2, \dots, X_n$ be drawn from a symmetric distribution with mean $\mu$ and variance $\sigma^2$. For every constant $C_1 > 0$ there exists a constant $C_2$ such that $\widehat{\mu} := \text{median}_{i\in[n/2]} \frac{X_{2i-1}+X_{2i}}{2}$ satisfies*

$$|\widehat{\mu} - \mu| \leq C_2 \sigma \cdot \sqrt{\frac{\log \frac{2}{\delta}}{n}}$$

*with probability $1 - \delta$, for all $\delta$ with $\log \frac{1}{\delta} \leq C_1 n$.*

**Algorithm 2** Global Estimation

**Input parameters:**

- Failure probability $\delta$, Samples $x_1, \dots, x_n \sim f^*$, smoothing parameter $r$, approximation parameter $\xi > 0$.

1. First, use the first $n/\xi$ samples to compute an initial estimate $\mu_1$ of the mean $\mu$ by using the Median-of-pairwise-means estimator in Lemma 9.

2. Use the next $n/\xi$ samples to compute the kernel density estimate $\widehat{f_r}$ of $f_r$ (as defined in (8)), along with the associated symmeterized, clipped KDE score $\widehat{s}_r^{\text{sym}}$ (as defined in (12)), clipped at $\frac{2}{r}\sqrt{\log \frac{n}{\xi \log \frac{\xi}{\delta}}}$ and symmetrized around the initial estimate $\mu_1$. Compute the Fisher information estimate $\widehat{\mathcal{I}}_r = \mathbb{E}_{x\sim\widehat{f_r}}\left[\widehat{s}_r^{\text{sym}}(x)^2\right]$.

3. Run Algorithm 1 using the remaining $n - \frac{2n}{\xi}$ samples, and return the final estimate $\hat{\mu}$.

**Lemma 10 (Global Estimation)** *Let $\xi > C$ for large enough constant $C > 3$ be a parameter, and suppose $\xi \leq \gamma \leq \left(\frac{n}{\xi \log \frac{1}{\delta}}\right)^{2/5-\alpha}$ for constant $\alpha > 0$. For any $r \leq \sigma$ and $\frac{n}{\log \frac{1}{\delta}} \geq \xi \left(\frac{\gamma^{5/12}\sigma}{r}\right)^{6+\alpha}$, with probability $1 - \delta$, Algorithm 2 outputs an estimate $\hat{\mu}$ with*

$$|\hat{\mu} - \mu| \leq \left(1 + O\left(\frac{1}{\sqrt{\gamma}}\right) + O\left(\frac{1}{\xi}\right)\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n\mathcal{I}_r}} + O\left(\frac{\sigma}{r\sqrt{\gamma}}\sqrt{\frac{\xi \log \frac{\xi}{\delta}}{n\mathcal{I}_r}}\right)$$

**Proof Sketch** Combining Lemmas 9, 5, 4, 6, we know that $\widehat{s}_r^{\text{sym}}(x)$ and $\widehat{\mathcal{I}}_r$ computed in Steps 1 and 2 satisfy Property 7 with high probability. Invoking Lemma 8 on Step 3 yields the result. ■

Theorem 1 follows by setting $\gamma = \frac{1}{\eta^2}$, $\xi = \frac{1}{\eta}$, and calculation.

## Acknowledgments

We thank a bunch of people and funding agency.

## References

Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, and Shengjun Pan. Competitive closeness testing. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 47–68, 2011.

Jayadev Acharya, Hirakendu Das, Alon Orlitsky, and Ananda Theertha Suresh. A unified maximum likelihood approach for estimating symmetric properties of discrete distributions. In *International Conference on Machine Learning*, pages 11–21. PMLR, 2017.

Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *J. Comput. Syst. Sci*, 58(1):137–147, 1999.

Nima Anari, Moses Charikar, Kirankumar Shiragur, and Aaron Sidford. Instance based approximations to profile maximum likelihood. *Advances in neural information processing systems*, 33: 20272–20285, 2020.

Anonymous. Neighborhood optimality in mean estimation: Beyond worst-case, beyond sub-gaussian and beyond $1 + \alpha$ moments. In *Concurrently submitted to COLT 2023*, 2023. We share a non-empty subset of authors with this paper.

Hilal Asi and John C Duchi. Near instance-optimality in differential privacy. *arXiv preprint arXiv:2005.10630*, 2020a.

Hilal Asi and John C Duchi. Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms. *Advances in neural information processing systems*, 33:14106–14117, 2020b.

Pranjal Awasthi, Abhimanyu Das, Weihao Kong, and Rajat Sen. Trimmed maximum likelihood estimation for robust generalized linear model. In *Proc. NeurIPS'22*, 2022.

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. doi: 10.1214/11-AIHP454. URL https://doi.org/10.1214/11-AIHP454.

Moses Charikar, Kirankumar Shiragur, and Aaron Sidford. Efficient profile maximum likelihood for universal symmetric property estimation. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 780–791, 2019.

Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L. Bartlett. Fast mean estimation with sub-gaussian rates. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 786–806. PMLR, 25–28 Jun 2019.

H Cramér et al. Mathematical methods of statistics. *Mathematical methods of statistics.*, 1946.

Constantinos Daskalakis, Themis Gouleakis, Chistos Tzamos, and Manolis Zampetakis. Efficient statistics, in high dimensions, from truncated samples. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 639–649. IEEE, 2018.

Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I. Oliveira. Sub-Gaussian mean estimators. *Ann. Stat*, 44(6):2695–2725, 2016.

Shivam Gupta, Jasper C.H. Lee, Eric Price, and Paul Valiant. Finite-sample maximum likelihood estimation of location. In *NeurIPS*, 2022. URL https://arxiv.org/abs/2206.02348.

Shivam Gupta, Jasper C.H. Lee, and Eric Price. High-dimensional location estimation via norm concentration for subgamma vectors, 2023. URL https://arxiv.org/abs/2302.02497.

Yi Hao and Alon Orlitsky. The broad optimality of profile maximum likelihood. *Advances in Neural Information Processing Systems*, 32, 2019.

Samuel B. Hopkins. Mean estimation with sub-Gaussian rates in polynomial time. *The Annals of Statistics*, 48(2):1193 – 1213, 2020. doi: 10.1214/19-AOS1843. URL https://doi.org/10.1214/19-AOS1843.

Ziyue Huang, Yuting Liang, and Ke Yi. Instance-optimal mean estimation under differential privacy. *Advances in Neural Information Processing Systems*, 34:25993–26004, 2021.

Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theor. Comput. Sci*, 43:169–188, 1986.

Jasper C.H. Lee and Paul Valiant. Optimal sub-gaussian mean estimation in very high dimensions. In *Proc. ITCS 2022*, 2022a.

Jasper C.H. Lee and Paul Valiant. Optimal sub-gaussian mean estimation in $\mathbb{R}$. In *Proc. FOCS'21*, pages 672–683, 2022b.

Gábor Lugosi and Shahar Mendelson. Sub-Gaussian estimators of the mean of a random vector. *Ann. Stat.*, 47(2):783–794, 2019.

Gregory T Minton and Eric Price. Improved concentration bounds for count-sketch. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 669–686. SIAM, 2014.

A.S. Nemirovsky and D.B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

Dmitri S Pavlichin, Jiantao Jiao, and Tsachy Weissman. Approximate profile maximum likelihood. *J. Mach. Learn. Res.*, 20(122):1–55, 2019.

Seth Pettie and Dingyu Wang. Information theoretic limits of cardinality estimation: Fisher meets shannon. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 556–569, 2021.

CR Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.*, 37:81–91, 1945.

Charles J Stone. Adaptive maximum likelihood estimators of a location parameter. *The Annals of Statistics*, 3(2):267–284, 1975.

Gregory Valiant and Paul Valiant. Instance optimal learning of discrete distributions. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 142–155, 2016.

Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017.

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

Ramya Korlakai Vinayak, Weihao Kong, Gregory Valiant, and Sham Kakade. Maximum likelihood estimation for learning populations of parameters. In *International Conference on Machine Learning*, pages 6448–6457. PMLR, 2019.

## Appendix A. Definitions

Let $f^*$ be an arbitrary symmetric distribution with mean $\mu$ and variance $\sigma^2$, and let $f_r$ be the $r$-smoothed version of $f^*$ with variance $\sigma_r^2 = \sigma^2 + r^2$. Let $s_r$ be the score function of $f_r$, so that $s_r(x) = \frac{f_r'(x)}{f_r(x)}$. Let $\mathcal{I}_r = \mathbb{E}_{x \sim f_r}\left[s_r(x)^2\right] = -\mathbb{E}_{x \sim f_r}\left[s_r'(x)\right]$ be the Fisher information of $f_r$.

Let $w_r$ be the density function of $\mathcal{N}(0, r^2)$. We recall the definition of the Kernel Density Estimate (KDE) $\widehat{f}_r$ from $N$ samples $Y_1, \dots Y_N \sim f^*$, from (5).

$$\widehat{f}_r(x) = \frac{1}{N}\sum_{i=1}^{N} w_r(x - Y_i) \tag{8}$$

It has score function $\widehat{s}_r$ with $\widehat{s}_r(x) = \frac{\widehat{f}_r'(x)}{\widehat{f}_r(x)}$. We recall the definition of $\widehat{s}_r^{\text{clip}}$, the clipped KDE score from $N$ samples with associated failure probability $\delta$, from (6).

$$\widehat{s}_r^{\text{clip}}(x) = \text{sign}(\widehat{s}_r(x)) \cdot \min\left(|\widehat{s}_r(x)|, \frac{2}{r}\sqrt{\log \frac{N}{\log \frac{1}{\delta}}}\right) \tag{9}$$

We also recall the definition of $\widehat{s}_r^{\text{sym}}$, the symmetrized clipped KDE score from $N$ samples, symmetrized around $y$, from (7).

$$\widehat{s}_r^{\text{sym}}(x) = \begin{cases} \widehat{s}_r^{\text{clip}}(x) & x \geq y \\ -\widehat{s}_r^{\text{clip}}(2y - x) & x < y \end{cases} \tag{10}$$

## Appendix B. Clipped Kernel Density Estimate

In this section, we will analyze the *clipped Kernel Density Estimate* score function $\widehat{s}_r^{\text{clip}}$ of a distribution. Our main result in this section is Lemma 3, which says that $\widehat{s}_r^{\text{clip}}$ is a good approximation to the true score function $s_r$ of the $r$-smoothed distribution in a specific sense.

### B.1. Pointwise guarantees

In this section, we show that for any individual point $x$, the KDE score $\widehat{s}_r(x)$ approximates $s_r(x)$ well, where $s_r$ is the true score function of our $r$-smoothed distribution $f_r$. We begin by showing that the KDE $\widehat{f_r}$ approximates the true density $f_r$ pointwise.

**Lemma 11 (Pointwise density estimate guarantee)** *Let $\widehat{f_r}$ be the kernel density estimate of $f_r$ from $N$ samples $Y_1, \dots, Y_N \sim f^*$, given by*

$$\widehat{f_r}(x) = \frac{1}{N} \sum_{i=1}^{N} w_r(x - Y_i)$$

*where $w_r$ is the pdf of $\mathcal{N}(0, r^2)$. For any fixed $x$, when $N \geq \frac{3 \log \frac{2}{\delta}}{f_r(x) r}$ we have that with probability $1 - \delta$*

$$\left| \widehat{f_r}(x) - f_r(x) \right| \leq \sqrt{\frac{3 f_r(x) \log \frac{2}{\delta}}{N r}}$$

*This holds even for asymmetric $f^*, f_r$.*

**Proof** For every $x$, we have

$$|w_r(x)| \leq \frac{1}{r}$$

So, by multiplicative Chernoff, we have for $0 \leq \varepsilon \leq 1$

$$\Pr_{Y_i \sim f^*} \left[ \left| \frac{1}{N} \sum_{i=1}^{N} w_r(x - Y_i) - \mathop{\mathbb{E}}_{Y \sim f^*}[w_r(x - Y)] \right| \geq \varepsilon \mathop{\mathbb{E}}_{Y \sim f^*}[w_r(x - Y)] \right] \leq 2 \exp\left( -\frac{\varepsilon^2 r N f_r(x)}{3} \right)$$

The claim follows.

■

The next Lemma shows that the derivative of the KDE $\widehat{f_r'}$ approximates the true derivative of the density function $f_r'$ pointwise.

**Lemma 12 (Pointwise density derivative estimate guarantee)** *Let $\widehat{f_r}$ be the kernel density estimate of $f_r$ from $N$ samples $Y_1, \dots, Y_N \sim f^*$. For a fixed $x$, letting $N \geq \frac{\log \frac{1}{\delta}}{r f_r(x)}$, we have that with probability $1 - \delta$,*

$$\left| \widehat{f_r'}(x) - f_r'(x) \right| \lesssim \sqrt{\frac{f_r(x) \log \frac{1}{\delta}}{N r^3}}$$

*This holds even for asymmetric $f^*, f_r$.*

**Proof**

$$\begin{aligned}
&\mathop{\mathbb{E}}_{Y\sim f^*}[w_r'(x-Y_i)^2] = \mathop{\mathbb{E}}_{Y\sim f^*}\left[\left(-w_r(x-Y)\frac{(x-Y)}{r^2}\right)^2\right]\\
&= \frac{1}{r^4}\mathop{\mathbb{E}}_{Y\sim f^*}\left[w_r(x-Y)^2(x-Y)^2\right]\\
&\lesssim \frac{1}{r^3}\mathop{\mathbb{E}}_{Y\sim f^*}[w_r(x-Y)] \quad \text{since } w_r(x-Y)(x-Y)^2 \le \frac{r}{\sqrt{2\pi e}}\\
&= \frac{f_r(x)}{r^3}
\end{aligned}$$

Also, $w_r'(x-Y_i) = -\frac{(x-Y_i)}{\sqrt{2\pi}r^3}e^{-\frac{(x-Y_i)^2}{2r^2}}$ is bounded in $[-1/r^2, 1/r^2]$. So, by Bernstein's inequality,

$$\Pr_{Y_i\sim f^*}\left[\left|\frac{1}{N}\sum_{i=1}^N w_r'(x-Y_i) - \mathop{\mathbb{E}}_{Y\sim f^*}[w_r'(x-Y)]\right| \ge \varepsilon\right] \le 2\exp\left(-\Omega\left(\frac{\varepsilon^2}{\frac{f_r(x)}{Nr^3}+\frac{\varepsilon}{Nr^2}}\right)\right)$$

So,

$$|\widehat{f}_r'(x) - f_r'(x)| \lesssim \sqrt{\frac{f_r(x)\log\frac{1}{\delta}}{Nr^3}} + \frac{\log\frac{1}{\delta}}{Nr^2}$$

Since $N \ge \frac{\log\frac{1}{\delta}}{f_r(x)r}$, the claim follows. ■

Finally, we have the main result of this section, which shows that the KDE score $\widehat{s}_r$ approximates the true score $s_r$ pointwise.

**Lemma 13 (Pointwise Score Estimate Guarantee)** *Let $\widehat{f}_r$ be the kernel density estimate of $f_r$ from $N$ samples $Y_1, \dots, Y_N \sim f^*$. For fixed $x$, $N \ge \frac{6\log\frac{4}{\delta}}{rf_r(x)}$, and the KDE score $\widehat{s}_r$ defined in (8), given by*

$$\widehat{s}_r(x) = \frac{\widehat{f}_r'(x)}{\widehat{f}_r(x)}$$

*we have that with probability $1-\delta$,*

$$|\widehat{s}_r(x) - s_r(x)| \lesssim \sqrt{\frac{\log\frac{1}{\delta}\log\frac{1}{rf_r(x)}}{Nr^3 f_r(x)}}$$

*This holds even for asymmetric $f^*$, $f_r$.*

**Proof** We have, by Lemmas 11 and 12, by a union bound, with probability $1-\delta$,

$$\frac{\widehat{f}_r'(x)}{\widehat{f}_r(x)} \le \frac{f_r'(x) + O\left(\sqrt{\frac{f_r(x)\log\frac{1}{\delta}}{Nr^3}}\right)}{f_r(x)\left(1-\sqrt{\frac{3\log\frac{4}{\delta}}{f_r(x)rN}}\right)}$$

Since $N \geq \frac{6 \log \frac{4}{\delta}}{r f_r(x)}$ we have that $\sqrt{\frac{3 \log \frac{4}{\delta}}{f_r(x) r N}} \leq 1/\sqrt{2}$. So,

$$\begin{aligned}\widehat{s}_r(x) - s_r(x) &\lesssim \sqrt{\frac{\log \frac{1}{\delta}}{f_r(x) N r^3}} + s_r(x)\sqrt{\frac{\log \frac{1}{\delta}}{f(x) r N}} \\ &\lesssim \sqrt{\frac{\log \frac{1}{\delta} \log \frac{1}{r f_r(x)}}{N r^3 f_r(x)}} \quad \text{by Lemma 33}\end{aligned}$$

Similarly,

$$\widehat{s}_r(x) - s_r(x) \gtrsim \sqrt{\frac{\log \frac{1}{\delta} \log \frac{1}{r f_r(x)}}{N r^3 f_r(x)}}$$

■

## B.2. Close-by scores are close

In this section, we show that for small enough $\varepsilon$, $s_r(x + \varepsilon)$ and $s_r(x)$ are close to each other for the score function $s_r$. We begin with the following utility lemma.

**Lemma 14** *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 $X = Y + Z_r \sim f_r$. For any $x$, and $t > 0$, we have*

$$\Pr_{Z_r}[|Z_r|> rt | X = x] \leq \frac{e^{-t^2/2}}{\sqrt{2\pi} r f_r(x)}$$

*This holds even for asymmetric $f^*$, $f_r$.*

**Proof** Recall that

$$f_r(x) = \mathbb{E}_{Y \sim f^*}[w_r(x - Y)]$$

where $w_r$ is the pdf of $\mathcal{N}(0, r^2)$. So, we have

$$\Pr_{X, Z_r}(\{X = x\} \cap \{|Z_r|> rt\}) = \mathbb{E}_{Y \sim f^*}\left[w_r(x - Y) 1_{|x - Y| > rt}\right] \leq \frac{1}{\sqrt{2\pi} r} e^{-t^2/2}$$

Since $w_r(x - Y) = \frac{1}{\sqrt{2\pi} r} e^{-\frac{(x-Y)^2}{2r^2}}$. Thus

$$\Pr_{Z_r}[|Z_r|> rt | X = x] = \frac{\Pr_{X, Z_r}[\{X = x\} \cap \{|Z_r|> rt\}]}{\Pr_{X \sim f_r}(X = x)} \leq \frac{e^{-t^2/2}}{f_r(x)\sqrt{2\pi} r}.$$

■

The next lemma shows that $f_r(x + \varepsilon)$ is close to $f_r(x)$ for small $\varepsilon$.

**Lemma 15** *Let $\varepsilon, x$ be such that $\frac{8|\varepsilon|}{r}\sqrt{\log \frac{1}{r f_r(x)}} \leq 1$. We have,*

$$\frac{f_r(x+\varepsilon)}{f_r(x)} \leq 1 + \frac{10|\varepsilon|}{r}\sqrt{\log \frac{1}{f_r(x) r}}$$

*This holds even for asymmetric $f^*, f_r$.*

**Proof** 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 $X = Y + Z_r \sim f_r$. For every $x$, by Lemma 36, we have

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

Without loss of generality, we assume that $\varepsilon > 0$. Now, since $\frac{8|\varepsilon|}{r}\sqrt{\log \frac{1}{r f_r(x)}} \leq 1$,

$$\begin{aligned}
\mathop{\mathbb{E}}_{Z_r|x}\left[e^{\varepsilon Z_r / r^2}\right] &\leq \left(1 + \frac{8\varepsilon}{r}\sqrt{\log \frac{1}{f_r(x) r}}\right) + \int_{1+\frac{8\varepsilon}{r}\sqrt{\log \frac{1}{f_r(x) r}}}^{\infty} \mathop{\Pr}_{Z_r|x}\left[e^{\varepsilon z/r^2} \geq u\right] du \\
&= 1 + \frac{8\varepsilon}{r}\sqrt{\log \frac{1}{f_r(x) r}} + \int_{1+\frac{8\varepsilon}{r}\sqrt{\log \frac{1}{f_r(x) r}}}^{\infty} \mathop{\Pr}_{Z_r|x}\left[Z_r \geq \frac{r^2 \log u}{\varepsilon}\right] du \\
&= 1 + \frac{8\varepsilon}{r}\sqrt{\log \frac{1}{r f_r(x)}} + \int_{\log\left(1+\frac{8\varepsilon}{r}\sqrt{\log \frac{1}{r f_r(x)}}\right)}^{\infty} \mathop{\Pr}_{Z_r|x}\left[Z_r \geq rv\right] \frac{\varepsilon}{r} e^{\varepsilon v / r} dv \\
&\quad \left(\text{Substituting } v = \frac{r \log u}{\varepsilon}\right) \\
&\leq 1 + \frac{8\varepsilon}{r}\sqrt{\log \frac{1}{f_r(x) r}} + \int_{4\sqrt{\log \frac{1}{f_r(x) r}}}^{\infty} \mathop{\Pr}_{Z_r|x}[Z_r \geq rv] \frac{|\varepsilon|}{r} e^{\varepsilon v / r} dv \\
&\leq 1 + \frac{8\varepsilon}{r}\sqrt{\log \frac{1}{f_r(x) r}} + \frac{|\varepsilon|}{\sqrt{2\pi}} \int_{4\sqrt{\log \frac{1}{f_r(x) r}}}^{\infty} \frac{e^{-v^2/2 + \varepsilon v / r}}{r^2 f_r(x)} dv \quad \text{by Lemma 14} \\
&\leq 1 + \frac{8\varepsilon}{r}\sqrt{\log \frac{1}{f_r(x) r}} + \frac{|\varepsilon| e^{\frac{2\varepsilon^2}{r^2}}}{r^2 f_r(x)} \int_{4\sqrt{\log \frac{1}{f_r(x) r}}}^{\infty} \frac{e^{-\frac{\left(v - \frac{2\varepsilon}{r}\right)^2}{2}}}{\sqrt{2\pi}} dv \\
&\leq 1 + \frac{8\varepsilon}{r}\sqrt{\log \frac{1}{f_r(x) r}} + \frac{|\varepsilon| e^{\frac{2\varepsilon^2}{r^2}}}{r^2 f_r(x)} \mathop{\Pr}_{W \sim \mathcal{N}(0,1)}\left[W \geq 4\sqrt{\log \frac{1}{f_r(x) r}} - \frac{2\varepsilon}{r}\right] \\
&\leq 1 + \frac{8\varepsilon}{r}\sqrt{\log \frac{1}{f_r(x) r}} + \frac{11|\varepsilon|}{10 r^2 f_r(x)} \mathop{\Pr}_{W \sim \mathcal{N}(0,1)}\left[W \geq \sqrt{2 \log \frac{1}{r f_r(x)}}\right] \\
&\quad \text{since } f_r(x) \leq \frac{1}{\sqrt{2\pi} r}, \text{ so that } \frac{2|\varepsilon|}{r} \leq \frac{1}{4\sqrt{\log \frac{1}{r f_r(x)}}} \leq \frac{1}{4 \log \sqrt{2\pi}}\sqrt{\log \frac{1}{r f_r(x)}} \\
&\quad \text{and since } |\varepsilon|/r \leq \frac{1}{8\sqrt{\log \frac{1}{r f_r(x)}}} \leq \frac{1}{8\sqrt{\log \sqrt{2\pi}}} \text{ so that } e^{2\frac{\varepsilon^2}{r^2}} \leq 11/10
\end{aligned}$$

$$\begin{aligned}
&\leq 1 + \frac{8\varepsilon}{r}\sqrt{\log\frac{1}{rf_r(x)}} + \frac{11|\varepsilon|}{10r} \\
&\leq 1 + \frac{10|\varepsilon|}{r}\sqrt{\log\frac{1}{rf_r(x)}} \quad \text{since } \frac{11|\varepsilon|}{10} \leq \frac{2|\varepsilon|}{r}\sqrt{\log\sqrt{2\pi}} \leq \frac{2|\varepsilon|}{r}\sqrt{\log\frac{1}{rf_r(x)}}
\end{aligned}$$

giving the result. ■

The next lemma shows that $f_r'(x+\varepsilon)$ is close to $f_r'(x)$ for small $\varepsilon$.

**Lemma 16 (Close-by density derivatives are close)** *Let $\varepsilon, x$ be such that $\frac{20|\varepsilon|}{r}\sqrt{\log\frac{1}{f_r(x)r}} < 1$. We have*

$$|f_r'(x+\varepsilon) - f_r'(x)| \lesssim \frac{|\varepsilon|}{r^2} f_r(x)\log\left(\frac{1}{rf_r(x)}\right)$$

*This holds even for asymmetric $f^*$, $f_r$.*

**Proof** Let $w_r$ be the pdf of $\mathcal{N}(0, r^2)$. We have that

$$w_r''(x) = \frac{-r^2+x^2}{\sqrt{2\pi}r^5}e^{-\frac{x^2}{2r^2}} = \frac{-r^2+x^2}{r^4}w_r(x) = \frac{-1+2\log\left(\frac{1}{\sqrt{2\pi}r\cdot w_r(x)}\right)}{r^2}w_r(x)$$

since $x^2 = 2r^2\log\left(\frac{1}{\sqrt{2\pi}r\cdot w_r(x)}\right)$. So, since $g(z) = z\log\left(\frac{1}{\sqrt{2\pi}r\cdot z}\right)$ is concave on $[0,1]$, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{Y\sim f^*}[w_r''(x-Y)] \\
&= \frac{1}{r^2}\mathop{\mathbb{E}}_{Y\sim f^*}\left[w_r(x-Y)\left(-1+2\log\left(\frac{1}{\sqrt{2\pi}r\cdot w_r(x-Y)}\right)\right)\right] \\
&\leq \frac{1}{r^2}\mathop{\mathbb{E}}_{Y\sim f^*}[w_r(x-Y)]\left(-1+2\log\left(\frac{1}{\sqrt{2\pi}r\,\mathbb{E}_{Y\sim f^*}[w_r(x-Y)]}\right)\right) \quad \text{by Jensen's inequality} \\
&\lesssim \frac{f_r(x)\log\left(\frac{1}{f_r(x)\sqrt{2\pi}r}\right)}{r^2}
\end{aligned}$$

So, by Taylor's theorem, for some $|\zeta| < |\varepsilon|$

$$w_r'(x+\varepsilon-Y) = w_r'(x-Y) + \varepsilon w_r''(x+\zeta-Y)$$

Now, by the above

$$\begin{aligned}
\mathop{\mathbb{E}}_{Y\sim f^*}\left[w_r''(x+\zeta-Y)\right] &\lesssim \frac{f_r(x+\zeta)\log\left(\frac{1}{f_r(x+\zeta)\sqrt{2\pi}r}\right)}{r^2} \\
&\lesssim \frac{f_r(x)\log\left(O\left(\frac{1}{f_r(x)r}\right)\right)}{r^2} \quad \text{Using Lemma 15} \\
&\lesssim \frac{f_r(x)\log\frac{1}{rf_r(x)}}{r^2} \quad \text{since } rf_r(x) \leq \frac{1}{\sqrt{2\pi}}
\end{aligned}$$

So, by the above,

$$\begin{aligned} f_r'(x+\varepsilon) - f_r'(x) &\lesssim \varepsilon \mathop{\mathbb{E}}_{Y\sim f^*}\left[w_r''(x+\zeta-Y)\right] \\ &\lesssim \frac{\varepsilon}{r^2} f_r(x)\log\left(\frac{1}{rf_r(x)}\right) \end{aligned}$$

Similarly, by Taylor's theorem, for some $|\zeta|< |\varepsilon|$,

$$w_r'(x-Y) = w_r'(x+\varepsilon-Y) - \varepsilon w_r''(x+\zeta-Y)$$

so that

$$f_r'(x) - f_r'(x+\varepsilon) \lesssim -\frac{\varepsilon}{r^2} f_r(x)\log\left(\frac{1}{rf_r(x)}\right)$$

The claim follows. ■

Finally, the main result of this section shows that $s_r(x+\varepsilon)$ is close to $s_r(x)$ for small $\varepsilon$.

**Lemma 17 (Close-by scores are close)** *Let $\varepsilon, x$ be such that $\frac{20|\varepsilon|}{r}\sqrt{\log\frac{1}{rf_r(x)}} < 1$. We have*

$$|s_r(x+\varepsilon) - s_r(x)| \lesssim \frac{|\varepsilon|}{r^2}\log\left(\frac{1}{rf_r(x)}\right)$$

*This holds even for asymmetric $f^*, f_r$.*

**Proof** By Lemma 15 and 16, we have

$$\frac{f_r'(x+\varepsilon)}{f_r(x+\varepsilon)} \le \frac{f_r'(x) + O\left(\frac{|\varepsilon|}{r^2} f_r(x)\log\left(\frac{1}{rf_r(x)}\right)\right)}{f_r(x)\left(1 - \frac{10|\varepsilon|}{r}\sqrt{\log\frac{1}{rf_r(x)}}\right)}$$

Since $\frac{10|\varepsilon|}{r}\sqrt{\log\frac{1}{r\alpha}} < 1/2$, we have

$$\frac{f_r'(x+\varepsilon)}{f_r(x+\varepsilon)} \le \frac{f_r'(x)}{f_r(x)} + O\left(\frac{|\varepsilon|}{r^2}\log\left(\frac{1}{rf_r(x)}\right)\right) + O\left(\frac{f_r'(x)}{f_r(x)}\frac{|\varepsilon|}{r}\sqrt{\log\frac{1}{rf_r(x)}}\right)$$

So,

$$\begin{aligned} s_r(x+\varepsilon) - s_r(x) &\lesssim +\frac{|\varepsilon|}{r^2}\log\left(\frac{1}{rf_r(x)}\right) + s_r(x)\frac{|\varepsilon|}{r}\sqrt{\log\frac{1}{rf_r(x)}} \\ &\lesssim \frac{|\varepsilon|}{r^2}\log\left(\frac{1}{rf_r(x)}\right) \quad \text{by Lemma 33} \end{aligned}$$

We can get the lower bound in the same way. ■

### B.3. Bounding the clipped KDE error

In this section, we show that the clipped KDE score function $\widehat{s}_r^{\text{clip}}$ approximates the true score function $\widehat{s}_r^{\text{sym}}$ in a specific sense. We begin by showing that sets that have small density under $f_r$ have small expected score.

**Lemma 18** *Let $S$ be any set with $\Pr_{f_r}[S] \leq \beta$. Let $|\varepsilon| \leq r/2$ and $|\varepsilon| \leq \frac{r}{\sqrt{\log \frac{1}{r^2 \mathcal{I}_r}}}$. Then, for the score function $s_r$ of $f_r$, we have*

$$\mathop{\mathbb{E}}_{x \sim f_r}[s_r(x+\varepsilon)^2 1_{x \in S}] \lesssim \frac{\beta}{r^2} \log \frac{1}{\beta}$$

*This holds even for asymmetric $f^*$, $f_r$.*

**Proof**

$$\begin{aligned}
\mathop{\mathbb{E}}_{x \sim f_r}[s_r^2(x+\varepsilon)\mathbb{1}_{x \in S}] &= \int_0^\infty \Pr_{x \sim f_r}[s_r^2(x+\varepsilon)\mathbb{1}_{x \in S} \geq t] dt \\
&\leq \int_0^\infty \min(\Pr_{x \sim f_r}[x \in S], \Pr_{x \sim f_r}[s_r^2(x+\varepsilon) \geq t]) dt \\
&\leq \int_0^\infty \min(\beta, \Pr_{x \sim f_r}[|s_r(x+\varepsilon)| \geq \sqrt{t}]) dt
\end{aligned}$$

So, by Lemma 35, for some explicit constant $C > 0$, we have:

$$\begin{aligned}
\mathop{\mathbb{E}}_{x \sim f_r}[s_r^2(x+\varepsilon)\mathbb{1}_{x \in S}] &\lesssim \int_0^\infty \min(\beta, e^{-Ctr^2}) dt \\
&\lesssim \beta B + \int_B^\infty e^{-Ctr^2} dt \\
&\lesssim \beta B + \frac{e^{-CBr^2}}{Cr^2}
\end{aligned}$$

Thus, setting $B = \frac{\log \frac{1}{\beta}}{Cr^2}$ gives

$$\mathbb{E}[s_r^2(x+\varepsilon) 1_{x \in S}] \lesssim \frac{\beta}{r^2} \log \frac{1}{\beta}$$

■

The next lemma shows that for every small width region, the KDE score $\widehat{s}$ approximates the true score well, as long as the density of that region is large.

**Lemma 19 (Generic Error estimate within bin)** *Let $f^*$ be an arbitrary distribution and let $f_r$ be the $r$-smoothed version of $f^*$. Let $\widehat{f_r}$ be the kernel density estimate of $f_r$ from $N$ samples $Y_1, \dots, Y_N \sim f^*$. Let $x, \varepsilon, N$ be such that $N \geq \frac{6 \log \frac{4}{\delta}}{r f_r(x)}$, and $\frac{20|\varepsilon|}{r}\sqrt{\log \frac{1}{r f_r(x)}} < 1$. Then, for the KDE score $\widehat{s}_r$ defined in (8), with probability $1-\delta$, we have that for all $|\zeta| \leq |\varepsilon|$ (simultaneously),*

$$|\widehat{s}_r(x+\zeta) - s_r(x+\zeta)| \lesssim \sqrt{\frac{\log \frac{1}{\delta} \log \frac{1}{r f_r(x)}}{N r^3 f_r(x)}} + \frac{|\varepsilon|}{r^2} \log\left(\frac{1}{r f_r(x)}\right)$$

**Proof** First, by Lemma 13, we have that with probability $1-\delta$,

$$|\widehat{s}_r(x) - s_r(x)| \lesssim \sqrt{\frac{\log\frac{1}{\delta}\log\frac{1}{rf_r(x)}}{Nr^3 f_r(x)}}$$

Now, by Lemma 17, since $|\zeta| \leq |\varepsilon|$,

$$|s_r(x+\zeta) - s_r(x)| \lesssim \frac{|\varepsilon|}{r^2}\log\left(\frac{1}{rf_r(x)}\right)$$

and

$$\begin{aligned}
|\widehat{s}_r(x+\zeta) - \widehat{s}_r(x)| &\lesssim \frac{|\varepsilon|}{r^2}\log\left(\frac{1}{r\widehat{f}_r(x)}\right) \\
&= \frac{|\varepsilon|}{r^2}\left(\log\left(\frac{1}{rf_r(x)}\right) + \log\left(\frac{f_r(x)}{\widehat{f}_r(x)}\right)\right) \\
&\leq \frac{|\varepsilon|}{r^2}\left(\log\left(\frac{1}{rf_r(x)}\right) + \log\left(1+\frac{1}{\sqrt{2}}\right)\right) \quad \text{by Lemma 11 and since } N \geq \frac{6\log\frac{4}{\delta}}{rf_r(x)} \\
&\lesssim \frac{|\varepsilon|}{r^2}\log\left(\frac{1}{rf_r(x)}\right) \quad \text{since } rf_r(x) \leq \frac{1}{\sqrt{2\pi}}, \text{ so } \log\frac{1}{rf_r(x)} = \Omega(1)
\end{aligned}$$

Putting everything together, with probability $1-\delta$, for all $|\zeta| \leq |\varepsilon|$,

$$\begin{aligned}
|\widehat{s}_r(x+\zeta) - s_r(x+\zeta)| &\leq |\widehat{s}_r(x+\zeta) - \widehat{s}_r(x)| + |\widehat{s}_r(x) - s_r(x)| + |s_r(x) - s_r(x+\zeta)| \\
&\lesssim \sqrt{\frac{\log\frac{1}{\delta}\log\frac{1}{rf_r(x)}}{Nr^3 f_r(x)}} + \frac{|\varepsilon|}{r^2}\log\left(\frac{1}{rf_r(x)}\right)
\end{aligned}$$

■

The next lemma shows that for all points with density larger than $\alpha$ within a $t\sigma$ radius around the true mean, the KDE score approximates the true score well.

**Lemma 20 (Generic Error estimate over large density region)** *Let $f^*$ be an arbitrary distribution with mean $\mu$ and variance $\sigma^2$, and let $f_r$ be the $r$-smoothed version of $f^*$, with variance $\sigma_r^2 = \sigma^2 + r^2$. Let $\widehat{s}_r$ be the score of the kernel density estimate of $f_r$ from $N$ samples $Y_1, \dots, Y_N \sim f^*$, as defined in (8). Let $\alpha > 0$ and let $N \geq \frac{6\log\left(\frac{4}{\delta}\left(\frac{2\sqrt{\alpha N}t\sigma_r}{\sqrt{r}}+1\right)\right)+400\log\frac{1}{\alpha r}}{\alpha r}$. Then with probability $1-\delta$, we have that,*

$$\mathop{\mathbb{E}}_{x\sim f_r}\left[(\widehat{s}_r(x) - s_r(x))^2 \mathbb{1}_{\{|x-\mu| \leq t\sigma_r \text{ and } f_r(x) \geq \alpha\}}\right] \lesssim \frac{\log\left(\frac{1}{\delta}\left(\frac{2t\sigma_r\sqrt{\alpha N}}{\sqrt{r}}+1\right)\right)}{\alpha N r^3}\log^2\frac{1}{\alpha r}$$

**Proof** Consider contiguous intervals of length $\varepsilon$ starting from $\mu - t\sigma_r$ so that the last interval covers $\mu + t\sigma_r$, and let $S$ be the set of the smallest $y$ such that $f(y) \geq \alpha$ in each of these intervals, if one exists. Note that $|S| \leq \frac{2t\sigma_r}{\varepsilon} + 1$. Then, we have that

$$\{x : |x-\mu| \leq t\sigma_r \text{ and } f_r(x) \geq \alpha\} \subseteq \{[y-\varepsilon, y+\varepsilon] | y \in S\}$$

Now, for $\varepsilon = \sqrt{\frac{r}{\alpha N}}$ and $y \in S$, since $N \geq \frac{6\log\left(\frac{4}{\delta}\left(\frac{2\sqrt{\alpha N}t\sigma_r}{\sqrt{r}}+1\right)\right)}{\alpha r} \geq \frac{6\log\frac{4|S|}{\delta}}{rf_r(y)}$ and $\frac{20|\varepsilon|}{r}\sqrt{\log\frac{1}{rf_r(y)}} \leq \frac{20}{\sqrt{\alpha r N}}\sqrt{\log\frac{1}{\alpha r}} \leq 1$, by Lemma 19, we have that with probability $1 - \frac{\delta}{|S|}$, for all $|\zeta| \leq \varepsilon$ (simultaneously),

$$|\widehat{s}_r(y+\zeta) - s_r(y+\zeta)| \lesssim \sqrt{\frac{\log\frac{|S|}{\delta}\log\frac{1}{rf_r(y)}}{Nr^3 f_r(y)}} + \sqrt{\frac{1}{\alpha N r^3}}\log\frac{1}{rf_r(y)}$$

$$\lesssim \sqrt{\frac{\log\left(\frac{1}{\delta}\left(\frac{2t\sigma_r\sqrt{\alpha N}}{\sqrt{r}}+1\right)\right)}{\alpha N r^3}}\log\frac{1}{\alpha r}$$

So by a union bound, with probability $1-\delta$, for all $x$ such that $|x-\mu| \leq t\sigma_r$ and $f(x) \geq \alpha$ simultaneously,

$$|\widehat{s}_r(x) - s_r(x)| \lesssim \sqrt{\frac{\log\left(\frac{1}{\delta}\left(\frac{2t\sigma_r\sqrt{\alpha N}}{\sqrt{r}}+1\right)\right)}{\alpha N r^3}}\log\frac{1}{\alpha r}$$

So,

$$\mathop{\mathbb{E}}_{x\sim f_r}\left[(\widehat{s}_r(x) - s_r(x))^2 \mathbb{1}_{|x-\mu|\leq t\sigma_r \text{ and } f_r(x)\geq\alpha}\right] \lesssim \frac{\log\left(\frac{1}{\delta}\left(\frac{2t\sigma_r\sqrt{\alpha N}}{\sqrt{r}}+1\right)\right)}{\alpha N r^3}\log^2\frac{1}{\alpha r}$$

■

The next lemma instantiates the previous one with a particular value of $t$ and $\alpha$ based on our desired failure probability and the number of samples.

**Lemma 21 (Error estimate over large density region (instantiated))** *Let $f^*$ be an arbitrary distribution with mean $\mu$ and variance $\sigma^2$, and let $f_r$ be the $r$-smoothed version of $f^*$, with variance $\sigma_r^2 = \sigma^2 + r^2$ and Fisher information $\mathcal{I}_r$. Let $\widehat{s}_r$ be the score of the kernel density estimate of $f_r$ from $N$ samples $Y_1, \dots, Y_N \sim f^*$, as defined in (9). Let $\gamma \geq C$ for large enough constant $C > 1$ be a parameter. Let $t = \sqrt{\frac{\gamma \log\frac{N}{\mathcal{I}_r r^2 \log\frac{1}{\delta}}}{\mathcal{I}_r r^2}}$, $\alpha = \frac{1}{t^3\sigma_r}$. Then for any $r \leq \sigma$ and $\frac{N}{\log\frac{1}{\delta}} \geq \left(\gamma^{5/12}\frac{\sigma}{r}\right)^{6+\beta}$ for any constant $\beta > 0$, with probability $1-\delta$, we have that,*

$$\mathop{\mathbb{E}}_{x\sim f_r}\left[(\widehat{s}_r(x) - s_r(x))^2 \mathbb{1}_{\{|x-\mu|\leq t\sigma_r \textit{ and } f_r(x)\geq\alpha\}}\right] \lesssim \frac{\mathcal{I}_r}{\gamma}$$

**Proof** First, note that since $r \leq \sigma$,

$$\sigma_r^2 = \sigma^2 + r^2 \leq 2\sigma^2$$

Also, our setting of $N$ implies that WLOG

$$\frac{N}{\log\frac{1}{\delta}} \geq \left(\frac{\gamma^{5/12}\sigma\log\frac{N}{\log\frac{1}{\delta}}}{r}\right)^6$$

since $N \geq C \log \frac{1}{\delta}$. So,

$$\frac{\sigma}{r} \leq \left(\frac{N}{\log \frac{1}{\delta}}\right)^{1/6} \cdot \frac{1}{\gamma^{5/12} \log \frac{N}{\log \frac{1}{\delta}}} \leq \left(\frac{N}{\log \frac{1}{\delta}}\right)^{1/6} \tag{11}$$

We will first check that this $N$ satisfies the condition required to invoke Lemma 20 that $N \geq \frac{6\log\left(\frac{4}{\delta}\left(\frac{2\sqrt{\alpha N}t\sigma_r}{\sqrt{r}}+1\right)\right)+400\log\frac{1}{\alpha r}}{\alpha r}$. To do this, we will individually upper bound $\frac{1}{\alpha r}$ and $\frac{2t\sigma_r\sqrt{\alpha N}}{\sqrt{r}}$.

We have,

$$\begin{aligned}
\frac{1}{\alpha r} &= \frac{\sigma_r}{r}\left(\frac{\gamma \log \frac{N}{\mathcal{I}_r r^2 \log \frac{1}{\delta}}}{\mathcal{I}_r r^2}\right)^{3/2} \quad \text{since } \alpha = \frac{1}{t^3 \sigma_r} \text{ and } t = \sqrt{\frac{\gamma \log \frac{N}{\mathcal{I}_r r^2 \log \frac{1}{\delta}}}{\mathcal{I}_r r^2}} \\
&\leq \left(\frac{\sigma_r}{r}\right)^4 \gamma^{3/2} \log^{3/2} \frac{N\sigma_r^2}{r^2 \log \frac{1}{\delta}} \quad \text{since } \mathcal{I}_r \geq \frac{1}{\sigma_r^2} \\
&\leq \left(\frac{2\sigma}{r}\right)^4 \gamma^{3/2} \log^{3/2}\left(\frac{2N\sigma^2}{r^2 \log \frac{1}{\delta}}\right) \quad \text{since } \sigma_r^2 \leq 2\sigma^2 \\
&\leq 16 \frac{N^{4/6}}{\gamma^{5/3} \log^4\left(\frac{N}{\log \frac{1}{\delta}}\right) \log^{4/6} \frac{1}{\delta}} \gamma^{3/2} \log^{3/2}\left(2\left(\frac{N}{\log \frac{1}{\delta}}\right)^{4/3}\right) \quad \text{by (11)} \\
&\leq \frac{N}{\gamma^{1/10} \log(\frac{1}{\delta}) \log^2\left(\frac{N}{\log \frac{1}{\delta}}\right)} \quad \text{since } \gamma \geq C \text{ for } C \text{ large enough constant, and } \frac{\log \frac{1}{\delta}}{N} \leq 1
\end{aligned}$$

To further justify the last line above, observe that $\frac{\gamma^{3/2}}{\gamma^{5/3}} \leq \frac{1}{\gamma^{1/6}} = \frac{1}{\gamma^{1/10} \cdot \gamma^{1/15}}$, and that for large enough constant $C$, since $\gamma > C$, $\gamma^{1/15}$ can be made larger than any fixed constant. Also note that $\log^{3/2}(2(\frac{N}{\log \frac{1}{\delta}})^{4/3}) \leq \log^2 \frac{N}{\log \frac{1}{\delta}}$ for large enough $C$ since $\frac{N}{\log \frac{1}{\delta}} \geq \gamma \geq C$. So, the inequality follows.

Next, we bound $\frac{2t\sigma_r\sqrt{\alpha N}}{\sqrt{r}}$.

$$\begin{aligned}
\frac{2t\sigma_r\sqrt{\alpha N}}{\sqrt{r}} &= 2\sqrt{\frac{N\sigma_r}{tr}} \quad \text{since } \alpha = \frac{1}{t^3\sigma_r} \\
&= 2\sqrt{\frac{N\sigma_r}{\sqrt{\frac{\gamma\log\frac{N}{\mathcal{I}_r r^2\log\frac{1}{\delta}}}{\mathcal{I}_r r^2}}\,r}} \quad \text{since } t = \sqrt{\frac{\gamma\log\frac{N}{\mathcal{I}_r r^2\log\frac{1}{\delta}}}{\mathcal{I}_r r^2}} \\
&\le 4\sqrt{\frac{N\sigma}{r\sqrt{\gamma\log\frac{N}{\log\frac{1}{\delta}}}}} \quad \text{since } \mathcal{I}_r \le \frac{1}{r^2} \text{ and } \sigma_r^2 \le 2\sigma^2 \\
&\le 4\sqrt{\frac{N}{\sqrt{\gamma\log\frac{N}{\log\frac{1}{\delta}}}}\cdot\left(\frac{N}{\log\frac{1}{\delta}}\right)^{1/6}} \quad \text{by (11)} \\
&\le 4N\cdot\left(\frac{N}{\log\frac{1}{\delta}}\right)^{1/12} \quad \text{since } \gamma \ge 1, \frac{N}{\log\frac{1}{\delta}} \ge 1
\end{aligned}$$

So, we can now check the condition required to invoke Lemma 20. We have,

$$\begin{aligned}
&\frac{6\log\left(\frac{4}{\delta}\left(\frac{2\sqrt{\alpha N}t\sigma_r}{\sqrt{r}}+1\right)\right)+400\log\frac{1}{\alpha r}}{\alpha r} \\
&\le \left(6\log\left(\frac{4}{\delta}\left(4N\cdot\left(\frac{N}{\log\frac{1}{\delta}}\right)^{1/12}+1\right)\right)+400\log\frac{N}{\log\frac{1}{\delta}}\right)\left(\frac{N}{\gamma^{1/10}\log^2\left(\frac{N}{\log\frac{1}{\delta}}\right)\log\frac{1}{\delta}}\right) \\
&\le N \quad \text{since } \gamma \ge C
\end{aligned}$$

So, by Lemma 20, we have

$$\mathop{\mathbb{E}}_{x\sim f_r}\left[(\widehat{s}_r(x)-s_r(x))^2\mathbb{1}_{\{|x-\mu|\le t\sigma_r \text{ and } f_r(x)\ge\alpha\}}\right] \lesssim \frac{\log\left(\frac{1}{\delta}\left(\frac{2\sigma_r\sqrt{\alpha N}}{\sqrt{r}}+1\right)\right)}{\alpha N r^3}\log^2\frac{1}{\alpha r}$$

To bound the RHS above by $O\left(\mathcal{I}_r/\gamma\right)$ as required, we will first bound $\frac{1}{\alpha N r^3}$.

$$
\begin{aligned}
\frac{1}{\alpha N r^3} &= \frac{\sigma_r}{N r^3}\left(\frac{\gamma \log \frac{N}{\mathcal{I}_r r^2 \log \frac{1}{\delta}}}{\mathcal{I}_r r^2}\right)^{3/2} \qquad \text{since } \alpha = \frac{1}{t^3\sigma_r} \text{ and } t = \sqrt{\frac{\gamma \log \frac{N}{\mathcal{I}_r r^2 \log \frac{1}{\delta}}}{\mathcal{I}_r r^2}} \\
&= \frac{\mathcal{I}_r}{\gamma}\left(\frac{\sigma_r}{N r^6}\frac{\gamma^{5/2}}{\mathcal{I}_r^{5/2}}\log^{3/2}\frac{N}{\mathcal{I}_r r^2 \log \frac{1}{\delta}}\right) \\
&\le \frac{\mathcal{I}_r}{\gamma}\left(\frac{\sigma_r^6\gamma^{5/2}\log^{3/2}\frac{N\sigma_r^2}{r^2\log\frac{1}{\delta}}}{N r^6}\right) \qquad \text{since } \mathcal{I}_r \ge \frac{1}{\sigma_r^2} \\
&\lesssim \frac{\mathcal{I}_r}{\gamma}\left(\frac{\sigma^6\gamma^{5/2}\log^{3/2}\frac{4N\sigma^2}{r^2\log\frac{1}{\delta}}}{N r^6}\right) \qquad \text{since } \sigma_r^2 \le 2\sigma^2 \\
&\le \frac{\mathcal{I}_r}{\gamma}\left(\frac{\log^{3/2}(4\left(\frac{N}{\log\frac{1}{\delta}}\right)^{4/3})}{\log^6\left(\frac{N}{\log\frac{1}{\delta}}\right)\log\frac{1}{\delta}}\right) \qquad \text{by (11)} \\
&\lesssim \frac{\mathcal{I}_r}{\gamma\log^4\left(\frac{N}{\log\frac{1}{\delta}}\right)\log\frac{1}{\delta}}
\end{aligned}
$$

So, plugging in everything,

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{x\sim f_r}\left[(\widehat{s}_r(x)-s_r(x))^2\mathbb{1}_{\{|x-\mu|\le t\sigma_r \text{ and } f_r(x)\ge\alpha\}}\right] \\
&\lesssim \frac{\log\left(\frac{1}{\delta}\left(\frac{2\sigma_r\sqrt{\alpha N}}{\sqrt{r}}+1\right)\right)}{\alpha N r^3}\log^2\frac{1}{\alpha r} \\
&\lesssim \frac{\mathcal{I}_r}{\gamma\log^4\left(\frac{N}{\log\frac{1}{\delta}}\right)\log\frac{1}{\delta}}\log\left(\frac{4N\cdot\left(\frac{N}{\log\frac{1}{\delta}}\right)^{1/12}+1}{\delta}\right)\log^2\left(\left(\frac{N}{\log\frac{1}{\delta}}\right)^{4/6}\right) \\
&\lesssim \frac{\mathcal{I}_r}{\gamma}
\end{aligned}
$$

■

The lemmas so far have shown that the KDE score $\widehat{s}_r$ approximates the true score $s_r$ well in large denstiy regions in the typical $t\sigma$ radius around the true mean. The next lemma shows that the same guarantee holds for the clipped KDE score $\widehat{s}_r^{\text{clip}}$.

**Lemma 22 (Error estimate over large density region for clipped KDE)** *Let $f^*$ be an arbitrary distribution with mean $\mu$ and variance $\sigma^2$, and let $f_r$ be the $r$-smoothed version of $f^*$, with variance $\sigma_r^2 = \sigma^2 + r^2$. Let $\widehat{s}_r^{clip}$ be the clipped kernel density estimate score from $N$ samples, as defined in* (8). *Let $\gamma \ge C$ for large enough constant $C > 1$ be a parameter. Let $t = \sqrt{\frac{\gamma\log\frac{N}{\mathcal{I}_r r^2\log\frac{1}{\delta}}}{\mathcal{I}_r r^2}}$,*

$\alpha = \frac{1}{t^3\sigma_r}$.*Then for any* $r \le \sigma$ *and* $\frac{N}{\log\frac{1}{\delta}} \ge \left(\frac{\gamma^{5/12}\sigma}{r}\right)^{6+\beta}$ *for* $\beta > 0$*, with probability* $1-\delta$*, we have that,*

$$\mathop{\mathbb{E}}_{x\sim f_r}\left[(\widehat{s}_r^{\mathit{clip}}(x) - s_r(x))^2 \mathbb{1}_{\{|x-\mu|\le t\sigma_r \text{ and } f_r(x)\ge\alpha\}}\right] \lesssim \frac{\mathcal{I}_r}{\gamma}$$

**Proof** Note that our condition that $\frac{N}{\log\frac{1}{\delta}} \ge \left(\frac{\gamma^{5/12}\sigma}{r}\right)^{6+\beta}$ implies

$$\frac{\sigma}{r} \le \left(\frac{N}{\log\frac{1}{\delta}}\right)^{1/6}$$

Also, since $r \le \sigma$, $\frac{N}{\log\frac{1}{\delta}} \ge (\gamma^{5/12})^{6+\beta} \ge \gamma^{5/2}$.

So, by Lemma 33, for $x$ such that $f_r(x) \ge \alpha$,

$$\begin{aligned}
|s_r(x)| &\le \frac{1}{r}\sqrt{2\log\frac{1}{\sqrt{2\pi}r\alpha}} \\
&= \frac{1}{r}\sqrt{2\log\left(\frac{\sigma_r}{\sqrt{2\pi}r}\left(\frac{\gamma\log\frac{N}{\mathcal{I}_r r^2\log\frac{1}{\delta}}}{\mathcal{I}_r r^2}\right)^{3/2}\right)} \quad \text{since } \alpha = \frac{1}{t^3\sigma_r} \text{ and } t = \sqrt{\frac{\gamma\log\frac{N}{\mathcal{I}_r r^2\log\frac{1}{\delta}}}{\mathcal{I}_r r^2}} \\
&\le \frac{1}{r}\sqrt{2\log\left(\frac{\sigma_r^4}{\sqrt{2\pi}r^4}\left(\gamma\log\frac{\sigma_r^2 N}{r^2\log\frac{1}{\delta}}\right)^{3/2}\right)} \quad \text{since } \mathcal{I}_r \ge \frac{1}{\sigma_r^2} \\
&\le \frac{1}{r}\sqrt{2\log\left(\frac{16}{\sqrt{2\pi}}\left(\frac{N}{\log\frac{1}{\delta}}\right)^{4/6}\left(\gamma\log\left(4\left(\frac{N}{\log\frac{1}{\delta}}\right)^{4/3}\right)\right)^{3/2}\right)} \\
&\qquad \text{since } \sigma_r^2 \le 2\sigma^2 \text{ and using } \sigma/r \le \left(\frac{N}{\log\frac{1}{\delta}}\right)^{1/6} \\
&\le \frac{2}{r}\sqrt{\log\frac{N}{\log\frac{1}{\delta}}} \quad \text{since } \frac{N}{\log\frac{1}{\delta}} \ge \gamma^{5/2} \ge C^{5/2} \text{ for a sufficiently large constant } C
\end{aligned}$$

Then, since $\widehat{s}_r^{\text{clip}}$ is clipped at $\frac{2}{r}\sqrt{\log\frac{N}{\log\frac{1}{\delta}}}$ by definition in (9), by Lemma 21, we have the claim. ■

The next lemma shows that for small density sets, for any function that is clipped appropriately, the error incurred relative to the true score function is small.

**Lemma 23** *Let* $f^*$ *be an arbitrary distribution, and let* $f_r$ *be the* $r$*-smoothed version of* $f^*$*. Let* $s_r$ *be the score function of* $f_r$*, and let* $\mathcal{I}_r$ *be the Fisher information. Let* $\gamma \ge C$ *for large enough constant* $C$*,* $\frac{N}{\log\frac{1}{\delta}} \ge \gamma$*, and let* $\tilde{s}$ *be any function with* $|\tilde{s}_r(x)| \le \frac{2}{r}\sqrt{\log\frac{N}{\log\frac{1}{\delta}}}$ *for all* $x$*. Let* $S$ *be a set* *with* $\Pr[S] \le \frac{1}{t^2}$ *for* $t = \sqrt{\frac{\gamma \log \frac{N}{\mathcal{I}_r r^2 \log \frac{1}{\delta}}}{\mathcal{I}_r r^2}}$. *Then, we have*

$$\mathop{\mathbb{E}}_{x\sim f_r}[(\tilde{s}_r(x) - s_r(x))^2 \mathbb{1}_{x\in S}] \lesssim \frac{\mathcal{I}_r}{\gamma}$$

**Proof** First, by Lemma 18,

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim f_r}\left[s_r(x)^2 \mathbb{1}_{x\in S}\right] &\lesssim \frac{1}{t^2 r^2}\log t^2 \\
&\le \frac{\mathcal{I}_r}{\gamma \log \frac{N}{\mathcal{I}_r r^2 \log \frac{1}{\delta}}} \log\left(\frac{\gamma}{\mathcal{I}_r r^2} \log \frac{N}{\mathcal{I}_r r^2 \log \frac{1}{\delta}}\right) \\
&\lesssim \frac{\mathcal{I}_r}{\gamma} \quad \text{since } \gamma \ge 1,\, \gamma \le \frac{N}{\log \frac{1}{\delta}} \text{ and } \mathcal{I}_r \le \frac{1}{r^2}
\end{aligned}$$

Now, by assumption,

$$|\tilde{s}_r(x)| \le \frac{2}{r}\sqrt{\log \frac{N}{\log \frac{1}{\delta}}}$$

So,

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim f_r}\left[\tilde{s}_r(x)^2 \mathbb{1}_{x\in S}\right] &\le \frac{4}{t^2 r^2}\log \frac{N}{\log \frac{1}{\delta}} \\
&= \frac{\mathcal{I}_r}{\gamma \log \frac{N}{\mathcal{I}_r r^2 \log \frac{1}{\delta}}} \log\left(\frac{N}{\log \frac{1}{\delta}}\right) \\
&\lesssim \frac{\mathcal{I}_r}{\gamma \log \frac{N}{\mathcal{I}_r r^2 \log \frac{1}{\delta}}} \log\left(\frac{N}{\mathcal{I}_r r^2 \log \frac{1}{\delta}}\right) \quad \text{since } \mathcal{I}_r \le \frac{1}{r^2} \\
&\lesssim \frac{\mathcal{I}_r}{\gamma}
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim f_r}\left[(\tilde{s}_r(x) - s_r(x))^2 \mathbb{1}_{x\in S}\right] &\lesssim \mathop{\mathbb{E}}_{x\sim f_r}\left[\tilde{s}_r(x)^2 \mathbb{1}_{x\in S}\right] + \mathop{\mathbb{E}}_{x\sim f_r}\left[s_r(x)^2 \mathbb{1}_{x\in S}\right] \\
&\lesssim \frac{\mathcal{I}_r}{\gamma}
\end{aligned}$$

■

The next lemma shows that within the typical $t\sigma$ radius around the mean, for the set of points with small density, the clipped KDE score $\widehat{s}_r^{\text{clip}}$ approximates the true score $s_r$ well.

**Lemma 24 (Error estimate over small density regions within typical region)** *Let $f^*$ be an arbitrary distribution with mean $\mu$ and variance $\sigma^2$, and let $f_r$ be the $r$-smoothed version of $f^*$, with variance $\sigma_r^2 = \sigma^2 + r^2$. Let $s_r$ be the score function of $f_r$, and let $\mathcal{I}_r$ be the Fisher information. Let*

*$\tilde{s}_r$ be any score function with $|\tilde{s}_r(x)|\leq \frac{2}{r}\sqrt{\log \frac{N}{\log \frac{1}{\delta}}}$ for all $x$. Let $t = \sqrt{\frac{\gamma \log \frac{N}{\mathcal{I}_r r^2 \log \frac{1}{\delta}}}{\mathcal{I}_r r^2}}$ for $\gamma \geq 1$, and let $\alpha = \frac{1}{t^3 \sigma_r}$. Then,*

$$\mathop{\mathbb{E}}_{x\sim f_r}[(\tilde{s}_r(x) - s_r(x))^2 \mathbb{1}_{\{|x-\mu|\leq t\sigma_r \textit{ and } f(x)<\alpha\}}] \lesssim \frac{\mathcal{I}_r}{\gamma}$$

**Proof** By our choice of $\alpha$,

$$\Pr_{x\sim f_r}[|x-\mu|\leq t\sigma_r \text{ and } f(x) < \alpha] \leq \alpha t \sigma_r = \frac{1}{t^2}$$

So, by Lemma 23, the claim follows ■

The next lemma shows that in the region outside the typical region of radius $t\sigma$ around the true mean, the clipped KDE score $\widehat{s}_r^{\text{clip}}$ approximates the true score well.

**Lemma 25 (Error estimate over atypical region)** *Let $f^*$ be an arbitrary distribution with mean $\mu$ and variance $\sigma^2$, and let $f_r$ be the $r$-smoothed version of $f^*$, with variance $\sigma_r^2 = \sigma^2 + r^2$. Let $s_r$ be the score function of $f_r$, and let $\mathcal{I}_r$ be the Fisher information. Let $\tilde{s}_r$ be any score function with $|\tilde{s}_r(x)|\leq \frac{2}{r}\sqrt{\log \frac{N}{\log \frac{1}{\delta}}}$ for all $x$. Let $t = \sqrt{\frac{\gamma \log \frac{N}{\mathcal{I}_r r^2 \log \frac{1}{\delta}}}{\mathcal{I}_r r^2}}$ for $\gamma \geq 1$, and let $\alpha = \frac{1}{t^3\sigma}$. Then,*

$$\mathop{\mathbb{E}}_{x\sim f_r}[(\tilde{s}_r(x) - s_r(x))^2 \mathbb{1}_{|x-\mu|>t\sigma_r}] \lesssim \frac{\mathcal{I}_r}{\gamma}$$

**Proof** By Chebyshev's inequality,

$$\Pr[|x-\mu|> t\sigma_r] \leq \frac{1}{t^2}$$

So, by Lemma 23, the claim follows. ■

The main result of this section as follows shows that the clipped KDE score approximates the true score well.

**Lemma 3 (Clipped KDE score error)** *Let $\widehat{s}_r^{clip}$ be the clipped Kernel Density estimate from $N$ samples, defined in (9). Let $\gamma > C$ be a parameter, for large enough constant $C \geq 1$. Then for any $r \leq \sigma$ and $\frac{N}{\log \frac{1}{\delta}} \geq \left(\frac{\gamma^{5/12}\sigma}{r}\right)^{6+\beta}$ for $\beta > 0$, with probability $1-\delta$, we have that,*

$$\mathop{\mathbb{E}}_{x\sim f_r}\left[(\widehat{s}_r^{clip}(x) - s_r(x))^2\right] \lesssim \frac{\mathcal{I}_r}{\gamma}$$

*This holds even for asymmetric $f^*$ and $f_r$.*

**Proof** Let $\mu, \sigma_r^2$ be the mean and variance of $f_r$. Let $t = \sqrt{\frac{\gamma \log \frac{N}{\mathcal{I}_r r^2 \log \frac{1}{\delta}}}{\mathcal{I}_r r^2}}$, $\alpha = \frac{1}{t^3\sigma_r}$. Now,

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim f_r}\left[(\widehat{s}_r^{\text{clip}}(x) - s_r(x))^2\right] &= \mathop{\mathbb{E}}_{x\sim f_r}\left[(\widehat{s}_r^{\text{clip}}(x) - s_r(x))^2 \mathbb{1}_{\{|x-\mu|\leq t\sigma_r \text{ and } f_r(x)\geq\alpha\}}\right] \\ &\quad + \mathop{\mathbb{E}}_{x\sim f_r}\left[(\widehat{s}_r^{\text{clip}}(x) - s_r(x))^2 \mathbb{1}_{\{|x-\mu|\leq t\sigma_r \text{ and } f_r(x)<\alpha\}}\right] \\ &\quad + \mathop{\mathbb{E}}_{x\sim f_r}[(\widehat{s}_r^{\text{clip}}(x) - s_r(x))^2 \mathbb{1}_{|x-\mu|>t\sigma_r}]\end{aligned}$$

So, by Lemmas 22, 24 and 25, we have the claim. ■

## Appendix C. Symmetrization

In this section we show that the expectation of our symmetrized, clipped KDE score function $\widehat{s}_r^{\text{sym}}$, symmetrized around a point $\mu + \varepsilon$ for small $\varepsilon$ has expectation close to $\varepsilon \mathcal{I}_r$, where $\mathcal{I}_r$ is the Fisher information of the $r$-smoothed distribution. We also show that the second moment of $\widehat{s}_r^{\text{sym}}$ is close to $\mathcal{I}_r$. We begin by recalling the definition of $\widehat{s}_r^{\text{sym}}$ from (7).

**Definition** Let the symmetrized, clipped KDE score, symmetrized around a point $y$ from $N$ samples be given by

$$\widehat{s}_r^{\text{sym}}(x) = \begin{cases} \widehat{s}_r^{\text{clip}}(x) & x \geq y \\ -\widehat{s}_r^{\text{clip}}(2y - x) & x < y \end{cases} \tag{12}$$

where $\widehat{s}_r^{\text{clip}}$ is the clipped KDE score from $N$ samples, as defined in (9).

First we show that the true score function centered at $-\varepsilon$ is close to the true score centered at 0 in $\ell_2$ distance.

**Lemma 26** *Let $f_r$ be an $r$-smoothed distribution with score function $s_r$ and Fisher information $\mathcal{I}_r$. Then, for $|\varepsilon| \leq r/60$,*

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

**Proof** By Lemma 36,

$$\begin{aligned} s_r(x+\varepsilon) - s_r(x) &= \frac{\mathbb{E}_{Z_r|x}\left[e^{\frac{\varepsilon Z_r}{r^2}} \frac{Z_r - \varepsilon}{r^2}\right]}{\mathbb{E}_{Z_r|x}\left[e^{\frac{\varepsilon Z_r}{r^2}}\right]} - \mathop{\mathbb{E}}_{Z_r|x}\left[\frac{Z_r}{r^2}\right] \\ &= -\frac{\varepsilon}{r^2} + \frac{\mathbb{E}_{Z_r|x}\left[e^{\frac{\varepsilon Z_r}{r^2}}(Z_r - \mathbb{E}_{Z_r|x}[Z_r])\right]}{r^2\, \mathbb{E}_{Z_r|x}\left[e^{\frac{\varepsilon Z_r}{r^2}}\right]} \end{aligned} \tag{13}$$

Now,

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

So,

$$\begin{aligned} \mathop{\mathbb{E}}_{Z_r|x}\left[e^{\frac{\varepsilon Z_r}{r^2}}(Z_r - \mathop{\mathbb{E}}_{Z_r|x}[Z_r])\right]^4 &\lesssim \mathop{\mathbb{E}}_{Z_r|x}\left[(e^{\frac{\varepsilon Z_r}{r^2}} - 1)Z_r\right]^4 + \mathop{\mathbb{E}}_{Z_r|x}\left[e^{\frac{\varepsilon Z_r}{r^2}} - 1\right]^4 \mathop{\mathbb{E}}_{Z_r|x}[Z_r]^4 \\ &\lesssim \mathop{\mathbb{E}}_{Z_r|x}\left[(e^{\frac{\varepsilon Z_r}{r^2}} - 1)^4\right] \mathop{\mathbb{E}}_{Z_r|x}[Z_r^4] \quad \text{by Cauchy-Schwarz and Jensen's inequalities} \end{aligned}$$

So, we have, by Cauchy-Schwarz and Jensen's inequalities,

$$\mathop{\mathbb{E}}_{x}\left[\mathop{\mathbb{E}}_{Z_r|x}\left[e^{\frac{\varepsilon Z_r}{r^2}}(Z_r - \mathop{\mathbb{E}}_{Z_r|x}[Z_r])\right]^4\right] \lesssim \sqrt{\mathop{\mathbb{E}}_{Z_r}\left[(e^{\frac{\varepsilon Z_r}{r^2}} - 1)^8\right] \mathop{\mathbb{E}}_{Z_r}[Z_r^8]} \tag{14}$$

We will now bound $\mathbb{E}_{Z_r}\left[(e^{\frac{\varepsilon Z_r}{r^2}}-1)^8\right]$. By a Taylor expansion, when $|\varepsilon z|\le r^2$

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

so that

$$\mathop{\mathbb{E}}_{Z_r\sim\mathcal{N}(0,r^2)}\left[(e^{\frac{\varepsilon Z_r}{r^2}}-1)^8\mathbb{1}_{|\varepsilon Z_r|\le r^2}\right]\lesssim\frac{\varepsilon^8}{r^{16}}\mathop{\mathbb{E}}_{Z_r}\left[Z_r^8\right]\lesssim\frac{\varepsilon^8}{r^8}$$

On the other hand, when $|\varepsilon z|\ge r^2$, we have $(e^{\frac{\varepsilon z}{r^2}}-1)^8\lesssim e^{\frac{8|\varepsilon z|}{r^2}}$, meaning that

$$\begin{aligned}\mathop{\mathbb{E}}_{Z_r\sim\mathcal{N}(0,r^2)}\left[(e^{\frac{\varepsilon Z_r}{r^2}}-1)^8\mathbb{1}_{|\varepsilon Z_r|>r^2}\right]&\lesssim\int_{|r^2/\varepsilon|}^{\infty}\frac{1}{\sqrt{2\pi}r}e^{\frac{8|\varepsilon z|}{r^2}}e^{-\frac{z^2}{2r^2}}dz\\&=e^{\frac{32\varepsilon^2}{r^2}}\int_{|r^2/\varepsilon|}^{\infty}\frac{1}{\sqrt{2\pi}r}e^{-\frac{(z-8|\varepsilon|)^2}{2r^2}}dz\\&\lesssim\mathop{\Pr}_{Z_r\sim\mathcal{N}(0,r^2)}\left[Z_r\ge r^2/|\varepsilon|-8|\varepsilon|\right]\\&\lesssim e^{-\frac{(|r^2/\varepsilon|-8|\varepsilon|)^2}{2r^2}}\\&\lesssim\frac{\varepsilon^8}{r^8}\end{aligned}$$

Also, $\mathbb{E}_{Z_r}[Z_r^8]\lesssim r^8$ So, we have shown in (14),

$$\mathop{\mathbb{E}}_{x}\left[\mathop{\mathbb{E}}_{Z_r|x}\left[e^{\frac{\varepsilon Z_r}{r^2}}(Z_r-\mathop{\mathbb{E}}_{Z_r|x}[Z_r])\right]^4\right]\lesssim\frac{\varepsilon^4}{r^4}\sqrt{\mathop{\mathbb{E}}_{Z_r}[Z_r^8]}\lesssim\varepsilon^4$$

Also, using Jensen's inequality

$$\begin{aligned}\mathop{\mathbb{E}}_{x}\left[\frac{1}{\mathbb{E}_{Z_r|x}\left[e^{\frac{\varepsilon Z_r}{r^2}}\right]^4}\right]&\le\mathop{\mathbb{E}}_{x}\left[e^{-4\varepsilon\,\mathbb{E}_{Z_r|x}[Z_r/r^2]}\right]\\&=\mathop{\mathbb{E}}_{x}\left[e^{-4\varepsilon s_r(x)}\right]\quad\text{since } s_r(x)=\mathop{\mathbb{E}}_{Z_r|x}[Z_r]/r^2\text{ by Lemma 36}\\&\le e^{8\varepsilon^2/r^2}\lesssim 1\quad\text{by Lemma 35 and since }|\varepsilon|\le r/60\end{aligned}$$

Then, using (13),

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim f_r}\left[(s_r(x+\varepsilon)-s_r(x))^2\right]&\lesssim\frac{\varepsilon^2}{r^4}+\frac{1}{r^4}\mathop{\mathbb{E}}_{x}\left[\frac{\mathbb{E}_{Z_r|x}\left[e^{\frac{\varepsilon Z_r}{r^2}}(Z_r-\mathbb{E}_{Z_r|x}[Z_r])\right]^2}{\mathbb{E}_{Z_r|x}[e^{\frac{\varepsilon Z_r}{r^2}}]^2}\right]\\&\le\frac{\varepsilon^2}{r^4}+\frac{1}{r^4}\sqrt{\mathop{\mathbb{E}}_{x}\left[\mathop{\mathbb{E}}_{Z_r|x}\left[e^{\frac{\varepsilon Z_r}{r^2}}(Z_r-\mathop{\mathbb{E}}_{Z_r|x}[Z_r])\right]^4\right]\mathop{\mathbb{E}}_{x}\left[\frac{1}{\mathbb{E}_{Z_r|x}\left[e^{\frac{\varepsilon Z_r}{r^2}}\right]^4}\right]}\\&\lesssim\frac{\varepsilon^2}{r^4}\end{aligned}$$

■

The next lemma shows that $\widehat{s}_r^{\text{sym}}$ approximates $s_r$ in a certain sense. Using this, we obtain that the second moment of $\widehat{s}_r^{\text{sym}}$ is close the $\mathcal{I}_r$.

**Lemma 27** *Let $f^*$ be an arbitrary symmetric distribution with mean $\mu$, and let $f_r$ be the $r$-smoothed version of $f^*$. Let $s_r$ be the score function of $f_r$, and let $\mathcal{I}_r$ be the Fisher information. Let $\widehat{s}_r^{sym}$ be the symmetrized clipped Kernel Density Estimate score from $N$ samples, symmetrized around $\mu+\varepsilon$ for $|\varepsilon|\le r/60$, as defined in (12). Let $\gamma > C$ be a parameter for large enough constant $C$. Then for any $r \le \sigma$ and $\frac{N}{\log\frac{1}{\delta}} \ge \left(\frac{\gamma^{5/12}\sigma}{r}\right)^{6+\beta}$ for $\beta > 0$, if $|\varepsilon|\le r^2\sqrt{\frac{\mathcal{I}_r}{\gamma}}$, with probability $1-\delta$,*

$$\mathop{\mathbb{E}}_{x\sim f_r}\left[(\widehat{s}_r^{sym}(x) - s_r(x))^2\right] \lesssim \frac{\mathcal{I}_r}{\gamma}$$

**Proof** By definition of $\widehat{s}_r^{\text{sym}}$, and using Lemma 3,

$$\mathop{\mathbb{E}}_{x\sim f_r}\left[(\widehat{s}_r^{\text{sym}}(x) - s_r(x))^2\mathbb{1}_{x\ge\mu+\varepsilon}\right] = \mathop{\mathbb{E}}_{x\sim f_r}\left[(\widehat{s}_r^{\text{clip}}(x) - s_r(x))^2\mathbb{1}_{x\ge\mu+\varepsilon}\right] \lesssim \frac{\mathcal{I}_r}{\gamma}$$

On the other hand, by Lemmas 3 and 26

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x\sim f_r}\left[(\widehat{s}_r^{\text{sym}}(x) - s_r(x))^2\mathbb{1}_{x<\mu+\varepsilon}\right]\\
&= \mathop{\mathbb{E}}_{x\sim f_r}\left[(-\widehat{s}_r^{\text{clip}}(2(\mu+\varepsilon)-x) + s_r(2\mu - x))^2\mathbb{1}_{x<\mu+\varepsilon}\right]\\
&\le \mathop{\mathbb{E}}_{x\sim f_r}\left[(-\widehat{s}_r^{\text{clip}}(2(\mu+\varepsilon)-x) + s_r(2(\mu+\varepsilon) - x))^2\mathbb{1}_{x<\mu+\varepsilon}\right] + \mathop{\mathbb{E}}_{x\sim f_r}\left[(s_r((2\mu-x)+2\varepsilon) - s_r(2\mu-x))^2\right]\\
&\lesssim \frac{\mathcal{I}_r}{\gamma} + \frac{\varepsilon^2}{r^4}
\end{aligned}$$

The claim follows since $\frac{\varepsilon^2}{r^4} \le \frac{\mathcal{I}_r}{\gamma}$. ■

**Lemma 4 (Symmetrized Clipped KDE score variance)** *Let $\widehat{s}_r^{sym}$ be the symmetrized clipped Kernel Density Estimate score from $N$ samples, symmetrized around $\mu+\varepsilon$ for $|\varepsilon|\le r/60$, as defined in (12). Let $\gamma > C$ be a parameter for large enough constant $C$. Then for any $r \le \sigma$ and $\frac{N}{\log\frac{1}{\delta}} \ge \left(\frac{\gamma^{5/12}\sigma}{r}\right)^{6+\beta}$ for $\beta > 0$, if $|\varepsilon|\le r^2\sqrt{\frac{\mathcal{I}_r}{\gamma}}$, with probability $1-\delta$,*

$$\left|\mathop{\mathbb{E}}_{x\sim f_r}\left[\widehat{s}_r^{sym}(x)^2\right] - \mathcal{I}_r\right|\lesssim \frac{\mathcal{I}_r}{\sqrt{\gamma}}$$

**Proof** We have

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim f_r}\left[\widehat{s}_r^{\text{sym}}(x)^2\right] &= \mathop{\mathbb{E}}_{x\sim f_r}\left[(s_r(x) + \widehat{s}_r^{\text{sym}}(x) - s_r(x))^2\right]\\
&= \mathop{\mathbb{E}}_{x\sim f_r}\left[s_r(x)^2\right] + 2\mathop{\mathbb{E}}_{x\sim f_r}\left[s_r(x)(\widehat{s}_r^{\text{sym}}(x) - s_r(x))\right] + \mathop{\mathbb{E}}_{x\sim f_r}\left[(\widehat{s}_r^{\text{sym}}(x) - s_r(x))^2\right]\\
&\le \mathop{\mathbb{E}}_{x\sim f_r}\left[s_r(x)^2\right] + 2\sqrt{\mathop{\mathbb{E}}_{x\sim f_r}\left[s_r(x)^2\right]\mathop{\mathbb{E}}_{x\sim f_r}\left[(\widehat{s}^{\text{sym}}(x) - s_r(x))^2\right]} + \mathop{\mathbb{E}}_{x\sim f_r}\left[(\widehat{s}_r^{\text{sym}}(x) - s_r(x))^2\right]
\end{aligned}$$

So, by Lemma 27,

$$\mathop{\mathbb{E}}_{x\sim f_r}\left[\widehat{s}_r^{\text{sym}}(x)^2\right]-\mathcal{I}_r\lesssim\frac{\mathcal{I}_r}{\sqrt{\gamma}}+\frac{\mathcal{I}_r}{\gamma}\lesssim\frac{\mathcal{I}_r}{\sqrt{\gamma}}$$

Similarly,

$$\mathop{\mathbb{E}}_{x\sim f_r}\left[s_r(x)^2\right]\le\mathop{\mathbb{E}}_{x\sim f_r}\left[\widehat{s}_r^{\text{sym}}(x)^2\right]+2\sqrt{\mathop{\mathbb{E}}_{x\sim f_r}[\widehat{s}_r^{\text{sym}}(x)^2]\mathop{\mathbb{E}}_{x\sim f_r}\left[(\widehat{s}_r^{\text{sym}}(x)-s_r(x))^2\right]}+\mathop{\mathbb{E}}_{x\sim f_r}\left[(\widehat{s}_r^{\text{sym}}(x)-s_r(x))^2\right]$$

So, since we showed $\mathbb{E}_{x\sim f_r}[\widehat{s}_r^{\text{sym}}(x)^2]\lesssim\mathcal{I}_r$, we have

$$\mathop{\mathbb{E}}_{x\sim f_r}[s_r(x)^2]\le\mathop{\mathbb{E}}_{x\sim f_r}\left[\widehat{s}_r^{\text{sym}}(x)^2\right]+O\left(\frac{\mathcal{I}_r}{\sqrt{\gamma}}\right)$$

so that

$$\mathop{\mathbb{E}}_{x\sim f_r}\left[\widehat{s}_r^{\text{sym}}(x)^2\right]-\mathcal{I}_r\gtrsim\frac{\mathcal{I}_r}{\sqrt{\gamma}}$$

The claim follows. ■

Finally, we show that the expectation of the symmetrized, clipped KDE score function $\widehat{s}_r^{\text{sym}}$ symmetrized around $\mu+\varepsilon$ for small $\varepsilon$ is close to $\varepsilon\mathcal{I}_r$.

**Lemma 5 (Symmetrized Clipped KDE score mean)** *Let $\widehat{s}_r^{sym}$ be the symmetrized clipped Kernel Density Estimate score from $N$ samples, symmetrized around $\mu+\varepsilon$ for $|\varepsilon|\le r/60$, as defined in (12). Let $\gamma>C$ be a parameter for large enough constant $C$. Then for any $r\le\sigma$ and $\frac{N}{\log\frac{1}{\delta}}\ge\left(\frac{\gamma^{5/12}\sigma}{r}\right)^{6+\beta}$ for $\beta>0$, if $|\varepsilon|\le r^2\sqrt{\frac{\mathcal{I}_r}{\gamma}}$, with probability $1-\delta$,*

$$\left|\mathop{\mathbb{E}}_{x\sim f_r}[\widehat{s}_r^{sym}(x)]-\varepsilon\mathcal{I}_r\right|\lesssim\frac{\varepsilon\mathcal{I}_r}{\sqrt{\gamma}}$$

**Proof** Since $f_r$ is symmetric around $\mu$, $f_r(x)=f_r(2\mu-x)$. So using the definition of $\widehat{s}_r^{\text{sym}}$, we have

$$\begin{aligned}\int_{-\infty}^{\mu+\varepsilon}f_r(x-\varepsilon)\widehat{s}_r^{\text{sym}}(x)dx&=-\int_{-\infty}^{\mu+\varepsilon}f_r(2\mu-x+\varepsilon)s_r^{\text{clip}}(2(\mu+\varepsilon)-x)dx\\&=\int_{\infty}^{\mu+\varepsilon}f_r(y-\varepsilon)s_r^{\text{clip}}(y)dy\quad\text{Substituting } y=2(\mu+\varepsilon)-x\\&=-\int_{\mu+\varepsilon}^{\infty}f_r(y-\varepsilon)s_r^{\text{sym}}(y)dy\quad\text{since } s_r^{\text{sym}}(x)=s_r^{\text{clip}}(x)\text{ for } x\ge\mu+\varepsilon\end{aligned}$$

So, we have

$$\mathop{\mathbb{E}}_{x\sim f_r}\left[\frac{f_r(x-\varepsilon)}{f_r(x)}\widehat{s}_r^{\mathsf{sym}}(x)\right] = \int_{-\infty}^{\mu+\varepsilon} f_r(x-\varepsilon)\widehat{s}_r^{\mathsf{sym}}(x)dx + \int_{\mu+\varepsilon}^{\infty} f_r(x-\varepsilon)s_r^{\mathsf{sym}}(x)dx = 0$$

So,

$$\mathop{\mathbb{E}}_{x\sim f_r}\left[\widehat{s}_r^{\mathsf{sym}}(x)\right] = \mathop{\mathbb{E}}_{x\sim f_r}\left[\frac{f_r(x)-f_r(x-\varepsilon)}{f_r(x)}\widehat{s}_r^{\mathsf{sym}}(x)\right]$$

Thus,

$$\begin{aligned}
&\left(\mathop{\mathbb{E}}_{x\sim f_r}\left[\widehat{s}_r^{\mathsf{sym}}(x)\right] - \varepsilon\mathop{\mathbb{E}}_{x\sim f_r}\left[\widehat{s}_r^{\mathsf{sym}}(x)^2\right]\right)^2\\
&= \mathop{\mathbb{E}}_{x\sim f_r}\left[\left(\frac{f_r(x)-f_r(x-\varepsilon)}{f_r(x)} - \varepsilon\widehat{s}_r^{\mathsf{sym}}(x)\right)\widehat{s}_r^{\mathsf{sym}}(x)\right]^2\\
&= \mathop{\mathbb{E}}_{x\sim f_r}\left[\left(\frac{f_r(x)-f_r(x-\varepsilon)-\varepsilon f_r'(x)}{f_r(x)} + \varepsilon(s_r(x)-\widehat{s}_r^{\mathsf{sym}}(x))\right)\widehat{s}_r^{\mathsf{sym}}(x)\right]^2 \qquad \text{since } s_r(x) = \frac{f_r'(x)}{f_r(x)}\\
&\le \mathop{\mathbb{E}}_{x\sim f_r}\left[\left(\frac{f_r(x)-f_r(x-\varepsilon)-\varepsilon f_r'(x)}{f_r(x)} + \varepsilon(s_r(x)-\widehat{s}_r^{\mathsf{sym}}(x))\right)^2\right]\mathop{\mathbb{E}}_{x\sim f_r}\left[\widehat{s}_r^{\mathsf{sym}}(x)^2\right]\\
&\lesssim \left(\mathop{\mathbb{E}}_{x\sim f_r}\left[\left(\frac{f_r(x)-f_r(x-\varepsilon)-\varepsilon f_r'(x)}{f_r(x)}\right)^2\right] + \varepsilon^2\mathop{\mathbb{E}}_{x\sim f_r}\left[(s_r(x)-\widehat{s}_r^{\mathsf{sym}}(x))^2\right]\right)\mathop{\mathbb{E}}_{x\sim f_r}\left[\widehat{s}_r^{\mathsf{sym}}(x)^2\right]\\
&= \left(\mathop{\mathbb{E}}_{x\sim f_r}\left[\Delta_{-\varepsilon}(x)^2\right] + \varepsilon^2\mathop{\mathbb{E}}_{x\sim f_r}\left[(s_r(x)-\widehat{s}_r^{\mathsf{sym}}(x))^2\right]\right)\mathop{\mathbb{E}}_{x\sim f_r}\left[\widehat{s}_r^{\mathsf{sym}}(x)^2\right]
\end{aligned}$$

where $\Delta_\varepsilon(x) = \frac{f_r(x+\varepsilon)-f_r(x)-\varepsilon f_r'(x)}{f_r(x)}$. But by Lemma 40, $\mathbb{E}_{x\sim f_r}\left[\Delta_{-\varepsilon}(x)^2\right] \lesssim \frac{\varepsilon^4}{r^4} \lesssim \varepsilon^2\frac{\mathcal{I}_r}{\gamma}$ (since $|\varepsilon| \le r^2\sqrt{\frac{\mathcal{I}_r}{\gamma}}$). So, by Lemma 27 and Lemma 4, we have

$$\begin{aligned}
\left|\mathop{\mathbb{E}}_{x\sim f_r}\left[\widehat{s}_r^{\mathsf{sym}}(x)\right] - \varepsilon\mathop{\mathbb{E}}_{x\sim f_r}\left[\widehat{s}_r^{\mathsf{sym}}(x)^2\right]\right| &\lesssim \sqrt{\left(\varepsilon^2\frac{\mathcal{I}_r}{\gamma} + \varepsilon^2\frac{\mathcal{I}_r}{\gamma}\right)\mathcal{I}_r}\\
&\lesssim \frac{\varepsilon\mathcal{I}_r}{\sqrt{\gamma}}
\end{aligned}$$

Then, using Lemma 4 once again on $\mathbb{E}_{x\sim f_r}[s_r^{\mathsf{sym}}(x)^2]$ on the LHS, the claim follows. ■

## Appendix D. Estimating $\mathcal{I}_r$

In this section, we show that $\widehat{\mathcal{I}}_r = \mathbb{E}_{x\sim\widehat{f}_r}[\widehat{s}_r^{\mathsf{sym}}(x)^2]$ provides a good estimate of $\mathcal{I}_r$.

**Lemma 28** *Let $\widehat{f}_r$ be the kernel density estimate of $f_r$ from $N$ samples $Y_1,\dots,Y_N \sim f^*$ as defined in (8). Let $x, \varepsilon, N$ be such that $N \ge \frac{12\log\frac{2}{\delta}}{f_r(x)r}$ and $\frac{16|\varepsilon|}{r}\sqrt{\log\frac{1}{rf_r(x)}} \le 1$. We have that with probability $1-\delta$, for all $|\zeta| \le |\varepsilon|$ simultaneously,*

$$|\widehat{f}_r(x+\zeta) - f_r(x+\zeta)| \lesssim \sqrt{\frac{f_r(x+\zeta)\log\frac{2}{\delta}}{Nr}} + \frac{|\varepsilon|f_r(x+\zeta)}{r}\sqrt{\log\frac{2}{f_r(x+\zeta)r}}$$

**Proof** By Lemma 11, with probability $1-\delta$,

$$|\widehat{f}_r(x) - f_r(x)|\leq \sqrt{\frac{3f_r(x)\log\frac{2}{\delta}}{Nr}}$$

Note that since $N \geq \frac{12\log\frac{2}{\delta}}{f_r(x)r}$, the RHS above is at most $\frac{f_r(x)}{2}$.

Also, by Lemma 15,

$$|f_r(x+\zeta) - f_r(x)|\leq \frac{10|\varepsilon|f_r(x)}{r}\sqrt{\log\frac{1}{f_r(x)r}}$$

Now, since $\frac{8|\varepsilon|}{r}\sqrt{\log\frac{1}{r\widehat{f}_r(x)}} \leq \frac{8|\varepsilon|}{r}\sqrt{\log\frac{2}{rf_r(x)}} \leq \frac{16|\varepsilon|}{r}\sqrt{\log\frac{1}{rf_r(x)}} < 1$, by Lemma 15,

$$\begin{aligned}|\widehat{f}_r(x+\zeta) - \widehat{f}_r(x)| &\leq \frac{10|\varepsilon|\widehat{f}_r(x)}{r}\sqrt{\log\frac{1}{\widehat{f}_r(x)r}}\\ &\leq \frac{15|\varepsilon|f_r(x)}{r}\sqrt{\log\frac{2}{f_r(x)r}}\end{aligned}$$

So, putting everything together

$$\begin{aligned}|\widehat{f}_r(x+\zeta) - f_r(x+\zeta)| &\leq |\widehat{f}_r(x+\zeta) - \widehat{f}_r(x)|+|\widehat{f}_r(x) - f_r(x)|+|f_r(x) - f_r(x+\zeta)|\\ &\leq \sqrt{\frac{3f_r(x)\log\frac{2}{\delta}}{Nr}} + \frac{30|\varepsilon|f_r(x)}{r}\sqrt{\log\frac{2}{f_r(x)r}}\end{aligned}$$

Now, since $\frac{16|\varepsilon|}{r}\sqrt{\log\frac{1}{rf_r(x)}} \leq 1$ so that by Lemma 15

$$|f_r(x+\zeta) - f_r(x)|\leq \frac{5}{8}f_r(x)$$

The claim follows. ■

**Lemma 29** *Let $f^*$ be an arbitrary distribution with mean $\mu$ and variance $\sigma^2$ and let $f_r$ be the $r$-smoothed version of $f^*$, with variance $\sigma_r^2 = \sigma^2 + r^2$. Let $\widehat{f}_r$ be the kernel density estimate of $f_r$ from $N$ samples $Y_1, \dots, Y_N \sim f^*$ as defined in (8). Let $0 < \alpha \leq \frac{1}{\sqrt{2\pi r}}$ and let $N \geq \frac{12\log\left(\frac{2}{\delta}\left(\frac{2t\sigma_r\sqrt{\alpha N}}{\sqrt{r}}+1\right)\right)+400\log\frac{1}{\alpha r}}{\alpha r}$. Then, with probability $1-\delta$, for all $x$ such that $|x-\mu|\leq t\sigma_r$ and $f_r(x) \geq \alpha$ simultaneously,*

$$|\widehat{f}_r(x) - f_r(x)|\lesssim f_r(x)\sqrt{\frac{1}{\alpha r N}\log\left(\frac{2}{\delta}\left(\frac{2t\sigma_r\sqrt{\alpha N}}{\sqrt{r}}+1\right)\right)\log\frac{2}{\alpha r}}$$

**Proof** Consider contiguous intervals of length $\varepsilon$ starting from $\mu - t\sigma_r$ so that the last interval covers $\mu + t\sigma_r$, and let $S$ be the set of the smallest $y$ such that $f_r(y) \geq \alpha$ in each of these intervals, if one exists. Note that $|S|\leq \frac{2t\sigma_r}{\varepsilon} + 1$. Then, we have that

$$\{x : |x-\mu|\leq t\sigma_r \text{ and } f_r(x) \geq \alpha\} \subseteq \{[y-\varepsilon, y+\varepsilon]|y \in S\}$$

Now for $\varepsilon = \sqrt{\frac{r}{\alpha N}}$ and $y \in S$, since $N \geq \frac{12\log\left(\frac{2}{\delta}\left(\frac{2t\sigma_r\sqrt{\alpha N}}{\sqrt{r}}+1\right)\right)}{\alpha r} \geq \frac{12\log\frac{2|S|}{\delta}}{f_r(y)r}$ and $\frac{16|\varepsilon|}{r}\sqrt{\log\frac{1}{rf_r(y)}} = \frac{16}{\sqrt{\alpha rN}}\sqrt{\log\frac{1}{f_r(y)r}} \leq \frac{16}{\sqrt{\alpha rN}}\sqrt{\log\frac{1}{\alpha r}} \leq 1$, we have by Lemma 28 that with probability $1 - \frac{\delta}{|S|}$, for all $|\zeta|\leq|\varepsilon|$ simultaneously,

$$\begin{aligned}
\left|\widehat{f}_r(y+\zeta) - f_r(y+\zeta)\right| &\lesssim \sqrt{\frac{f_r(y+\zeta)\log\frac{2|S|}{\delta}}{Nr}} + f_r(y+\zeta)\sqrt{\frac{1}{\alpha Nr}\log\frac{2}{\alpha r}} \\
&\lesssim f_r(y+\zeta)\sqrt{\frac{1}{\alpha Nr}\log\left(\frac{2}{\delta}\left(\frac{2t\sigma_r\sqrt{\alpha N}}{\sqrt{r}}+1\right)\right)\log\frac{2}{\alpha r}} \\
&\text{since } f_r(y+\zeta) \gtrsim \alpha \text{ and } \alpha \leq \frac{1}{\sqrt{2\pi r}} \text{ so that } \log\frac{2}{\alpha r} > 1
\end{aligned}$$

So by a union bound, with probability $1-\delta$, for all $x$ such that $|x-\mu|\leq t\sigma_r$ and $f_r(x) \geq \alpha$ simultaneously,

$$|\widehat{f}_r(x) - f_r(x)|\lesssim f_r(x)\sqrt{\frac{1}{\alpha Nr}\log\left(\frac{2}{\delta}\left(\frac{2t\sigma_r\sqrt{\alpha N}}{\sqrt{r}}+1\right)\right)\log\frac{2}{\alpha r}}$$

■

**Lemma 30** *Let $S$ be a set and let $\widehat{f}_r$ be the kernel density estimate of $f_r$ from $N$ samples $Y_1, \ldots, Y_N \sim f^*$. Then, with probability $1-\delta$,*

$$\left|\Pr_{\widehat{f}_r}[S] - \Pr_{f_r}[S]\right| \lesssim \sqrt{\frac{\log\frac{2}{\delta}}{N}}$$

**Proof**

$$\mathop{\mathbb{E}}_{Y_i}\left[\left[\Pr_{\widehat{f}_r}[S]\right]\right] = \mathop{\mathbb{E}}_{Y_i}\left[\int_S \widehat{f}_r(x)dx\right] = \int_S \mathop{\mathbb{E}}_{Y_i}\left[\widehat{f}_r(x)\right]dx = \int_S f_r(x)dx = \Pr_{f_r}[S]$$

Furthermore $0 \leq \Pr_{\widehat{f}_r}[S] \leq 1$. So, by Hoeffding's inequality, with probability $1-\delta$,

$$\left|\Pr_{\widehat{f}_r}[S] - \Pr_{f_r}[S]\right| \lesssim \sqrt{\frac{\log\frac{2}{\delta}}{N}}$$

■

**Lemma 31** *Let $\gamma \geq C$ for large enough constant $C$, and let $\frac{N}{\log \frac{1}{\delta}} \geq \left(\frac{\gamma}{r^4 \mathcal{I}_r^2}\right)^{1+\alpha}$ for small constant $\alpha > 0$. Let $\tilde{s}$ be any function with $|\tilde{s}_r(x)| \leq \frac{2}{r}\sqrt{\log \frac{N}{\log \frac{1}{\delta}}}$ for all $x$. Let $S$ be a set with $\Pr_{f_r}[S] \lesssim \frac{1}{t^2}$ for $t = \gamma^{1/4}\sqrt{\frac{\log \frac{N}{\mathcal{I}_r r^2 \log \frac{1}{\delta}}}{\mathcal{I}_r r^2}}$. Let $\widehat{f}_r$ be the Kernel density estimate of $f_r$ as defined in (8) from $N$ samples. Then, we have*

$$\int_S \left(\widehat{f}_r(x) - f_r(x)\right) \tilde{s}_r(x)^2 dx \lesssim \frac{\mathcal{I}_r}{\sqrt{\gamma}}$$

**Proof** Since $|\tilde{s}_r(x)| \leq \frac{2}{r}\sqrt{\log \frac{N}{\log \frac{1}{\delta}}}$ and $\Pr_{f_r}[S] \leq \frac{1}{t^2}$, we have

$$\begin{aligned}
\int_S \left(\widehat{f}_r(x) - f_r(x)\right) \tilde{s}_r(x)^2 dx &\leq \int_S \widehat{f}_r(x) \tilde{s}_r(x)^2 dx \\
&\leq \frac{4}{r^2} \log\left(\frac{N}{\log \frac{1}{\delta}}\right) \Pr_{\widehat{f}_r}[S] \\
&\lesssim \frac{4}{r^2} \log\left(\frac{N}{\log \frac{1}{\delta}}\right) \left(\Pr_{f_r}[S] + \sqrt{\frac{\log \frac{2}{\delta}}{N}}\right) \quad \text{by Lemma 30} \\
&\lesssim \frac{1}{t^2 r^2} \log\left(\frac{N}{\log \frac{1}{\delta}}\right) + \frac{1}{r^2}\sqrt{\frac{\log \frac{1}{\delta}}{N}} \log\left(\frac{N}{\log \frac{1}{\delta}}\right)
\end{aligned}$$

Now,

$$\begin{aligned}
\frac{1}{t^2 r^2} \log \frac{N}{\log \frac{1}{\delta}} &= \frac{\mathcal{I}_r}{\sqrt{\gamma} \log \frac{N}{\mathcal{I}_r r^2 \log \frac{1}{\delta}}} \log\left(\frac{N}{\log \frac{1}{\delta}}\right) \quad \text{by our setting of } t \\
&\lesssim \frac{\mathcal{I}_r}{\sqrt{\gamma} \log \frac{N}{\mathcal{I}_r r^2 \log \frac{1}{\delta}}} \log\left(\frac{N}{\mathcal{I}_r r^2 \log \frac{1}{\delta}}\right) \quad \text{since } \mathcal{I}_r \leq \frac{1}{r^2} \\
&\lesssim \frac{\mathcal{I}_r}{\sqrt{\gamma}}
\end{aligned}$$

Also,

$$\begin{aligned}
\frac{1}{r^2}\sqrt{\frac{\log \frac{2}{\delta}}{N}} \log\left(\frac{N}{\log \frac{1}{\delta}}\right) &\leq \frac{1}{r^2}\left(\frac{\log \frac{2}{\delta}}{N}\right)^{\frac{1}{2} - \frac{\alpha}{4}} \\
&\lesssim \frac{\mathcal{I}_r}{\sqrt{\gamma}} \quad \text{since } \frac{N}{\log \frac{2}{\delta}} \geq \left(\frac{\gamma}{r^4 \mathcal{I}_r^2}\right)^{1+\alpha}
\end{aligned}$$

So, the claim follows. ■

**Lemma 32** *Let $\widehat{f_r}$ be the kernel density estimate of $f_r$ from $N$ samples, as defined in (8). Let $\gamma \geq C$ for large constant $C \geq 1$ be a parameter. Suppose $\tilde{s}_r$ is a function such that $\mathbb{E}_{x \sim f_r}[\tilde{s}_r(x)] \lesssim \mathcal{I}_r$. Let $t = \gamma^{1/4}\sqrt{\frac{\log \frac{N}{\mathcal{I}_r r^2 \log \frac{1}{\delta}}}{\mathcal{I}_r r^2}}$, $\alpha = \frac{1}{t^3 \sigma_r}$. Then for any $r \leq \sigma$ and $\frac{N}{\log \frac{1}{\delta}} \geq \left(\gamma^{5/12}\frac{\sigma}{r}\right)^{6+\beta}$ for some small constant $\beta > 0$, with probability $1-\delta$, we have*

$$\int_{-\infty}^{\infty} \left|\widehat{f_r}(x) - f_r(x)\right| \tilde{s}_r(x)^2 \mathbb{1}_{\{|x-\mu| \leq t\sigma_r \textit{ and } f_r(x) \geq \alpha\}} dx \lesssim \frac{\mathcal{I}_r}{\sqrt{\gamma}}$$

**Proof** This proof is similar to the proof of Lemma 21. First note that since $r \leq \sigma$,

$$\sigma_r^2 = \sigma^2 + r^2 \leq 2\sigma^2$$

Note also that $\frac{N}{\log \frac{1}{\delta}} \geq 1$ since $\gamma \geq 1$ and $r \leq \sigma$. So our setting of $N$ implies WLOG

$$\frac{N}{\log \frac{1}{\delta}} \geq \left(\frac{\gamma^{5/12}\sigma \log \frac{N}{\log \frac{1}{\delta}}}{r}\right)^6$$

or

$$\frac{\sigma}{r} \leq \left(\frac{N}{\log \frac{1}{\delta}}\right)^{1/6} \cdot \frac{1}{\gamma^{5/12} \log \frac{N}{\log \frac{1}{\delta}}} \leq \left(\frac{N}{\log \frac{1}{\delta}}\right)^{1/6} \tag{15}$$

We will first check that this $N$ satisfies the condition required to invoke Lemma 29 that $N \geq \frac{12 \log\left(\frac{2}{\delta}\left(\frac{2t\sigma_r\sqrt{\alpha N}}{\sqrt{r}}+1\right)\right) + 400 \log \frac{1}{\alpha r}}{\alpha r}$. To do this, we individually upper bound $\frac{1}{\alpha r}$ and $\frac{2t\sigma_r\sqrt{\alpha N}}{\sqrt{r}}$. We have,

$$\begin{aligned}
\frac{1}{\alpha r} &= \frac{\sigma_r \gamma^{3/4}}{r}\left(\frac{\log \frac{N}{\mathcal{I}_r r^2 \log \frac{1}{\delta}}}{\mathcal{I}_r r^2}\right)^{3/2} && \text{since } \alpha = \frac{1}{t^3\sigma_r} \text{ and } t = \gamma^{1/4}\sqrt{\frac{\log \frac{N}{\mathcal{I}_r r^2 \log \frac{1}{\delta}}}{\mathcal{I}_r r^2}} \\
&\leq \left(\frac{\sigma_r}{r}\right)^4 \gamma^{3/4} \log^{3/2} \frac{N\sigma_r^2}{r^2 \log \frac{1}{\delta}} && \text{since } \mathcal{I}_r \geq \frac{1}{\sigma_r^2} \\
&\leq \left(\frac{2\sigma}{r}\right)^4 \gamma^{3/4} \log^{3/2} \frac{2N\sigma^2}{r^2 \log \frac{1}{\delta}} && \text{since } \sigma_r^2 \leq 2\sigma^2 \\
&\leq 16 \frac{N^{4/6}}{\gamma^{5/3} \log^4 \frac{N}{\log \frac{1}{\delta}} \log^{4/6} \frac{1}{\delta}} \cdot \gamma^{3/4} \log^{3/2}\left(2\left(\frac{N}{\log \frac{1}{\delta}}\right)^{4/3}\right) && \text{by (15)} \\
&\leq \frac{16N}{\gamma^{5/2} \log^4 \frac{N}{\log \frac{1}{\delta}} \log \frac{1}{\delta}} \gamma^{3/4} \log^{3/2}\left(2\left(\frac{N}{\log \frac{1}{\delta}}\right)^{4/3}\right) && \text{since } \frac{N}{\log \frac{1}{\delta}} \geq \gamma^{5/2} \\
&\leq \frac{N}{\gamma^{3/2} \log^2 \frac{N}{\log \frac{1}{\delta}} \log \frac{1}{\delta}} && \text{since } \gamma \geq C \text{ for large enough constant } C
\end{aligned}$$

To further justify the last line above, observe that $\frac{\gamma^{3/4}}{\gamma^{5/2}} \leq \frac{1}{\gamma^{3/2}.\gamma^{1/4}}$, and that for large enough constant $C$, since $\gamma \geq C$, $\gamma^{1/4}$ can be made larger than any fixed constant. Also note that $\log^{3/2}\left(2\left(\frac{N}{\log\frac{1}{\delta}}\right)^{4/3}\right) \leq \log^2\frac{N}{\log\frac{1}{\delta}}$ for large enough constant $C$ since $\frac{N}{\log\frac{1}{\delta}} \geq \gamma \geq C$. So the inequality follows. Next we bound $\frac{2t\sigma_r\sqrt{\alpha N}}{\sqrt{r}}$.

$$\begin{aligned}
\frac{2t\sigma_r\sqrt{\alpha N}}{\sqrt{r}} &= 2\sqrt{\frac{N\sigma_r}{tr}} \quad \text{since } \alpha = \frac{1}{t^3\sigma_r} \\
&= 2\sqrt{\frac{N\sigma_r}{r\gamma^{1/4}\sqrt{\frac{\log\frac{N}{\mathcal{I}_r r^2\log\frac{1}{\delta}}}{\mathcal{I}_r r^2}}}} \quad \text{since } t = \gamma^{1/4}\sqrt{\frac{\log\frac{N}{\mathcal{I}_r r^2\log\frac{1}{\delta}}}{\mathcal{I}_r r^2}} \\
&\leq 4\sqrt{\frac{N\sigma}{r\gamma^{1/4}\sqrt{\log\frac{N}{\log\frac{1}{\delta}}}}} \quad \text{since } \mathcal{I}_r \leq \frac{1}{r^2} \text{ and } \sigma_r^2 \leq 2\sigma^2 \\
&\leq 4\sqrt{\frac{N}{\gamma^{1/4}\sqrt{\log\frac{N}{\log\frac{1}{\delta}}}} \cdot \left(\frac{N}{\log\frac{1}{\delta}}\right)^{1/6}} \quad \text{by (15)} \\
&\leq 4N \cdot \left(\frac{N}{\log\frac{1}{\delta}}\right)^{1/12} \quad \text{since } \gamma \geq 1, \frac{N}{\log\frac{1}{\delta}} \geq 1
\end{aligned}$$

So, we can now check the condition required to invoke Lemma 29. We have

$$\begin{aligned}
&\frac{12\log\left(\frac{2}{\delta}\left(\frac{2t\sigma_r\sqrt{\alpha N}}{\sqrt{r}}+1\right)\right)+400\log\frac{1}{\alpha r}}{\alpha r} \\
&\leq \left(12\log\left(\frac{2}{\delta}\left(4N\cdot\left(\frac{N}{\log\frac{1}{\delta}}\right)^{1/12}+1\right)\right)+400\log\frac{N}{\log\frac{1}{\delta}}\right)\left(\frac{N}{\gamma^{3/2}\log^2\frac{N}{\log\frac{1}{\delta}}\log\frac{1}{\delta}}\right) \\
&\leq N \quad \text{since } \gamma \geq C
\end{aligned}$$

So, by Lemma 29, we have

$$|\widehat{f}_r(x) - f_r(x)| \lesssim f_r(x)\sqrt{\frac{1}{\alpha r N}\log\left(\frac{2}{\delta}\left(\frac{2t\sigma_r\sqrt{\alpha N}}{\sqrt{r}}+1\right)\right)\log\frac{2}{\alpha r}}$$

We will show that the RHS above is bounded by $O\left(\frac{f_r(x)}{\sqrt{\gamma}}\right)$. Since we showed that $\frac{1}{\alpha r} \leq \frac{N}{\gamma^{3/2}\log^2\frac{N}{\log\frac{1}{\delta}\log\frac{1}{\delta}}}$, we have that

$$\frac{1}{rN} \leq \frac{\alpha}{\gamma^{3/2}\log^2\frac{N}{\log\frac{1}{\delta}}\log\frac{1}{\delta}}$$

So, plugging this into the RHS above, along with our bounds for $\frac{2t\sigma_r\sqrt{\alpha N}}{\sqrt{r}}$ and $\frac{1}{\alpha r}$, we have,

$$\begin{aligned}
&|\widehat{f}_r(x) - f_r(x)| \\
&\lesssim \frac{f_r(x)}{\sqrt{\gamma}} \cdot \sqrt{\frac{1}{\gamma^{1/2}\log^2\frac{N}{\log\frac{1}{\delta}}\log\frac{1}{\delta}} \log\left(\frac{2}{\delta}\left(4N\cdot\left(\frac{N}{\log\frac{1}{\delta}}\right)^{1/12}\right)+1\right)\log\left(\frac{N}{\gamma^{3/2}\log^2\frac{N}{\log\frac{1}{\delta}}\log\frac{1}{\delta}}\right)} \\
&\lesssim \frac{f_r(x)}{\sqrt{\gamma}}
\end{aligned}$$

So finally,

$$\begin{aligned}
&\int_{-\infty}^{\infty}\left|\widehat{f}_r(x) - f_r(x)\right|\tilde{s}_r(x)^2 \mathbb{1}_{\{|x-\mu|\le t\sigma_r \text{ and } f_r(x)\ge\alpha\}}dx \\
&\lesssim \frac{1}{\sqrt{\gamma}}\int_{-\infty}^{\infty} f_r(x)\tilde{s}_r(x)^2\mathbb{1}_{\{|x-\mu|\le t\sigma_r \text{ and } f_r(x)\ge\alpha\}}dx \\
&\le \frac{1}{\sqrt{\gamma}}\mathop{\mathbb{E}}_{x\sim f_r}[\tilde{s}_r(x)^2] \\
&\lesssim \frac{\mathcal{I}_r}{\sqrt{\gamma}} \quad \text{since } \mathop{\mathbb{E}}_{x\sim f_r}[\tilde{s}_r(x)^2] \lesssim \mathcal{I}_r \text{ by assumption}
\end{aligned}$$

■

**Lemma 6 (Smoothed Fisher information Estimation)** *Let $\gamma \ge C$ for large constant $C \ge 1$ be a parameter. Suppose we have a function $\tilde{s}_r$ that satisfies for $r \le \sigma$*

$$\left|\mathop{\mathbb{E}}_{x\sim f_r}[\tilde{s}_r(x)^2] - \mathcal{I}_r\right| \lesssim \frac{\mathcal{I}_r}{\sqrt{\gamma}}$$

*and that $|\tilde{s}_r(x)| \le \frac{2}{r}\sqrt{\log\frac{N}{\log\frac{1}{\delta}}}$ for all $x$. Let $\widehat{f}_r$ be the kernel density estimate for $f_r$ from $N$ samples, as defined in (8). Then, for $\frac{N}{\log\frac{1}{\delta}} \ge \left(\gamma^{5/12}\frac{\sigma}{r}\right)^{6+\beta}$ for some small constant $\beta > 0$, with probability $1-\delta$, we have*

$$\left|\mathop{\mathbb{E}}_{x\sim \widehat{f}_r}[\tilde{s}_r(x)^2] - \mathcal{I}_r\right| \lesssim \frac{\mathcal{I}_r}{\sqrt{\gamma}}$$

**Proof** We have

$$\mathop{\mathbb{E}}_{x\sim \widehat{f}_r}\left[\tilde{s}_r(x)^2\right] = \mathop{\mathbb{E}}_{x\sim f_r}\left[\tilde{s}_r(x)^2\right] + \int_{-\infty}^{\infty}\left(\widehat{f}_r(x) - f_r(x)\right)\tilde{s}_r(x)^2 dx$$

So, by our assumption,

$$\left|\mathop{\mathbb{E}}_{x\sim \widehat{f}_r}[\tilde{s}_r(x)^2] - \mathcal{I}_r\right| \lesssim \frac{\mathcal{I}_r}{\sqrt{\gamma}} + \int_{-\infty}^{\infty}\left(\widehat{f}_r(x) - f_r(x)\right)\tilde{s}_r(x)^2 dx$$

It remains to bound the integral in the RHS above by $O\left(\frac{\mathcal{I}_r}{\sqrt{\gamma}}\right)$. Let $t = \gamma^{1/4}\sqrt{\frac{\log \frac{N}{\mathcal{I}_r r^2 \log\frac{1}{\delta}}}{\mathcal{I}_r r^2}}$, $\alpha = \frac{1}{t^3\sigma_r}$. We have

$$\begin{aligned}\int_{-\infty}^{\infty}\left(\widehat{f}_r(x) - f_r(x)\right)\tilde{s}_r(x)^2 dx = &\int_{-\infty}^{\infty}\left(\widehat{f}_r(x) - f_r(x)\right)s_r(x)^2\mathbb{1}_{\{|x-\mu|\le t\sigma_r \text{ and } f_r(x)\ge\alpha\}}dx\\ &+\int_{-\infty}^{\infty}\left(\widehat{f}_r(x) - f_r(x)\right)\tilde{s}_r(x)^2\mathbb{1}_{\{|x-\mu|\le t\sigma_r \text{ and } f_r(x)<\alpha\}}dx\\ &+\int_{-\infty}^{\infty}\left(\widehat{f}_r(x) - f_r(x)\right)\tilde{s}_r(x)^2\mathbb{1}_{\{|x-\mu|> t\sigma_r\}}dx\end{aligned}$$

By Lemma 32, the first term in the RHS above is bounded by $\frac{\mathcal{I}_r}{\sqrt{\gamma}}$. To bound the other two terms, note that

$$\Pr_{x\sim f_r}\left[|x-\mu|\le t\sigma_r \text{ and } f_r(x) < \alpha\right] \le \alpha t\sigma_r \lesssim \frac{1}{t^2}$$

and by Chebyshev's inequality,

$$\Pr_{x\sim f_r}\left[|x-\mu|> t\sigma_r\right] \le \frac{1}{t^2}$$

Also, for small constant $\beta > 0$,

$$\begin{aligned}\left(\frac{\gamma}{r^4\mathcal{I}_r^2}\right)^{1+\beta} &\le \left(\frac{\gamma\sigma_r^4}{r^4}\right)^{1+\beta} && \text{since } \mathcal{I}_r \ge \frac{1}{\sigma_r^2}\\ &\le \left(\frac{4\gamma\sigma^4}{r^4}\right)^{1+\beta} && \text{since } \sigma_r^2 = \sigma^2 + r^2 \le 2\sigma^2\\ &\le \left(\gamma^{5/12}\frac{\sigma}{r}\right)^{6+\beta} && \text{since } \gamma \ge C \text{ for large enough constant } C\text{, and } r\le\sigma\\ &\le \frac{N}{\log\frac{1}{\delta}}\end{aligned}$$

So, the conditions of Lemma 31 hold, and applying it to the second and third term in the RHS above gives the claim. ■

## Appendix E. Local Estimation

In this section, we describe our local estimation procedure, which takes a symmetrized and clipped KDE score function along with symmetrization point $\mu_1 = \mu + \varepsilon$, and produces a refined estimate $\hat{\mu}$ of the mean.

```
Algorithm 1 Local Estimation

Input Parameters:
  • $n$ samples $x_1, \dots, x_n \sim f^*$, the symmetrized and clipped KDE score function $\widehat{s}_r^{\mathrm{sym}}$, symmetrization point $\mu_1$, Fisher information estimate $\widehat{\mathcal{I}}_r$

  1. 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.

  2. Compute $\hat{\varepsilon} = \frac{1}{\widehat{\mathcal{I}}_r n} \sum_{i=1}^n \widehat{s}_r^{\mathrm{sym}}(x'_i)$. Return $\hat{\mu} = \mu_1 - \hat{\varepsilon}$.
```

**Property 7 (KDE Estimation Properties)** *Let $f_r$ be the $r$-smoothed version of symmetric distribution $f^*$ in Algorithm 1, with Fisher information $\mathcal{I}_r$. For parameters $\gamma > C$ for some sufficiently large constant $C$ and $\xi$, $\widehat{s}_r^{sym}$ satisfies that for symmetrization point $\mu_1 = \mu + \varepsilon$,*

$$\left| \mathop{\mathbb{E}}_{x \sim f_r} \left[\widehat{s}_r^{sym}(x)\right] - \varepsilon \mathcal{I}_r \right| \lesssim \frac{\varepsilon \mathcal{I}_r}{\sqrt{\gamma}} \quad \textit{and} \quad \left| \mathop{\mathbb{E}}_{x \sim f_r} \left[\widehat{s}_r^{sym}(x)^2\right] - \mathcal{I}_r \right| \lesssim \frac{\mathcal{I}_r}{\sqrt{\gamma}}$$

*and $|\widehat{s}_r^{sym}(x)| \leq \frac{2}{r}\sqrt{\log \frac{n}{\xi \log \frac{\xi}{\delta}}}$ for all $x$. Furthermore, the Fisher information estimate $\widehat{\mathcal{I}}_r$ satisfies*

$$\left| \widehat{\mathcal{I}}_r - \mathcal{I}_r \right| \lesssim \frac{\mathcal{I}_r}{\sqrt{\gamma}}$$

**Lemma 8 (Local Estimation)** *In Algorithm 1, let $f_r$ be the $r$-smoothed version of symmetric distribution $f^*$, with score function $s_r$ and Fisher information $\mathcal{I}_r$. Suppose for parameters $\gamma, \xi$, and symmetrized clipped KDE score $\widehat{s}^{sym}$ symmetrized around $\mu_1$, Property 7 is satisfied. Then, with probability $1 - \delta$, the output $\hat{\mu}$ of Algorithm 1 satisfies*

$$|\hat{\mu} - \mu| \leq \left(1 + O\left(\frac{1}{\sqrt{\gamma}}\right)\right) \sqrt{\frac{2 \log \frac{2}{\delta}}{n \mathcal{I}_r}} + O\left(\frac{\sqrt{\log \frac{n}{\xi \log \frac{\xi}{\delta}}}}{r \mathcal{I}_r} \cdot \frac{\log \frac{2}{\delta}}{n}\right) + O\left(\frac{\varepsilon}{\sqrt{\gamma}}\right)$$

**Proof** Let $\mu_1 = \mu + \varepsilon$. First, since by Property 7,

$$\frac{1}{\widehat{\mathcal{I}}_r} |\widehat{s}_r^{\mathrm{sym}}(x)| \leq \frac{2}{r \widehat{\mathcal{I}}_r} \sqrt{\log \frac{n}{\xi \log \frac{\xi}{\delta}}} \leq \left(1 + O\left(\frac{1}{\sqrt{\gamma}}\right)\right) \frac{2}{r \mathcal{I}_r} \sqrt{\log \frac{n}{\xi \log \frac{\xi}{\delta}}}$$

for all $x$, by Bernstein's inequality, the estimate $\hat{\varepsilon}$ satisfies that with probability $1 - \delta$,

$$\left| \hat{\varepsilon} - \frac{1}{\widehat{\mathcal{I}}_r} \mathop{\mathbb{E}}_{x \sim f_r} \left[\widehat{s}_r^{\mathrm{sym}}(x)\right] \right| \leq \frac{1}{\widehat{\mathcal{I}}_r} \sqrt{\mathop{\mathbb{E}}_{x \sim f_r} \left[\widehat{s}_r^{\mathrm{sym}}(x)^2\right]} \sqrt{\frac{2 \log \frac{2}{\delta}}{n}} + O\left(\frac{\sqrt{\log \frac{n}{\xi \log \frac{\xi}{\delta}}}}{r \mathcal{I}_r} \cdot \frac{\log \frac{2}{\delta}}{n}\right)$$

Since $|\widehat{\mathcal{I}}_r - \mathcal{I}_r| \lesssim \frac{\mathcal{I}_r}{\sqrt{\gamma}}$, for $\gamma > C$ for sufficiently large constant $C$, we have

$$\left| \frac{1}{\mathcal{I}_r} \mathop{\mathbb{E}}_{x \sim f_r} \left[\widehat{s}_r^{\mathrm{sym}}(x)\right] - \frac{1}{\widehat{\mathcal{I}}_r} \mathop{\mathbb{E}}_{x \sim f_r} \left[\widehat{s}_r^{\mathrm{sym}}(x)\right] \right| \lesssim \frac{1}{\mathcal{I}_r \sqrt{\gamma}} \left| \mathop{\mathbb{E}}_{x \sim f_r} \left[\widehat{s}_r^{\mathrm{sym}}(x)\right] \right| \lesssim \frac{\varepsilon}{\sqrt{\gamma}}$$

Combining this with the above and the fact that $|\widehat{\mathcal{I}}_r - \mathcal{I}_r| \lesssim \frac{\mathcal{I}_r}{\sqrt{\gamma}}$ yields

$$\begin{aligned}
&\left|\hat{\varepsilon} - \frac{1}{\mathcal{I}_r} \mathop{\mathbb{E}}_{x \sim f_r}\left[\widehat{s}_r^{\text{sym}}(x)\right]\right| \\
&\leq \left(1 + O\left(\frac{1}{\sqrt{\gamma}}\right)\right) \frac{1}{\mathcal{I}_r} \sqrt{\mathop{\mathbb{E}}_{x \sim f_r}\left[\widehat{s}_r^{\text{sym}}(x)^2\right]} \sqrt{\frac{2\log\frac{2}{\delta}}{n}} + O\left(\frac{\sqrt{\log\frac{n}{\xi \log \frac{\xi}{\delta}}}}{r\mathcal{I}_r} \cdot \frac{\log\frac{2}{\delta}}{n}\right) + O\left(\frac{\varepsilon}{\sqrt{\gamma}}\right)
\end{aligned}$$

Then, combined with Property 7 and the triangle inequality, this implies that with probability $1-\delta$,

$$|\hat{\varepsilon} - \varepsilon| \leq \left(1 + O\left(\frac{1}{\sqrt{\gamma}}\right)\right) \sqrt{\frac{2\log\frac{2}{\delta}}{n\mathcal{I}_r}} + O\left(\frac{\sqrt{\log\frac{n}{\xi \log \frac{\xi}{\delta}}}}{r\mathcal{I}_r} \cdot \frac{\log\frac{2}{\delta}}{n}\right) + O\left(\frac{\varepsilon}{\sqrt{\gamma}}\right)$$

So, since $\mu_1 = \mu + \varepsilon$ and $\hat{\mu} = \mu + \hat{\varepsilon}$, we have the claim. ■

## Appendix F. Global Estimation

In this section, we describe our global estimation procedure and show that it provides a good estimate of the mean. It uses a small number of samples to compute an initial estimate $\mu_1$ of $\mu$, and uses another small set of samples to compute the symmetrized, clipped KDE score function $\widehat{s}_r^{\text{sym}}$ symmetrized around $\mu_1$. It then uses our local estimation procedure to produce the final estimate $\hat{\mu}$.

**Algorithm 2** Global Estimation

**Input parameters:**

- Failure probability $\delta$, Samples $x_1, \dots, x_n \sim f^*$, smoothing parameter $r$, approximation parameter $\xi > 0$.

1. First, use the first $n/\xi$ samples to compute an initial estimate $\mu_1$ of the mean $\mu$ by using the Median-of-pairwise-means estimator in Lemma 9.
2. Use the next $n/\xi$ samples to compute the kernel density estimate $\widehat{f_r}$ of $f_r$ (as defined in (8)), along with the associated symmeterized, clipped KDE score $\widehat{s}_r^{\text{sym}}$ (as defined in (12)), clipped at $\frac{2}{r}\sqrt{\log\frac{n}{\xi \log\frac{\xi}{\delta}}}$ and symmetrized around the initial estimate $\mu_1$. Compute the Fisher information estimate $\widehat{\mathcal{I}}_r = \mathbb{E}_{x \sim \widehat{f}_r}\left[\widehat{s}_r^{\text{sym}}(x)^2\right]$.
3. Run Algorithm 1 using the remaining $n - \frac{2n}{\xi}$ samples, and return the final estimate $\hat{\mu}$.

**Lemma 10 (Global Estimation)** *Let $\xi > C$ for large enough constant $C > 3$ be a parameter, and suppose $\xi \leq \gamma \leq \left(\frac{n}{\xi \log\frac{1}{\delta}}\right)^{2/5-\alpha}$ for constant $\alpha > 0$. For any $r \leq \sigma$ and $\frac{n}{\log\frac{1}{\delta}} \geq \xi\left(\frac{\gamma^{5/12}\sigma}{r}\right)^{6+\alpha}$,* *with probability* $1-\delta$*, Algorithm* 2 *outputs an estimate* $\hat{\mu}$ *with*

$$|\hat{\mu}-\mu|\leq \left(1+O\left(\frac{1}{\sqrt{\gamma}}\right)+O\left(\frac{1}{\xi}\right)\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n\mathcal{I}_r}}+O\left(\frac{\sigma}{r\sqrt{\gamma}}\sqrt{\frac{\xi\log\frac{\xi}{\delta}}{n\mathcal{I}_r}}\right)$$

**Proof** Let $\varepsilon=\mu_1-\mu$ where $\mu_1$ is our median of means estimate in Step 1. First, note that by Lemma 9, $\varepsilon$ satisfies that with probability $1-\delta/\xi$,

$$|\varepsilon|\lesssim \sigma\cdot\sqrt{\frac{\xi\log\frac{\xi}{\delta}}{n}}$$

We condition on Step 1 succeeding so that the above holds. To obtain bounds on the expectation and variance of $\widehat{s}^{\text{sym}}$, we will now check that the following conditions required to invoke Lemma 4 and Lemma 5 hold:

- $|\varepsilon|\leq r/60$
- $|\varepsilon|\leq r^2\sqrt{\frac{\mathcal{I}_r}{\gamma}}$

Note that $\frac{\log\frac{1}{\delta}}{n}\leq 1$ and $\gamma\geq 1$. Note also that $\xi\leq\gamma$. So, by our setting of $n$,

$$r\geq \sigma\gamma^{5/12}\xi^{1/6}\left(\frac{\log\frac{1}{\delta}}{n}\right)^{1/6}\gg O\left(\sigma\cdot\sqrt{\frac{\xi\log\frac{\xi}{\delta}}{n}}\right)\geq|\varepsilon|$$

so that $|\varepsilon|\leq r/60$. Similarly

$$\begin{aligned}
r^2\sqrt{\frac{\mathcal{I}_r}{\gamma}} &\geq \sigma^2\gamma^{5/6}\xi^{1/3}\left(\frac{\log\frac{1}{\delta}}{n}\right)^{1/3}\sqrt{\frac{\mathcal{I}_r}{\gamma}}\\
&\geq \frac{\sigma^2\gamma^{5/6}\xi^{1/3}}{\sqrt{\sigma^2+r^2}}\left(\frac{\log\frac{1}{\delta}}{n}\right)^{1/3} \qquad \text{since } \mathcal{I}_r\geq\frac{1}{\sigma^2+r^2}\\
&\geq \frac{\sigma}{2}\gamma^{5/6}\xi^{1/3}\left(\frac{\log\frac{1}{\delta}}{n}\right)^{1/3} \qquad \text{since } r\leq\sigma\\
&\gg O\left(\sigma\cdot\sqrt{\frac{\xi\log\frac{\xi}{\delta}}{n}}\right) \qquad \text{since } \frac{\log\frac{1}{\delta}}{n}\leq 1\text{, and } \xi\geq\gamma\geq 1\\
&\geq|\varepsilon|
\end{aligned}$$

Also, our choice of $n$ implies that

$$\frac{n}{\xi}\geq\left(\frac{\gamma^{5/12}\sigma}{r}\right)^{6+\alpha/2}\log\frac{\xi}{\delta}$$

So, we can invoke Lemma 4 and Lemma 5 to obtain the following bounds on the mean and second moment of $\widehat{s}^{\text{sym}}$ in Step 2, which hold with probability $1 - \frac{\delta}{\xi}$.

$$\left| \mathop{\mathbb{E}}_{x \sim f_r} \left[\widehat{s}_r^{\text{sym}}(x)\right] - \varepsilon \mathcal{I}_r \right| \lesssim \frac{\varepsilon \mathcal{I}_r}{\sqrt{\gamma}} \quad \text{and} \quad \left| \mathop{\mathbb{E}}_{x \sim f_r} \left[\widehat{s}_r^{\text{sym}}(x)^2\right] - \mathcal{I}_r \right| \lesssim \frac{\mathcal{I}_r}{\sqrt{\gamma}}$$

Also, for $\widehat{s}_r^{\text{sym}}$ that satisfies the above, since $\widehat{s}_r^{\text{sym}}$ is clipped at $\frac{2}{r}\sqrt{\log \frac{n}{\xi \log \frac{\xi}{\delta}}}$, and $\frac{n}{\xi} \geq \left(\frac{\gamma^{5/12}\sigma}{r}\right)^{6+\alpha/2} \log \frac{\xi}{\delta}$, the assumptions of Lemma 6 are satisfied for $N = \frac{n}{\xi}$ and failure probability $\frac{\delta}{\xi}$. So conditioned on the success of Step 1, by a union bound, with probability $1 - \frac{2\delta}{\xi}$, Property 7 is satisfied, and simultaneously, by Lemma 6, the Fisher information estimate $\widehat{\mathcal{I}}_r$ in Step 2 satisfies

$$\left| \widehat{\mathcal{I}}_r - \mathcal{I}_r \right| \lesssim \frac{\mathcal{I}_r}{\sqrt{\gamma}}$$

Conditioned on the above, since Property 7 is satisfied, by Lemma 8, our final estimate $\hat{\mu}$ satisfies that for $n' = n\left(1 - \frac{2}{\xi}\right)$ and $\delta' = \delta\left(1 - \frac{3}{\xi}\right)$, with probability $1 - \delta'$,

$$\begin{aligned} |\hat{\mu} - \mu| &\leq \left(1 + O\left(\frac{1}{\sqrt{\gamma}}\right)\right)\sqrt{\frac{2\log \frac{2}{\delta'}}{n'\mathcal{I}_r}} + O\left(\frac{\sqrt{\log \frac{n'}{\xi \log \frac{\xi}{\delta'}}}}{r\mathcal{I}_r} \cdot \frac{\log \frac{2}{\delta'}}{n'}\right) + O\left(\frac{\varepsilon}{\sqrt{\gamma}}\right) \\ &\leq \left(1 + O\left(\frac{1}{\sqrt{\gamma}}\right) + O\left(\frac{1}{\xi}\right)\right)\sqrt{\frac{2\log \frac{2}{\delta}}{n\mathcal{I}_r}} + O\left(\frac{\sqrt{\log \frac{n}{\xi \log \frac{\xi}{\delta}}}}{r\mathcal{I}_r} \cdot \frac{\log \frac{2}{\delta}}{n}\right) + O\left(\frac{\varepsilon}{\sqrt{\gamma}}\right) \end{aligned}$$

Now, we bound $\frac{\varepsilon}{\sqrt{\gamma}}$. We have,

$$\begin{aligned} \frac{\varepsilon}{\sqrt{\gamma}} &\lesssim \frac{\sigma\sqrt{\mathcal{I}_r}}{\sqrt{\gamma}}\sqrt{\frac{\xi \log \frac{\xi}{\delta}}{n\mathcal{I}_r}} \\ &\leq \frac{\sigma}{r\sqrt{\gamma}}\sqrt{\frac{\xi \log \frac{\xi}{\delta}}{n\mathcal{I}_r}} \quad \text{since } \mathcal{I}_r \leq \frac{1}{r^2} \end{aligned}$$

Finally, we bound $\frac{\sqrt{\log \frac{n}{\xi \log \frac{\xi}{\delta}}} \log \frac{1}{\delta}}{nr\mathcal{I}_r}$. First note that

$$\begin{aligned} r^2 &\geq \sigma^2 \gamma^{5/6} \xi^{1/3} \left(\frac{\log \frac{1}{\delta}}{n}\right)^{1/3} \\ &\gtrsim \frac{1}{\mathcal{I}_r} \gamma^{5/6} \xi^{1/3} \left(\frac{\log \frac{1}{\delta}}{n}\right)^{1/3} \quad \text{since } \mathcal{I}_r \geq \frac{1}{\sigma^2 + r^2} \gtrsim \frac{1}{\sigma^2} \text{ since } r \leq \sigma \end{aligned}$$

So, we have

$$
\begin{aligned}
\frac{\sqrt{\log \frac{n}{\xi \log \frac{\xi}{\delta}}} \log \frac{1}{\delta}}{r\mathcal{I}_r n} &\le \frac{1}{r\mathcal{I}_r}\left(\frac{\log \frac{1}{\delta}}{n}\right)^{1-\alpha} \quad \text{since } \frac{n}{\log \frac{1}{\delta}} \ge 1 \\
&\lesssim \frac{1}{\sqrt{\mathcal{I}_r}\gamma^{5/12}\xi^{1/3}}\left(\frac{\log \frac{1}{\delta}}{n}\right)^{5/6-\alpha} \\
&= \frac{\gamma^{1/12}}{\sqrt{\mathcal{I}_r}\gamma\xi^{1/3}}\left(\frac{\log \frac{1}{\delta}}{n}\right)^{5/6-\alpha} \\
&\lesssim \frac{1}{\sqrt{\mathcal{I}_r}\gamma\xi^{1/3}}\left(\frac{\log \frac{1}{\delta}}{n}\right)^{4/5-\alpha} \quad \text{since } \gamma \le \left(\frac{n}{\xi \log \frac{1}{\delta}}\right)^{2/5-\alpha} \le \left(\frac{n}{\log \frac{1}{\delta}}\right)^{2/5} \\
&\lesssim \frac{1}{\sqrt{\gamma}}\sqrt{\frac{\log \frac{1}{\delta}}{n\mathcal{I}_r}}
\end{aligned}
$$

So we have shown that

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

Thus, by a union bound, with probability $1-\delta$ in total, the claim follows. ■

**Theorem 1** *Let $\eta = (\frac{\log \frac{1}{\delta}}{n})^{\frac{1}{13}} < 1$, and let $\log \frac{1}{\delta} \le n/C$ for sufficiently large constant $C > 1$. Let $f^*$ be an arbitrary symmetric distribution with variance $\sigma^2$ and mean $\mu$. For $\eta\sigma \le r \le \sigma$, we have*

$$
|\widehat{\mu} - \mu| \le (1+\eta)\sqrt{\frac{2\log \frac{2}{\delta}}{n\mathcal{I}_r}}
$$

*with probability $1-\delta$.*

**Proof** Set $\xi = \frac{1}{\eta}$, and let $\gamma = \frac{1}{\eta^2}$. First we check the conditions of Lemma 10 that

- $\xi > C_1$ for sufficiently large constant $C_1$
- $\xi \le \gamma \le \left(\frac{n}{\xi \log \frac{1}{\delta}}\right)^{2/5-\alpha}$ for constant $\alpha > 0$
- $\frac{n}{\log \frac{1}{\delta}} \ge \xi \left(\frac{\gamma^{5/12}\sigma}{r}\right)^{6+\alpha}$ for constant $\alpha > 0$

We have,

$$
\xi = \frac{1}{\eta} = \left(\frac{n}{\log \frac{1}{\delta}}\right)^{\frac{1}{13}-\beta} \ge C^{\frac{1}{13}-\beta} \ge C_1
$$

for large enough constant $C$, by assumption. We also have,

$$\gamma = \frac{1}{\eta^2} \geq \frac{1}{\eta} = \xi$$

and

$$\gamma = \frac{1}{\eta^2} = \left(\frac{n}{\log\frac{1}{\delta}}\right)^{\frac{2}{13}-2\beta} = \eta^{\frac{2}{5}-\beta}\left(\frac{n}{\log\frac{1}{\delta}}\right)^{\frac{8}{65}-\Omega(\beta)} \leq \left(\frac{n}{\xi\log\frac{1}{\delta}}\right)^{2/5-\alpha}$$

Finally, we have

$$\begin{aligned}
\xi\left(\frac{\gamma^{5/12}\sigma}{r}\right)^{6+\alpha} &\leq \frac{1}{\eta}\left(\frac{1}{\eta^{11/6}}\right)^{6+\alpha} \quad \text{since } \xi = \frac{1}{\eta}, \gamma = \frac{1}{\eta^2}, \text{ and } r \geq \eta\sigma \\
&= \frac{1}{\eta^{12+\frac{11}{6}\alpha}} \\
&\leq \frac{n}{\log\frac{1}{\delta}} \quad \text{since } \eta = \left(\frac{\log\frac{1}{\delta}}{n}\right)^{1/13}
\end{aligned}$$

So we have verified the above conditions, which, along with the fact that $r \leq \sigma$, allow us to invoke Lemma 10. The Lemma gives that with probability $1-\delta$,

$$|\widehat{\mu}-\mu| \leq \left(1 + O\left(\frac{1}{\sqrt{\gamma}}\right) + O\left(\frac{1}{\xi}\right)\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n\mathcal{I}_r}} + \frac{\sigma}{r\sqrt{\gamma}}\sqrt{\frac{\xi\log\frac{\xi}{\delta}}{n\mathcal{I}_r}}$$

We bound the last term above.

$$\begin{aligned}
\frac{\sigma}{r\sqrt{\gamma}}\sqrt{\frac{\xi\log\frac{\xi}{\delta}}{n\mathcal{I}_r}} &\leq \frac{\eta}{\sqrt{\gamma}}\sqrt{\frac{\xi\log\frac{\xi}{\delta}}{n\mathcal{I}_r}} \quad \text{since } r \geq \eta\sigma \\
&= \eta^2\sqrt{\frac{\xi\log\frac{\xi}{\delta}}{n\mathcal{I}_r}} \quad \text{substituting } \gamma \\
&\leq \eta\sqrt{\frac{\log\frac{1}{\delta}}{n}} \quad \text{since } \xi = \frac{1}{\eta}
\end{aligned}$$

Along with the above and the setting of $\gamma$ and $\xi$, we have

$$|\widehat{\mu}-\mu| \leq (1 + O(\eta))\sqrt{\frac{2\log\frac{2}{\delta}}{n\mathcal{I}_r}}$$

Reparametrizing $\eta$ gives the claim. ■

# Appendix G. Properties of $r$-smoothed distributions

## G.1. Score bound in terms of density

The next lemma shows that $s_r$ is bounded in terms of $f_r$ and $r$.

**Lemma 33** *Let $f^*$ be an arbitrary distribution, and let $f_r$ be the $r$-smoothed version of $f^*$. Let $s_r$ be the score function of $f_r$. We have*

$$|s_r(x)| \leq \frac{1}{r}\sqrt{2\log\frac{1}{\sqrt{2\pi}rf_r(x)}}$$

**Proof** Let $w_r$ be the pdf of $\mathcal{N}(0, r^2)$. By definition of $r$-smoothing, we have that when $X \sim f_r$, $X = Y + Z_r$ where $Y \sim f^*$ and $Z_r \sim w_r$ for independent $Y, Z_r$. So,

$$f_r(x) = \Pr_{X\sim f_r}[X = x] = \int_{-\infty}^{\infty} \Pr_{Y\sim f^*}[Y = y] \Pr_{Z_r\sim w_r}[Z = Y - x]dy = \mathop{\mathbb{E}}_{Y\sim f^*}[w_r(x - Y)]$$

So, we have

$$\Pr[X = x|Y = y] = \mathbb{E}[w_r(x - Y)|Y = y] = w_r(x - y)$$

Now, since $w_r(x) = \frac{1}{\sqrt{2\pi}r}e^{-\frac{-x^2}{2r^2}}$

$$(x - Y) = r\sqrt{2\log\frac{1}{\sqrt{2\pi}r \cdot w_r(x - Y)}}$$

So, by Lemma 36,

$$\begin{aligned}
s_r(x) &= \mathbb{E}\left[\frac{Z_r}{r^2}\middle|X = x\right] \\
&= \frac{1}{r^2}\mathbb{E}\left[x - Y|X = x\right] \quad \text{since } X = Y + Z_r \\
&= \frac{1}{r}\mathbb{E}\left[\sqrt{2\log\frac{1}{\sqrt{2\pi}r \cdot w_r(x - Y)}}\middle|X = x\right] \\
&= \frac{1}{r}\int_{-\infty}^{\infty}\sqrt{2\log\frac{1}{\sqrt{2\pi}r \cdot w_r(x - y)}}\Pr[Y = y|X = x]dy \\
&= \frac{1}{r}\int_{-\infty}^{\infty}\sqrt{2\log\frac{1}{\sqrt{2\pi}r \cdot w_r(x - y)}}\frac{\Pr[Y = y]\Pr[X = x|Y = y]}{\Pr[X = x]}dy \quad \text{(by Bayes' Theorem)} \\
&= \frac{1}{r}\int_{-\infty}^{\infty}\frac{w_r(x - y)}{f_r(x)}\sqrt{2\log\frac{1}{\sqrt{2\pi}r \cdot w_r(x - y)}}\Pr[Y = y]dy \\
&= \frac{1}{r}\mathbb{E}\left[\frac{w_r(x - Y)}{f_r(x)}\sqrt{2\log\frac{1}{\sqrt{2\pi}r \cdot w_r(x - Y)}}\right]
\end{aligned}$$

Now, $g(z) = z\sqrt{2\log\frac{1}{\sqrt{2\pi}r\cdot z}}$ is concave on $[0,1]$. So, by Jensen's inequality,

$$\begin{aligned} s_r(x) &\le \frac{\mathbb{E}_{Y\sim f^*}[w_r(x-Y)]}{rf_r(x)}\sqrt{2\log\frac{1}{\sqrt{2\pi}r\cdot\mathbb{E}[w_r(x-Y)]}} \\ &= \frac{1}{r}\sqrt{2\log\frac{1}{\sqrt{2\pi}r\cdot f_r(x)}} \quad \text{since } f_r(x)=\mathbb{E}[w_r(x-Y)] \end{aligned}$$

■

## G.2. $r$-smoothed score is $O(1/r)$-subgaussian

The next two lemmas together show that the score function of an $r$-smoothed distribution is $O\left(\frac{1}{r}\right)$-subgaussian.

**Lemma 34** *Consider the distribution $f_r$ which is the $r$-smoothed version of distribution $f$. That is, $f_r$ has density $f_r(x) = \mathbb{E}_{y\leftarrow h}[\frac{1}{\sqrt{2\pi r^2}}e^{-\frac{(x-y)^2}{2r^2}}]$. Then, with probability at least $1-(1+\tau)\delta$, we sample a point $x\sim f_r$ such that*

$$\begin{aligned} &\mathop{\mathbb{E}}_{y\sim f}\left[\mathbb{1}\left[(x-y)^2 > 2r^2\log\frac{1}{\delta}\right]\frac{|x-y|}{r^2}\frac{1}{\sqrt{2\pi r^2}}e^{-\frac{(x-y)^2}{2r^2}}\right] \\ &\le \frac{1}{\tau}\mathop{\mathbb{E}}_{y\sim f}\left[\mathbb{1}\left[(x-y)^2 \le 2r^2\log\frac{1}{\delta}\right]\frac{|x-y|}{r^2}\frac{1}{\sqrt{2\pi r^2}}e^{-\frac{(x-y)^2}{2r^2}}\right] \end{aligned}$$

**Proof** Observe that, at any point $x$ violating the above inequality, we have

$$\begin{aligned} f_r(x) &= \mathop{\mathbb{E}}_{y\leftarrow f}\left[\mathbb{1}\left[(x-y)^2 \le 2r^2\log\frac{1}{\delta}\right]\frac{|x-y|}{r^2}\frac{1}{\sqrt{2\pi r^2}}e^{-\frac{(x-y)^2}{2r^2}}\right] \\ &\quad + \mathop{\mathbb{E}}_{y\leftarrow f}\left[\mathbb{1}\left[(x-y)^2 > 2r^2\log\frac{1}{\delta}\right]\frac{|x-y|}{r^2}\frac{1}{\sqrt{2\pi r^2}}e^{-\frac{(x-y)^2}{2r^2}}\right] \\ &\le (1+\tau)\mathop{\mathbb{E}}_{y\leftarrow f}\left[\mathbb{1}\left[(x-y)^2 > 2r^2\log\frac{1}{\delta}\right]\frac{|x-y|}{r^2}\frac{1}{\sqrt{2\pi r^2}}e^{-\frac{(x-y)^2}{2r^2}}\right] \end{aligned}$$

We wish to bound the probability of sampling $x$ violating the lemma inequality, which is bounded by the integral of the above right hand side. We can further bound it using the following:

$$\begin{aligned} &\int\mathop{\mathbb{E}}_{y\leftarrow f}\left[\mathbb{1}\left[(x-y)^2 > 2r^2\log\frac{1}{\delta}\right]\frac{|x-y|}{r^2}\frac{1}{\sqrt{2\pi r^2}}e^{-\frac{(x-y)^2}{2r^2}}\right]\mathrm{d}x \\ &= \mathop{\mathbb{E}}_{y\leftarrow f}\left[\int\mathbb{1}\left[(x-y)^2 > 2r^2\log\frac{1}{\delta}\right]\frac{|x-y|}{r^2}\frac{1}{\sqrt{2\pi r^2}}e^{-\frac{(x-y)^2}{2r^2}}\mathrm{d}x\right] \\ &\le \mathop{\mathbb{E}}_{y\leftarrow f}[\delta] = \delta \end{aligned}$$

Thus the probability of sampling a point $x$ violating the lemma inequality is upper bounded by the integral of $f_r(x)$ over those points, which is in turn upper bounded by $(1+\tau)\delta$. ■

**Lemma 35 (Score is $O(1/r)$-subgaussian)** *Let $s_r$ be the score function of an $r$-smoothed distribution $f_r$. We have that for $x \sim f_r$, with probability $1-\delta$, $|s_r(x)| \lesssim \frac{1}{r}\sqrt{\log \frac{2}{\delta}}$.*

**Proof**

By Lemma 34 using $\tau = 1$, with probability $1 - 2\delta$ over sampling a single point $x \leftarrow f_r$, the point $x$ satisfies

$$\begin{aligned}
&\mathop{\mathbb{E}}_{y \leftarrow f}\left[\mathbb{1}\left[(x-y)^2 > 2r^2 \log \frac{1}{\delta}\right] \frac{|x-y|}{r^2} \frac{1}{\sqrt{2\pi r^2}} e^{-\frac{(x-y)^2}{2r^2}}\right] \\
\leq &\mathop{\mathbb{E}}_{y \leftarrow f}\left[\mathbb{1}\left[(x-y)^2 \leq 2r^2 \log \frac{1}{\delta}\right] \frac{|x-y|}{r^2} \frac{1}{\sqrt{2\pi r^2}} e^{-\frac{(x-y)^2}{2r^2}}\right]
\end{aligned}$$

And so,

$$\begin{aligned}
s_r(x) = \frac{f_r'(x)}{f_r(x)} &= \frac{\mathbb{E}_{y \leftarrow f}\left[\frac{y-x}{r^2} \frac{1}{\sqrt{2\pi r^2}} e^{-\frac{(x-y)^2}{2r^2}}\right]}{f_r(x)} \\
&\leq \frac{\mathbb{E}_{y \leftarrow f}\left[\frac{|x-y|}{r^2} \frac{1}{\sqrt{2\pi r^2}} e^{-\frac{(x-y)^2}{2r^2}}\right]}{f_r(x)} \\
&= \frac{\mathbb{E}_{y \leftarrow f}\left[\mathbb{1}\left[(x-y)^2 \leq 2r^2 \log \frac{1}{\delta}\right] \frac{|x-y|}{r^2} \frac{1}{\sqrt{2\pi r^2}} e^{-\frac{(x-y)^2}{2r^2}}\right]}{f_r(x)} \\
&+ \frac{\mathbb{E}_{y \leftarrow f}\left[\mathbb{1}\left[(x-y)^2 > 2r^2 \log \frac{1}{\delta}\right] \frac{|x-y|}{r^2} \frac{1}{\sqrt{2\pi r^2}} e^{-\frac{(x-y)^2}{2r^2}}\right]}{f_r(x)} \\
&\leq 2 \cdot \frac{\mathbb{E}_{y \leftarrow f}\left[\mathbb{1}\left[(x-y)^2 \leq 2r^2 \log \frac{1}{\delta}\right] \frac{|x-y|}{r^2} \frac{1}{\sqrt{2\pi r^2}} e^{-\frac{(x-y)^2}{2r^2}}\right]}{f_r(x)} \\
&\leq \frac{2\sqrt{2}}{r}\sqrt{\log \frac{1}{\delta}} \frac{\mathbb{E}_{y \leftarrow f}\left[\mathbb{1}\left[(x-y)^2 \leq 2r^2 \log \frac{1}{\delta}\right] \frac{1}{\sqrt{2\pi r^2}} e^{-\frac{(x-y)^2}{2r^2}}\right]}{f_r(x)} \\
&\leq \frac{2\sqrt{2}}{r}\sqrt{\log \frac{1}{\delta}}
\end{aligned}$$

Reparameterizing from $2\delta$ to $\delta$ gives the lemma result. ■

### G.3. Lemmas from Gupta et al. (2022)

Here, we recall some of the properties of $r$-smoothed distributions shown in Gupta et al. (2022)

**Lemma 36 (From Gupta et al. (2022))** *Let $s_r$ be the score function of $r$-smoothed distribution $f_r$. Then,*

$$\frac{f_r(x+\varepsilon)}{f_r(x)} = \mathop{\mathbb{E}}_{Z_r|x}\left[e^{\frac{2\varepsilon Z_r - \varepsilon^2}{2r^2}}\right] \quad \textit{and in particular} \quad s_r(x) = \frac{1}{r^2} \mathop{\mathbb{E}}_{Z_r|x}[Z_r]$$

*and*

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

**Lemma 37 (From Gupta et al. (2022))** *Let $s_r$ be the score function of an $r$-smoothed distribution $f_r$ with Fisher information $\mathcal{I}_r$. Then for any $|\varepsilon|\leq r/2$,*

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

**Lemma 38 (From Gupta et al. (2022))** *Let $s_r$ be the score function of an $r$-smoothed distribution $f_r$ with Fisher information $\mathcal{I}_r$. Then, for any $|\varepsilon|\leq r/2$,*

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

**Lemma 39 (From Gupta et al. (2022))** *Let $\mathcal{I}_r$ be the Fisher information of an $r$-smoothed distribution $f_r$. Then $\mathcal{I}_r \leq 1/r^2$.*

**Lemma 40 (From Gupta et al. (2022))** *Let $f^*$ be an arbitrary distribution, and let $f_r$ 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 $|\varepsilon|\leq r/2$,*

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

## Appendix H. Median of Pairwise Means Estimator

Using results in Minton and Price (2014), we show that the median of pairwise means is a good estimator for symmetric random variables. In particular, it matches the convergence of the median-of-means estimator for all $(\varepsilon, \delta)$ without needing to specify $\varepsilon$ and $\delta$.

**Lemma 9 (Median of pairwise means estimator)** *Let $X_1, X_2, \dots, X_n$ be drawn from a symmetric distribution with mean $\mu$ and variance $\sigma^2$. For every constant $C_1 > 0$ there exists a constant $C_2$ such that $\widehat{\mu} := \mathrm{median}_{i\in[n/2]}\frac{X_{2i-1}+X_{2i}}{2}$ satisfies*

$$|\widehat{\mu} - \mu|\leq C_2\sigma\cdot\sqrt{\frac{\log\frac{2}{\delta}}{n}}$$

*with probability $1-\delta$, for all $\delta$ with $\log\frac{1}{\delta}\leq C_1 n$.*

**Proof** Let $Y_i = \frac{1}{2}(X_{2i-1} + X_{2i})$ for $i \in [n/2]$. Let $p$ be the pdf of $X - \mu$, and $q$ be the pdf of $Y - \mu$. Since $X - \mu$ is symmetric about $0$, the Fourier transform $\widehat{p}$ of $p$ is real-valued. By the Fourier convolution theorem, $q$ has nonnegative Fourier transform. Then by Lemma 3.1 of Minton and Price (2014), for any $\varepsilon < 1$,

$$\Pr[|Y_i|< \varepsilon\sigma] \geq C_3\varepsilon$$

for a universal constant $C_3$. Then it is easy to show (e.g., Lemma 3.3 of Minton and Price (2014)):

$$\Pr[|\widehat{\mu} - \mu|> \varepsilon\sigma] \leq 2e^{-\frac{C_3^2}{4}\varepsilon^2 n}.$$

Setting $\varepsilon = \frac{2}{C_3}\sqrt{\frac{\log\frac{2}{\delta}}{n}}$ gives the result, as long as $n > \frac{4}{C_3^2}\log\frac{2}{\delta}$ so this $\varepsilon < 1$.

There's a remaining regime of $\frac{C_3^2}{4} \leq \frac{\log\frac{1}{\delta}}{n} \leq C_1$ for which we need to prove a $\Theta(\sigma)$ bound on $|\widehat{\mu} - \mu|$. Note that $Y_i$ has variance $\sigma^2/2$, so for any $a > 0$, with probability $1 - a$ we have $|Y_i - \mu|\leq \frac{\sigma}{\sqrt{2a}}$. Let $E_i$ denote the event that $|Y_i - \mu|> \frac{\sigma}{\sqrt{2a}}$. Then

$$\Pr[|\widehat{\mu} - \mu|> \frac{\sigma}{\sqrt{2a}}] \leq \Pr[\sum_{i=1}^{n/2} E_i \geq \frac{n}{4}] \leq \binom{n/2}{n/4} a^{n/4} \leq (4a)^{n/4}.$$

which is $\delta$ for $a = \frac{1}{4}e^{-\frac{4}{n}\log\frac{1}{\delta}} \geq \frac{1}{4}e^{-4C_1}$. Thus with probability $1 - \delta$, $|\widehat{\mu} - \mu|\leq \sqrt{2}e^{2C_1}\sigma \eqsim \sigma$. ■