Choosing the p in L_p loss: adaptive rates for symmetric mean estimation

Yu-Chun Kao YK495@STAT.RUTGERS.EDU
Department of Statistics, Rutgers University

Min Xu MX76@STAT.RUTGERS.EDU
Department of Statistics, Rutgers University

Cun-Hui Zhang CZHANG@STAT.RUTGERS.EDU
Department of Statistics, Rutgers University

Editors: Shipra Agrawal and Aaron Roth

Abstract

When we have a univariate distribution that is symmetric around its mean, the mean can be estimated with a rate (sample complexity) much faster than $O(1/\sqrt{n})$ in many cases. For example, given univariate random variables $Y_1, \ldots, Y_n$ distributed uniformly on $[\theta_0 - c, \theta_0 + c]$, the sample midrange $\frac{Y_{(n)}+Y_{(1)}}{2}$ maximizes likelihood and has expected error $\mathbb{E}\left|\theta_0 - \frac{Y_{(n)}+Y_{(1)}}{2}\right| \leq 2c/n$, which is optimal and much lower than the error rate $O(1/\sqrt{n})$ of the sample mean. What the optimal rate is depends on the distribution and it is generally attained by the maximum likelihood estimator (MLE). However, MLE requires exact knowledge of the underlying distribution; if the underlying distribution is *unknown*, it is an open question whether an estimator can adapt to the optimal rate. In this paper, we propose an estimator of the symmetric mean θ_0 with the following properties: it requires no knowledge of the underlying distribution; it has a rate no worse than $1/\sqrt{n}$ in all cases (assuming a finite second moment) and, when the underlying distribution is compactly supported, our estimator can attain a rate of $n^{-\frac{1}{\alpha}}$ up to polylog factors, where the rate parameter α can take on any value in $(0, 2]$ and depends on the moments of the underlying distribution. Our estimator is formed by minimizing the L_γ-loss with respect to the data, for a power $\gamma \geq 2$ chosen in a data-driven way – by minimizing a criterion motivated by the asymptotic variance. Our approach can be directly applied to the regression setting where θ_0 is a function of observed features and motivates the use of L_γ loss function with a data-driven γ in certain settings.

Keywords: Mean estimation, adaptive rate, irregular estimation

1. Introduction

Symmetric mean estimation is a fundamental problem in statistics. Given IID random variables $Y_1, \ldots, Y_n \overset{d}{\sim} P(\cdot - \theta_0)$, where P is an unknown distribution symmetric around 0, the question is how to best estimate θ_0. The existing theory focuses on the case where P has a density p that is *smooth* so that the Fisher information $\mathcal{I} = \int (p')^2/p$ for θ_0 is finite, which implies that the optimal rate is $O(1/\sqrt{n})$. For example, asymptotic theory from Stein (1956), Stone (1975), Beran (1978), and many others construct estimators $\widehat{\theta}$ that are asymptotically normal in the sense that $\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} N(0, V(p))$, where the asymptotic variance $V(p) = 1/\mathcal{I}$ is the inverse of Fisher information and is known to be optimal by generalizations of the Cramer-Rao lower bound, see, e.g., the Hayek-Le Cam convolution theorem (Van der Vaart, 2000). The estimator $\widehat{\theta}$ is typically constructed by first non-parametrically estimating the density p. Recently, a line of work (Gupta

et al., 2023, 2022) proved a finite sample bound for similar estimators (we discuss their result more in Remark 12).

In this paper, we are interested in settings where the density p is *nonsmooth* (also known as *irregular*) or where P does not have a density. In these settings, Fisher information can be infinite and the optimal rate for estimating θ_0 can be much faster than $O(1/\sqrt{n})$. For example, given random variables $Y_1, \dots, Y_n \overset{d}{\sim} \text{Uniform}(\theta_0 - c, \theta_0 + c)$ for some $c > 0$, the optimal estimator for the center θ_0 is not the usual sample mean $\bar{Y}$ but rather the sample midrange $Y_{\text{mid}} = \frac{Y_{(n)}+Y_{(1)}}{2}$. Indeed, we have

$$\mathbb{E}\left|\frac{Y_{(n)} + Y_{(1)}}{2} - \theta_0\right| \le \mathbb{E}\left|\frac{Y_{(n)} - \theta_0 - c}{2}\right| + \mathbb{E}\left|\frac{Y_{(1)} - \theta_0 + c}{2}\right| = \frac{2c}{n+1},$$

which is far smaller than the $1/\sqrt{n}$ error of the sample mean; a two points argument in Le Cam (1973) shows that the $1/n$ rate is optimal in this case. However, sample midrange is a poor choice when $Y_1, \dots, Y_n \overset{d}{\sim} N(\theta_0, 1)$, where we have that $\mathbb{E}|Y_{\text{mid}} - \theta_0|$ is of order $1/\sqrt{\log n}$. These observations naturally motivate the following question: let P be a univariate distribution symmetric around 0, possibly nonsmooth, and suppose $Y_1, \dots, Y_n$ has the distribution $P(\cdot - \theta_0)$ which is the location shift of P, can we construct an estimator of the location θ_0 whose rate of convergence adapts to the unknown P? The existing theory does not have an answer to this question because P is not assumed to have a smooth density p. Even the problem of choosing between only the sample mean $\bar{Y}$ and the sample midrange Y_{mid} is nontrivial, as we show in this paper that the tried and true method of cross-validation fails in this setting (see Remark 2 for a detailed discussion).

If P has a density p which is known, the optimal rate in estimating the location θ_0 is governed by the speed with which the function $\Delta \mapsto H\big(p(\cdot), p(\cdot - \Delta)\big)$ decreases as Δ goes to zero, where $H(p, q) := \big\{\int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx\big\}^{1/2}$ is the Hellinger distance. To be precise, for any estimator $\widehat{\theta}$, we have

$$\liminf_{n\to\infty} \sup_{\theta_0} \mathbb{E}_{\theta_0}\big\{\sqrt{n} H\big(p(\cdot - \theta_0), p(\cdot - \widehat{\theta})\big)\big\} > 0,$$

where the supremum can be taken in a local ball of shrinking radius around any point in $\mathbb{R}$; see, for example, Theorem 6.1 in Chapter I of Ibragimov and Has' Minskii (2013) for an exact statement. Le Cam (1973) also showed that the oracle MLE attains this convergence rate under mild conditions. Therefore, if $H^2\big(p(\cdot), p(\cdot - \Delta)\big)$ is of order $|\Delta|^\alpha$ for some $\alpha > 0$, then the optimal rate of the error $\mathbb{E}_{\theta_0}|\widehat{\theta} - \theta_0|$ is $n^{-\frac{1}{\alpha}}$. If the underlying density p is differentiable in quadratic mean (DQM), then we have that $\alpha = 2$ which yields the usual rate of $n^{-\frac{1}{2}}$. However, if p is the uniform density on $[-1, 1]$, we have $\alpha = 1$ that gives an optimal rate of n^{-1}. The behavior of the function $\Delta \mapsto H\big(p(\cdot), p(\cdot - \Delta)\big)$ depends on the smoothness of the underlying density p. In the extreme case where p has a Dirac delta point mass at 0 for instance, $H\big(p(\cdot), p(\cdot - \Delta)\big)$ is bounded away from 0 for any $\Delta > 0$. This is expected since, in this case, we can estimate θ_0 perfectly by localizing the discrete point mass. More generally, discontinuities in the density function or singularities in its first derivative anywhere can increase $H\big(p(\cdot), p(\cdot - \Delta)\big)$ and thus lead to a faster rate in estimating the location θ_0. Interested readers can find a detailed discussion and a large class of examples in Chapter VI of Ibragimov and Has' Minskii (2013).

When the underlying density p is unknown, it becomes unclear how to design a rate adaptive location estimator. One possible approach is to nonparametrically estimate p, but we would need our density estimator to be able to accurately recover the points of discontinuities in p or singularities

in p' – this goes beyond the scope of existing theory on nonparametric density estimation which largely deals with estimating a smooth density p. Because of the clear difficulty in analyzing the rate adaptive location estimation problem in its full generality, we focus on rate adaptivity among compactly supported distributions which exhibit discontinuity or singularity at the boundary points of the support; the uniform density on $[-1, 1]$ for instance has discontinuities at points -1 and 1. Moreover, we would like the rate to be no worse than $\sqrt{n}$ (up to polylog factors) for all distributions with a finite variance.

With the more precise goal in mind, we study a simple class of estimators of the form $\widehat{\theta}_\gamma = \arg\min_\theta \sum_{i=1}^n |Y_i - \theta|^\gamma$ where the power $\gamma \geq 2$ is selected in a data-driven way. Estimators of this form cover both the sample mean $\bar{Y}$, with $\gamma = 2$, and the sample midrange, with $\gamma \to \infty$. These estimators are easy to interpret, easy to compute, and can be extended in a straightforward way to the regression setting where θ_0 is a linear function of some observed covariates.

The key step is to select the optimal power γ from the data; in particular, γ must be allowed to diverge with n to allow the resulting estimator to have an adaptive rate. Since $\widehat{\theta}_\gamma$ is unbiased for any $\gamma \geq 2$, the ideal selection criterion is to minimize the variance. In this work, we approximate the variance of $\widehat{\theta}_\gamma$ by its *asymptotic variance*, which has a finite sample empirical analog that can be computed from the empirical central moments of the data. We then select γ by minimizing the empirical asymptotic variance, using Lepski's method to ensure that we consider only those γ's for which the empirical asymptotic variance is a good estimate of the population version. Our main results are finite sample bounds stated in Theorem 9 and Theorem 10. We give an informal statement below.

Main result: (informal) Let $\widehat{\theta}$ be our estimator on $Y_i \sim P(\cdot - \theta_0)$, fully data-driven (no tuning).

1. Write $\mathcal{F}_\sigma$ is the set of all distributions (not necessarily symmetric) with second moment σ^2 and mean 0. We have $\sup_{P \in \mathcal{F}_\sigma} \mathbb{E}|\widehat{\theta} - \theta_0| \leq O(\sigma\sqrt{\log n/n})$.
2. Write $\mathcal{F}_{c,a_1,a_2,\alpha}$ as the set of symmetric distributions P supported on $[-c, c]$ and satisfying $a_1 \frac{c^\gamma}{\gamma^\alpha} \leq \int_{-c}^{c} |x|^\gamma dP(x) \leq a_2 \frac{c^\gamma}{\gamma^\alpha}$ for $\alpha \in (0, 2]$. We have $\sup_{p \in \mathcal{F}_{c,a_1,a_2,\alpha}} \mathbb{E}|\widehat{\theta} - \theta_0| \leq O((\log^2 n/n)^{1/\alpha})$.

Our estimation procedure can be easily adapted to the linear regression setting where we have $Y_i = X_i^\top \beta_0 + Z_i$ where Z_i has a symmetric distribution around 0. It is computationally fast using second order methods and can be directly applied on real data. Importantly, it is robust to violation of the symmetry assumption. More precisely, if $Y_i = \theta_0 + Z_i$ and the noise Z_i has a distribution that is *asymmetric* around 0 but still has mean zero, then our estimator will converge to $\mathbb{E}Y_i = \theta_0$ nevertheless.

Literature Review: Starting from the seminal paper by Stein (1956), a long series of work, for example Stone (1975), Beran (1978), and many others (Van Eeden, 1970; Bickel, 1982; Schick, 1986; Mammen and Park, 1997; Dalalyan et al., 2006) showed, under the regular DQM setting, we can attain an asymptotically efficient estimator $\widehat{\theta}$ by taking a pilot estimator $\widehat{\theta}_{\text{init}}$, applying a density estimation method on the residues $\widetilde{Z}_i = Y_i - \widehat{\theta}_{\text{init}}$ to obtain a density estimate $\widehat{p}$, and then construct $\widehat{\theta}$ either by maximizing the estimated log-likelihood, by taking one Newton step using an estimate of the Fisher information, or by various other related schemes; see Bickel et al. (1993) for more discussion on adaptive efficiency. Interestingly, Laha (2021) recently showed that the smoothness assumption can be substituted by a log-concavity condition instead.

Our work is very closely related to a line of work from Gupta et al. (2023) and Gupta et al. (2022). They observe (rightly) that many asymptotic results hide the fact that the convergence is pointwise, that is, dependent on the unknown density p in opaque ways. They provide finite sample

bound which gives convergence uniform over all densities but they require a smoothing parameter that is difficult to choose in a data-dependent way. We compare our result with these in Remark 12.

Also motivated in part by the contrast between sample midrange and sample mean, Baraud et al. (2017) and Baraud and Birgé (2018) propose the ρ-estimator. When the underlying density p is known, the ρ-estimator has optimal rate in estimating the location. When p is unknown, the ρ-estimator would need to estimate p nonparametrically; it is not clear under what conditions it would attain adaptive rate. Moreover, computing the ρ-estimator in practice is often difficult.

Our estimator is related to methods in robust statistics (Huber, 2011), although our aim is different. Our asymptotic variance based selector can be seen as a generalization of a procedure proposed by Lai et al. (1983), which uses the asymptotic variance to select between the sample mean and the median. Another somewhat related line of work is that of Chierichetti et al. (2014) and Pensia et al. (2019), which study location estimation when $Z_1, \dots, Z_n$ are allowed to have different distributions.

Notation: We write $[n] := \{1, 2, \dots, n\}$. We write $a \wedge b := \min(a, b)$, $a \vee b := \max(a, b)$, $(a)_+ := a \vee 0$ and $(a)_- := -(a \wedge 0)$. For two functions f, g, we write $f \lesssim g$ if there exists a universal constant $C > 0$ such that $f \le Cg$; we write $f \lesssim_\alpha g$ if there exists a $C_\alpha > 0$, which depends on α, such that $f \le C_\alpha g$. we write $f \asymp g$ or $f \propto g$ if $f \lesssim g$ and $g \lesssim f$; $f \asymp_\alpha g$ is defined similarly. We use the $\widetilde{O}(\cdot)$ notation to represent rate of convergence ignoring poly-log factors.

2. Method

We observe random variables $Y_1, \dots, Y_n$ such that $Y_i = \theta_0 + Z_i$ for $i \in [n]$, where $\theta_0 \in \mathbb{R}$ is the unknown location and $Z_1, \dots, Z_n \sim P$ where P is an unknown distribution with density $p(\cdot)$ symmetric around zero. Our goal is to estimate θ_0 from the observations $Y_1, \dots, Y_n$.

2.1. A class of estimators

Our approach is motivated by the fact that both the sample mean and the sample midrange minimize the ℓ^γ norm of the residual for different values of γ. More precisely,

$$\bar{Y} := \frac{1}{n}\sum_{i=1}^n Y_i = \arg\min_{\theta\in\mathbb{R}} \sum_{i=1}^n |Y_i - \theta|^2, \quad \text{and}$$
$$Y_{\text{mid}} := \frac{Y_{(n)} + Y_{(1)}}{2} = \arg\min_{\theta\in\mathbb{R}} \max_{i\in[n]} |Y_i - \theta| = \lim_{\gamma\to\infty} \arg\min_{\theta\in\mathbb{R}} \sum_{i=1}^n |Y_i - \theta|^\gamma.$$

This suggests an estimation scheme where we first select the power $\gamma \ge 2$ in a data-driven way and then output the empirical center with respect to the ℓ^γ norm:

$$\widehat{\theta}_\gamma := \arg\min_{\theta\in\mathbb{R}} \sum_{i=1}^n |Y_i - \theta|^\gamma.$$

It is clear that $\bar{Y} = \widehat{\theta}_2$ and that $\widehat{\theta}_\gamma$ approaches Y_{mid} as γ increases, that is, $Y_{\text{mid}} \equiv \widehat{\theta}_\infty := \lim_{\gamma\to\infty} \widehat{\theta}_\gamma$. We in fact have a deterministic bound of $|\widehat{\theta}_\gamma - Y_{\text{mid}}|$ in the following lemma:

Lemma 1 *Let $Y_1, \dots, Y_n$ be n arbitrary points on $\mathbb{R}$, then $|\widehat{\theta}_\gamma - Y_{mid}| \le 2(Y_{(n)} - Y_{(1)})\frac{\log n}{\gamma}$.*

We prove Lemma 1 in Section S3 of the appendix. Lemma 1 suggests that we need to consider γ as large as n to approximate Y_{mid} with error that is of order $\frac{\log n}{n}$. Therefore, in settings where Y_{mid} is optimal, we need γ to be able to diverge with n.

Estimators of form $\widehat{\theta}_\gamma$ is simple, easy to compute via Newton's method (see Section S3.4 of the appendix), and interpretable even for asymmetric distributions. The key question is of course, how do we select the power γ? It is necessary to allow γ to increase with n to attain adaptive rate but selecting a power γ that is too large can introduce tremendous excess variance. As is often said, *"with great power comes great responsibility"*.

Before describing our approach in the next subsection, we give some remarks on two approaches that seem reasonable but in fact do not work well.

Remark 2 *(Suboptimality of Cross-validation)*
Cross-validation is a natural method for choosing the best estimator among a family of estimators, but this fails in our problem. To illustrate why, we consider the simpler problem where we choose between only the sample mean $\widehat{\theta}_2$ and the sample midrange $\widehat{\theta}_\infty$. We consider held-out validation where we divide our data into training data D^{train} and test data D^{test} each with n data points. We compute $\widehat{\theta}_2^{train}, \widehat{\theta}_\infty^{train}$ on training data, evaluate test data MSE

$$\widehat{R}(\widehat{\theta}_j^{train}) := \frac{1}{n}\sum_{i=1}^{n}(Y_i^{test} - \widehat{\theta}_j^{train})^2 = (\bar{Y}^{test} - \widehat{\theta}_j^{train})^2 + \frac{1}{n}\sum_{i=1}^{n}(Y_i^{test} - \bar{Y}^{test})^2, \tag{1}$$

for $j \in \{2, \infty\}$. Since the second term on the right hand side of (1) is constant, we select $\gamma = 2$ if $(\bar{Y}^{test} - \widehat{\theta}_2^{train})^2 < (\bar{Y}^{test} - \widehat{\theta}_\infty^{train})^2$.

Now assume that the data follows the uniform distribution on $[\theta_0 - 1, \theta_0 + 1]$, so that the optimal estimator is the sample midrange $\widehat{\theta}_\infty$. We observe that $\sqrt{n}(\widehat{\theta}_2^{train} - \theta_0) \xrightarrow{d} N(0, 1/3)$ and $\sqrt{n}(\bar{Y}^{test} - \theta_0) \xrightarrow{d} N(0, 1/3)$ whereas $\sqrt{n}(\widehat{\theta}_\infty^{train} - \theta_0) \to 0$ in probability. Hence, by the Portmanteau Theorem,

$$\begin{aligned}
\liminf_{n\to\infty} \mathbb{P}(\textit{selecting } \widehat{\theta}_2) &= \liminf_{n\to\infty} \mathbb{P}(|\bar{Y}^{test} - \widehat{\theta}_2^{train}| < |\bar{Y}^{test} - \widehat{\theta}_\infty^{train}|) \\
&= \liminf_{n\to\infty} \mathbb{P}(|\sqrt{n}(\bar{Y}^{test} - \theta_0) - \sqrt{n}(\widehat{\theta}_2^{train} - \theta_0)| \\
&\qquad\qquad < |\sqrt{n}(\bar{Y}^{test} - \theta_0) - \sqrt{n}(\widehat{\theta}_\infty^{train} - \theta_0)|) \\
&\geq \mathbb{P}(|W_1 - W_2| < |W_2|) > 0,
\end{aligned}$$

where W_1 and W_2 are independent $N(0, 1/3)$ random variables. In other words, held-out validation has a non-vanishing probability of incorrectly selecting $\widehat{\theta}_2$ over $\widehat{\theta}_\infty$ even as $n \to \infty$ and thus has an error of order $1/\sqrt{n}$, which is far larger than the optimal $1/n$ rate. It is straightforward to extend the argument to the setting of K-fold cross-validation for any fixed K.

Remark 3 *(Suboptimality of MLE with respect to the generalized Gaussian family)*
We observe that $\widehat{\theta}_\gamma$ is the maximum likelihood estimator for the center when the data follow the Generalized Normal GN(θ, σ, γ) distribution, which is also known as the Subbotin distribution (Subbotin, 1923), whose density is of the form $p(x\,;\,\theta, \sigma, \gamma) = \frac{1}{2\sigma\Gamma(1+1/\gamma)} \exp\left(-\left|\frac{x-\theta}{\sigma}\right|^\gamma\right)$, where $\Gamma(t) := \int_0^\infty x^{t-1}e^{-x}dx$ denotes the Gamma function. This suggests a potential approach where we determine γ by fitting the data to the potentially misspecified Generalized Gaussian family via likelihood

maximization:

$$\arg\min_{\gamma} \min_{\theta,\sigma} \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i-\theta}{\sigma}\right|^{\gamma} + \log\sigma + \log(2\Gamma(1+1/\gamma)).$$

This approach works well if the underlying density p of the noise Z_i belongs in the Generalized Gaussian family. Otherwise, it may be suboptimal: it may select a γ that is too small when the optimal γ is large and it may select a γ that is too large when the optimal γ is small. We give a precise and detailed discussion of the drawbacks of the generalized Gaussian MLE in Section S1.1.

2.2. Approximate variance

Under the assumption that the noise Z_i has a distribution symmetric around 0, it is easy to see by symmetry that $\mathbb{E}\widehat{\theta}_\gamma = \theta_0$ for any fixed $\gamma > 0$. We thus propose a selection scheme based on minimizing the variance. The finite sample variance of $\widehat{\theta}_\gamma$ is intractable to compute, but for any fixed $\gamma > 1$, assuming $\mathbb{E}|Y-\theta_0|^{2(\gamma-1)} < \infty$, we have that $\sqrt{n}(\widehat{\theta}_\gamma - \theta_0) \xrightarrow{d} N(0, V(\gamma))$ as $n \to \infty$, where

$$V(\gamma) := \frac{\mathbb{E}|Y-\theta_0|^{2(\gamma-1)}}{\left[(\gamma-1)\mathbb{E}|Y-\theta_0|^{\gamma-2}\right]^2} \tag{2}$$

is the asymptotic variance of $\widehat{\theta}_\gamma$. Thus, from an asymptotic perspective, $\widehat{\theta}_\gamma$ is a better estimator of θ_0 if $V(\gamma)$ is small. When γ is allowed to depend on n, $V(\gamma)$ may not be a good approximation of the finite sample variance of $\widehat{\theta}_\gamma$, but the next example suggests that $V(\cdot)$ is still a sensible criterion.

Example 1 *When $Y_1, \dots, Y_n \overset{d}{\sim}$ Uniform$[\theta_0-1, \theta_0+1]$, straightforward calculation yields that $\mathbb{E}|Y-\theta_0|^q = \frac{1}{q+1}$ for any $q \in \mathbb{N}$ and thus, we have $V(\gamma) = \frac{1}{2\gamma-1}$. We see that $V(\gamma)$ is minimized when $\gamma \to \infty$, in accordance with the fact that the sample midrange Y_{mid} is the optimal estimator among the class of estimators $\{\widehat{\theta}_\gamma\}_{\gamma\geq 2}$. More generally, if Y_i has a density $p(\cdot)$ supported on $[\theta_0-1, \theta_0+1]$ which is symmetric around θ_0 and satisfies the property that $p(x)$ is bounded away from 0 and ∞ for all $x \in [\theta_0-1, \theta_0+1]$, then one may show that $V(\gamma) \propto \frac{1}{\gamma}$. On the other hand, if $Y_i \sim N(\theta_0, 1)$, then, using the fact that $\mathbb{E}|Y-\theta_0|^\gamma \asymp \gamma^{\gamma/2}e^{-\gamma/2}$, we can directly calculate that that $V(\gamma)$ is of order $\frac{2^\gamma}{\gamma}$, which goes to infinity as $\gamma \to \infty$ as expected. Since $\widehat{\theta}_2$ is the MLE, we have that $V(\gamma)$ is minimized at $\gamma = 2$ in the Gaussian case (Van der Vaart, 2000, Chapter 5.5).*

2.3. Proposed procedure

We thus propose to select γ by minimizing an estimate of the asymptotic variance $V(\gamma)$. It is important to note that although we use $V(\gamma)$ in our procedure, our error bounds are non-asymptotic.

For simplicity, we restrict our attention to $\gamma \geq 2$ in the main paper and discuss how to select $\gamma \in [1, 2)$ in Remark 7. A natural estimator of $V(\gamma)$ is

$$\widehat{V}(\gamma) := \frac{\min_\theta \frac{1}{n}\sum_{i=1}^{n}|Y_i-\theta|^{2(\gamma-1)}}{\left[(\gamma-1)\min_\theta \frac{1}{n}\sum_{i=1}^{n}|Y_i-\theta|^{\gamma-2}\right]^2}. \tag{3}$$

Although $\widehat{V}(\gamma)$ has pointwise consistency in that it is a consistent estimator of $V(\gamma)$ for any fixed γ (see Lemma S3.3 in Section S3.2 of the appendix), we require uniform consistency since

our goal is to minimize $\widehat{V}(\gamma)$ as a surrogate of $V(\gamma)$. This unfortunately does not hold; if we allow γ to diverge with n, the error $|\widehat{V}(\gamma) - V(\gamma)|$ can be arbitrarily large. This occurs because, if we fix n and increase γ, the finite average $\frac{1}{n}\sum_{i=1}^n |Y_i - \theta|^\gamma$ does not approximate the population mean and behaves closer to $\frac{1}{n}\max_i |Y_i - \theta|^\gamma$ instead. Indeed, for any fixed n and any deterministic set of points $Y_1, \ldots, Y_n$, we have

$$\begin{aligned}\widehat{V}(\infty) := \lim_{\gamma\to\infty} \widehat{V}(\gamma) &= \lim_{\gamma\to\infty} \frac{n}{(\gamma-1)^2} \frac{|Y_{(n)} - Y_{\text{mid}}|^{2(\gamma-1)} + |Y_{(1)} - Y_{\text{mid}}|^{2(\gamma-1)}}{\{|Y_{(n)} - Y_{\text{mid}}|^{\gamma-2} + |Y_{(1)} - Y_{\text{mid}}|^{\gamma-2}\}^2} \\ &= \lim_{\gamma\to\infty} \frac{n}{2(\gamma-1)^2} \left|\frac{Y_{(n)} - Y_{(1)}}{2}\right|^2 = 0. \end{aligned} \tag{4}$$

Therefore, unconstrained minimization of $\widehat{V}(\gamma)$ over all $\gamma \geq 1$ would select $\gamma = \infty$. See for example Figure 1(*a*), where we generate Gaussian noise $Z_i \sim N(0,1)$ and plot $\widehat{V}(\gamma)$ for a range of γ's; although the population $V(\gamma)$ tends to infinity when γ is large, the empirical $\widehat{V}(\gamma)$ increases for moderately large γ but then, as γ further increases, $\widehat{V}(\gamma)$ decreases and tends to 0.

Luckily, we can overcome this issue by restricting our attention to γ's that are not too large. To be precise, we add an upper bound $\gamma_{\max} \geq 2$ and minimize $\widehat{V}(\gamma)$ only among $\gamma \in [2, \gamma_{\max}]$. We select $\gamma_{\max}$ using Lepski's method, which is typically used to select smoothing parameters in nonparametric estimation problems (Lepskii, 1990, 1991) but can be readily adapted to our setting. The idea is to construct confidence intervals $\text{CI}_\gamma := \left[\widehat{\theta}_\gamma - \tau\sqrt{\widehat{V}(\gamma)/n},\ \widehat{\theta}_\gamma + \tau\sqrt{\widehat{V}(\gamma)/n}\right]$ for a set of γ's, starting with $\gamma = 2$, and take $\gamma_{\max}$ to be the largest γ such that the confidence intervals all intersect, i.e. $\gamma_{\max} := \sup\{\widetilde{\gamma} : \cap_{\gamma\leq\widetilde{\gamma}}\text{CI}_\gamma \neq \emptyset\}$. We would thus exclude γ for which $\widehat{\theta}_\gamma$ is far from θ_0 and $\widehat{V}(\gamma)$ is too small. This leads to our full estimation procedure below, which we refer to as **CAVS** (Constrained Asymptotic Variance Selector):

Algorithm 1 Constrained Asymptotic Variance Selection (CAVS) algorithm

---

Let $\tau > 0$ be a tuning parameter and let $\mathcal{N}_n \subseteq [2, \infty]$ be the set of candidate γ's. Define $\widehat{V}(\gamma)$ as (3) for $\gamma \in [2, \infty)$ and define $\widehat{V}(\infty) := 0$.

1. Define $\gamma_{\max}$ as the largest $\gamma \in \mathcal{N}_n$ such that

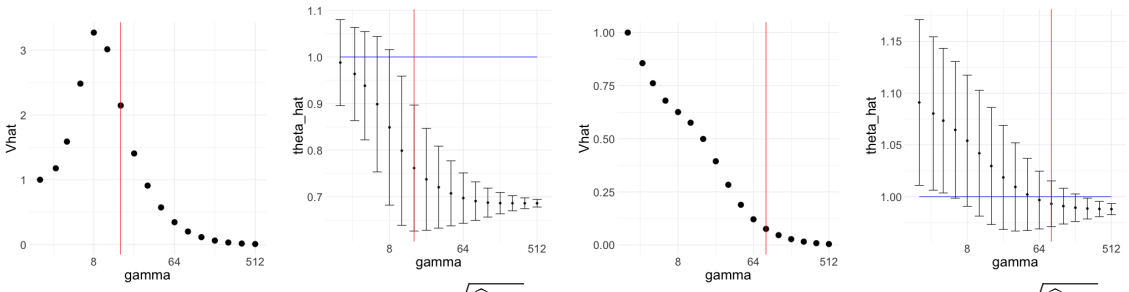
$$\bigcap_{\gamma\in\mathcal{N}_n,\, \gamma\leq\gamma_{\max}} \left[\widehat{\theta}_\gamma - \tau\sqrt{\frac{\widehat{V}(\gamma)}{n}},\ \widehat{\theta}_\gamma + \tau\sqrt{\frac{\widehat{V}(\gamma)}{n}}\right] \neq \emptyset.$$

2. Select $\widehat{\gamma} := \arg\min_{\gamma\in\mathcal{N}_n,\, \gamma\leq\gamma_{\max}} \widehat{V}(\gamma)$.

3. Output $\widehat{\theta} \equiv \widehat{\theta}_{\widehat{\gamma}} = \arg\min_{\theta\in\mathbb{R}} \sum_{i=1}^n |Y_i - \theta|^{\widehat{\gamma}}$.

The candidate set $\mathcal{N}_n$ can be the entire half-line $[2, \infty]$. In practice, we take $\mathcal{N}_n$ to be a finite set so that we are able to compute the minimizer of $\widehat{V}(\gamma)$. A convenient and computationally efficient choice is $\mathcal{N}_n = \{2, 4, 8, \ldots, n, \infty\}$, which we also use in our theory.

We illustrate how the CAVS procedure works with two examples in Figure 1. In Figure 1(*a*), we generate Gaussian noise $Z_i \sim N(0,1)$; we plot $\widehat{V}(\gamma)$ for a exponentially increasing sequence of



(a) $\widehat{V}(\gamma)$ for Gaussian (b) $\widehat{\theta}_\gamma \pm 2\sqrt{\frac{\widehat{V}(\gamma)}{n}}$ (c) $\widehat{V}(\gamma)$ for trunc Gaussian (d) $\widehat{\theta}_\gamma \pm 2\sqrt{\frac{\widehat{V}(\gamma)}{n}}$

Figure 1: Red line gives $\gamma_{\max}$; blue line is the true θ_0. We use $n = 500$ and select $\gamma_{\max}$ as the largest γ such that all confidence intervals to the left have a nonempty intersection.

γ's ranging from 2 to 512. The constraint upper bound $\gamma_{\max}$ is given by the red line in the figure. Unconstrained minimization of $\widehat{V}(\gamma)$ leads to $\widehat{\gamma} = 512$. Figure 1(*b*) illustrates the Lepski method that we use to choose upper bound $\gamma_{\max}$: we compute confidence intervals of width $\tau\sqrt{\widehat{V}(\gamma)/n}$ around $\widehat{\theta}_\gamma$ for the whole range of γ's. To get $\gamma_{\max}$, we pick the largest γ such that the intersection of all the confidence intervals to the left of $\gamma_{\max}$ is non-empty. This allows us to avoid the region where $\widehat{V}(\gamma)$ is very small but the actual $V(\gamma)$ is very large. Indeed, if $V(\gamma)$ is much larger than the variance $\mathrm{Var}(Z)$, then $\widehat{\theta}_\gamma$ likely to be far from the sample mean $\widehat{\theta}_2$ and thus, if $\widehat{V}(\gamma)$ is also small, then $\widehat{\theta}_\gamma \pm \tau\sqrt{\widehat{V}(\gamma)/n}$ is unlikely to overlap with the confidence interval around the sample mean.

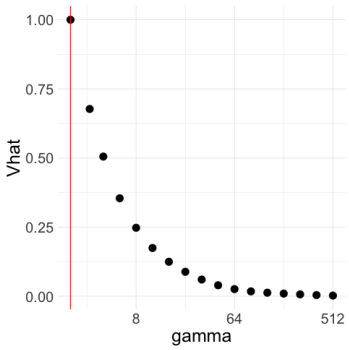
Therefore, with Gaussian noise, CAVS selects $\widehat{\gamma} = 2$ by minimizing $\widehat{V}(\gamma)$ for only those γ to the left of $\gamma_{\max}$ (red line) in Figure 1(*a*). In contrast, if $V(\gamma)$ decreases as γ increases, then $\widehat{\theta}_\gamma$ remains close to the sample mean $\widehat{\theta}_2$ and the confidence interval $\widehat{\theta}_\gamma \pm \tau\sqrt{\widehat{V}(\gamma)/n}$ overlaps with that of the sample mean even when γ is large, which means we would select a large $\gamma_{\max}$ as desired. We illustrate this in Figure 1(*c*) and 1(*d*), where we generate truncated Gaussian noise Z by truncating at $|Z| \le 2$; that is, we generate Gaussian samples and keep only those that lie in the interval $[-2, 2]$. From Figure 1(*c*), we see that our procedure picks a large $\widehat{\gamma} = 128$. We provide extensive numerical experiments in Section S2.

Remark 4 *(Selecting τ parameter) Our proposed CAVS procedure has a tuning parameter τ which governs the strictness of the $\gamma_{\max}$ constraint. Smaller τ will in general result in a smaller $\gamma_{\max}$ and hence a stronger constraint. For our theoretical results, namely Theorem 10, it suffices to choose τ to be very slowly growing so that $\frac{\tau}{\sqrt{\log\log n}} \to \infty$. For practical data analysis applications, we recommend $\tau = 1$ as a conservative choice based on simulation studies in Section S2.1*

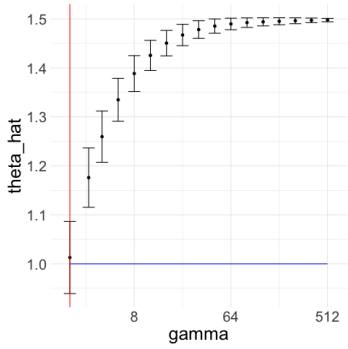
Remark 5 *(Robustness to asymmetry) One important aspect of CAVS is that it is robust to violations of the symmetry assumption. If the density p of the noise has mean zero but is asymmetric (so that θ_0 is the mean of Y_i), then, there may exist γ greater than 2 where the γ-th center of $Y_i = \theta_0 + Z_i$ is different from θ_0 (i.e. $\theta^*_\gamma := \arg\min_\theta \mathbb{E}|Y - \theta|^\gamma \neq \theta_0$) and so $\widehat{\theta}_\gamma$ is a biased estimator of θ_0.*

In such cases however, the confidence interval $\widehat{\theta}_\gamma \pm \tau\sqrt{\frac{\widehat{V}(\gamma)}{n}}$ will, for large enough n, be concentrated around $\mathbb{E}\widehat{\theta}_\gamma = \theta^*_\gamma$ and thus not overlap with the confidence interval about the sample mean

$\widehat{\theta}_2 \pm \tau\sqrt{\frac{\widehat{V}(2)}{n}}$, *which will concentrated around* $\mathbb{E}\widehat{\theta}_2 = \theta_0$. *Therefore, we would have* $\gamma_{\max} < \gamma$ *and the constraint would thus exclude any biased* $\widehat{\theta}_\gamma$. *We illustrate an example in Figure 2 where because* $\widehat{\theta}_3$ *is biased, we have that* $\gamma_{\max} = 2$ *and thus, we select* $\widehat{\gamma} = 2$ *and the resulting estimator* $\widehat{\theta}_{\widehat{\gamma}}$ *still converges to* θ_0. *Indeed, Theorem 9 does not require the noise distribution to be symmetric around* 0, *it only requires the noise to have mean zero.*



(*a*) $\widehat{V}$ for asymmetric density



(*b*) $\widehat{\theta}_\gamma \pm 2\sqrt{\frac{\widehat{V}(\gamma)}{n}}$

Figure 2: We generate mean zero Z_i from an asymmetric mixture distribution $\frac{2}{3}\text{Unif}[-1,0] + \frac{1}{3}\text{Unif}[0,2]$. Note that $\widehat{\theta}_3 \pm 2\sqrt{\frac{\widehat{V}(3)}{n}}$ does not overlap with $\widehat{\theta}_2 \pm 2\sqrt{\frac{\widehat{V}(2)}{n}}$ because $\mathbb{E}\widehat{\theta}_2 \neq \mathbb{E}\widehat{\theta}_3$ due to the asymmetry. Red line gives $\gamma_{\max}$; blue line is the true θ_0.

Remark 6 *(Extension to the regression setting) We can directly extend our estimation procedure to the linear regression setting. Suppose we observe* (Y_i, X_i) *for* $i = 1, 2, \ldots, n$ *where* X_i *is a random vector on* $\mathbb{R}^d$, $Y_i = X_i^\top \beta_0 + Z_i$, *and* Z_i *is an independent noise with a distribution symmetric around* 0. *Then, we would compute, for each* γ *in a set* $\mathcal{N}_n \subset [2, \infty]$,

$$\widehat{\beta}_\gamma = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \sum_{i=1}^n |Y_i - X_i^\top \beta|^\gamma \text{ and } \widehat{V}(\gamma) = \frac{\min_{\beta \in \mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n |Y_i - X_i^\top \beta|^{2(\gamma-1)}}{(\gamma-1)^2 \{\min_{\beta \in \mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n |Y_i - X_i^\top \beta|^{\gamma-2}\}^2}.$$

We define $\widehat{\Sigma}_X := \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$. *Using Taylor expansion, it is straightforward to show that* $\sqrt{n}\widehat{\Sigma}_X^{1/2}(\widehat{\beta}_\gamma - \beta_0) \xrightarrow{d} N(0, V(\gamma) I_d)$. *Thus, for a given* $\tau > 0$, *our estimation procedure first computes* $\gamma_{\max}$ *as the largest* $\gamma \in \mathcal{N}_n$ *such that*

$$\bigcap_{\gamma \in \mathcal{N}_n, \gamma \leq \gamma_{\max}} \bigotimes_{j=1}^p \left[\left(\widehat{\Sigma}_X^{1/2}\widehat{\beta}_\gamma\right)_j - \tau\sqrt{\frac{\widehat{V}(\gamma)}{n}}, \left(\widehat{\Sigma}_X^{1/2}\widehat{\beta}_\gamma\right)_j + \tau\sqrt{\frac{\widehat{V}(\gamma)}{n}} \right] \neq \emptyset,$$

where we use the $\otimes$ *notation to denote the Cartesian product. Then, we select the minimizer* $\widehat{\gamma} = \arg\min_{\gamma \in \mathcal{N}_n, \gamma \leq \gamma_{\max}} \widehat{V}(\gamma)$ *and output* $\widehat{\beta}_{\widehat{\gamma}}$.

Remark 7 *(Selecting* $\gamma \in [1, 2)$*) When the noise* Z_i *is heavy-tailed, it is desirable to allow consideration of* $\gamma \in [1, 2)$*; note that* $\gamma = 1$ *corresponds to the sample median* $\widehat{\theta}_1 = \arg\min_\theta \sum_{i=1}^n |Y_i -$

$\theta|$. *For* $\gamma \in [1,2)$, *the estimator* $\widehat{V}(\gamma)$ *given in* (3) *is not appropriate. In particular, if* Z_i *has a density* p *and population median 0 and that* $p(0) > 0$, *then the asymptotic variance of sample median is* $V(1) = \frac{1}{4p(0)^2}$ *instead of* (2). *For* $\gamma \in (1,2)$, *expression* (2) *holds but the estimator* $\widehat{V}(\gamma)$ *may behave poorly because of the negative power in the denominator. We do not have a general way of estimating* $V(\gamma)$ *for* $\gamma < 2$. *In the specific case of the sample median* ($\gamma = 1$), *there are various good estimators of the variance. For instance, Bloch and Gastwirth (1968) proposed an approach based on density estimation and Lai et al. (1983) proposed an approach based on the bootstrap. The general idea of selecting an estimator using asymptotic variance is not specific to the* L_γ*-centers; one can also add say Huber loss minimizers into the set of candidate estimators provided that there is a good way to estimate the asymptotic variance.*

Remark 8 *An important property of* $\widehat{\gamma}$ *is that it is shift and scale invariant in the following sense: if we apply the transformation* $\widetilde{Y}_i = bY_i + a$ *where* $b > 0$ *and* $a \in \mathbb{R}$ *and then compute* $\widetilde{\gamma}$ *on* $\{\widetilde{Y}_1, \dots, \widetilde{Y}_n\}$, *then* $\widetilde{\gamma} = \widehat{\gamma}$. *This follows from the fact that* $\widehat{V}(\gamma)/\widehat{V}(2)$ *is shift and scale invariant. Likewise, we see that* $\widehat{\theta}_{\widehat{\gamma}}$ *is shift and scale equivariant in that if we compute* $\widetilde{\theta}_{\widetilde{\gamma}}$ *on* $\{\widetilde{Y}_1, \dots, \widetilde{Y}_n\}$, *then* $\widetilde{\theta}_{\widetilde{\gamma}} = b\widehat{\theta}_{\widehat{\gamma}} + a$.

2.4. Error rate is at least $1/\sqrt{n}$

Using the definition of $\widehat{\gamma}$, we can directly show that $\widehat{\theta}_{\widehat{\gamma}}$ must be close to the sample mean $\bar{Y}$ and that the error of $\widehat{\theta}_{\widehat{\gamma}}$ is at most $O(\tau\sqrt{\sigma^2/n})$ where $\sigma^2 := \mathrm{Var}(Z)$.

Theorem 9 *Let* $\widehat{\sigma}^2$ *be the empirical variance of* $Y_1, \dots, Y_n$. *For any* n, *it holds surely that* $|\widehat{\theta}_{\widehat{\gamma}} - \bar{Y}| \le 2\tau\sqrt{\frac{\widehat{\sigma}^2}{n}}$. *Therefore, if we additionally have that* $\sigma^2 := \mathbb{E}|Z|^2 < \infty$, *then, writing* $\theta_0 = \mathbb{E}Y_1$,

$$\mathbb{E}|\widehat{\theta}_{\widehat{\gamma}} - \theta_0| \le C\tau\sqrt{\frac{\sigma^2}{n}}$$

for a universal constant $C > 0$.

Proof Since $\widehat{\gamma} \le \gamma_{\max}$, we have by the definition of $\gamma_{\max}$ that $\widehat{\theta}_{\widehat{\gamma}} + \tau\sqrt{\frac{\widehat{V}(\widehat{\gamma})}{n}} \ge \widehat{\theta}_2 - \tau\sqrt{\frac{\widehat{V}(2)}{n}}$ and $\widehat{\theta}_{\widehat{\gamma}} - \tau\sqrt{\frac{\widehat{V}(\widehat{\gamma})}{n}} \le \widehat{\theta}_2 - \tau\sqrt{\frac{\widehat{V}(2)}{n}}$. Since $\widehat{V}(\widehat{\gamma}) \le \widehat{V}(2)$ by definition of $\widehat{\gamma}$ and since $\widehat{\theta}_2 = \bar{Y}$ and $\widehat{V}(2) = \widehat{\sigma}^2$, the first claim immediately follows. The second claim directly follows from the first claim. ■

It is important to note that Theorem 9 does not require symmetry of the noise distribution P. If Y_i has a distribution *asymmetric* around θ_0 but $\mathbb{E}Y = \theta_0$, then Theorem 9 implies that $\widehat{\theta}_{\widehat{\gamma}}$ converges to θ_0 as might be desired.

3. Adaptive rate on compacted supported densities

Theorem 9 shows that, so long as τ is small and the noise Z_i has finite variance, then our proposed estimator has an error $\mathbb{E}|\widehat{\theta}_{\widehat{\gamma}} - \theta_0|$ that is at most $\widetilde{O}(n^{-1/2})$. In this section, we analyze the behavior of our estimator when the density $p(\cdot)$ is supported on the interval $[-c, c]$. Since our estimator is scale-equivariant, we can without loss of generality assume that the support of $p(\cdot)$ is $[-1, 1]$; note then that $\mathbb{E}|Z_i|^\gamma \le 1$ for all $\gamma \ge 0$ and that $\mathbb{E}|Z_i|^\gamma$ is non-increasing in γ.

We prove that if the moments decrease at a rate $\mathbb{E}|Z_i|^\gamma \propto 1/\gamma^\alpha$ where $\alpha \in (0, 2]$, then our estimator, without knowing α, can attain an adaptive rate of convergence of $\widetilde{O}(n^{-\frac{1}{\alpha}})$.

Theorem 10 *Suppose $Z_1, Z_2, \ldots, Z_n$ are independent and identically distributed with a distribution P symmetric around 0. Suppose there exists $\alpha \in (0, 2]$, $a_1 \in (0, 1]$ and $a_2 \geq 1$ such that $\frac{a_1}{\gamma^\alpha} \leq \mathbb{E}|Z|^\gamma \leq \frac{a_2}{\gamma^\alpha}$ for all $\gamma \geq 1$. Let $\mathcal{N}_n$ be a subset of $[2, \infty]$ with $M_n := \sup \mathcal{N}_n$ and suppose $\mathcal{N}_n$ contains 2^k for all integer $k \leq n \wedge \log_2 M_n$.*

Let $C_{a_1,a_2,\alpha} > 0$ be a constant that depends only on a_1, a_2, α; let $\widehat{\theta}_{\widehat{\gamma}}$ be defined as in Algorithm 1. The following then hold:

1. *If $\frac{\tau}{\sqrt{\log \log n}} \to \infty$, then*
$$|\widehat{\theta}_{\widehat{\gamma}} - \theta_0| \leq O_p\left(C_{a_1,a_2,\alpha}\left\{\left(\frac{\log^{\alpha+1} n}{n}\right)^{\frac{1}{\alpha}} \vee \frac{\log n}{M_n}\right\}\right).$$
2. *If $\tau \geq \sqrt{\log n}$, then*
$$\mathbb{E}|\widehat{\theta}_{\widehat{\gamma}} - \theta_0| \leq C_{a_1,a_2,\alpha}\left\{\left(\frac{\log^{\alpha+1} n}{n}\right)^{\frac{1}{\alpha}} \vee \frac{\log n}{M_n}\right\}.$$

Therefore, we can choose $M_n \geq 2^n$ and $\tau = \sqrt{\log n}$, without any knowledge of α, so that our estimator has an adaptive rate of convergence $\mathbb{E}|\widehat{\theta}_{\widehat{\gamma}} - \theta_0| \lesssim_{a_1,a_2,\alpha} \left(\frac{\log^{\alpha+1} n}{n}\right)^{\frac{1}{\alpha}}$ where α can take on any value in $(0, 2]$ depending on the underlying noise distribution. The adaptive rate $(\frac{\log^{\alpha+1}}{n})^{1/\alpha}$ is, up to log-factors, minimax optimal for the class of densities satisfying $\mathbb{E}|Z|^\gamma \propto \gamma^{-\alpha}$; see Remark 16 for more details.

We relegate the proof of Theorem 10 to Section S4.1 of the appendix, but give a sketch of the proof ideas here. First, by using the moment condition $\frac{a_1}{\gamma^\alpha} \leq \mathbb{E}|Z|^\gamma \leq \frac{a_2}{\gamma^\alpha}$ as well as Talagrand's inequality, we give the following uniform bound for $\widehat{V}(\gamma)$: that $\widehat{V}(\gamma) \asymp_{a_1,a_2,\alpha} \gamma^{\alpha-2} \asymp_{a_1,a_2,\alpha} V(\gamma)$ for all $2 \leq \gamma \leq \left(\frac{n}{\log n}\right)^{\frac{1}{\alpha}}$ with high probability. Using this bound in conjunction with another uniform bound on $|\widehat{\theta}_\gamma - \theta_0|$, we then can guarantee that $\gamma_{\max}$ is large enough in that $\gamma_{\max} \gtrsim_{a_1,a_2} \left(\frac{n}{\log n}\right)^{\frac{1}{\alpha}} \vee M_n$. These results in turn yields the key fact that $\widehat{\gamma}$ is also sufficiently large in that $\widehat{\gamma} \gtrsim_{a_1,a_2} \left(\frac{n}{\log n}\right)^{\frac{1}{\alpha}} \vee M_n$. We then bound the error of $\widehat{\theta}_{\widehat{\gamma}}$ by the inequality $|\widehat{\theta}_{\widehat{\gamma}} - \theta_0| \leq |\widehat{\theta}_{\widehat{\gamma}} - Y_{\text{mid}}| + |\theta_0 - Y_{\text{mid}}|$, where Y_{mid} is the sample midrange. We control the first term $|\widehat{\theta}_{\widehat{\gamma}} - Y_{\text{mid}}|$ through Lemma 1 and the second term $|\theta_0 - Y_{\text{mid}}|$ using the moment condition. The resulting bound gives the desired conclusion of Theorem 10. We also see that the oracle choice of a data dependent γ is any value in the range $[(n/\log n)^{1/\alpha}, \infty)$.

Remark 11 *A direct implication of Theorem 10 and the scale-equivariance of our estimator is that if Z_i takes value on $[-c, c]$ for any $c > 0$ and satisfies $\mathbb{E}|Z|^\gamma \propto \frac{c^\gamma}{\gamma^\alpha}$, then we have that $\mathbb{E}|\widehat{\theta}_{\widehat{\gamma}} - \theta_0| \leq c \cdot C_{a_1,a_2,\alpha}\left\{\left(\frac{\log^{\alpha+1} n}{n}\right)^{\frac{1}{\alpha}} \vee \frac{\log n}{M_n}\right\}$.*

Remark 12 *Gupta et al. (2023) constructed an estimator $\widehat{\theta}$, using on nonparametric density estimation, that has finite sample error bound $O_p(\sqrt{1/n\mathcal{I}_r})$ (Theorem 1 in Gupta et al. (2023)). They do not assume smoothness in the underlying density p but their error bound is in terms of a smooth Fisher information $\mathcal{I}_r$, which is the Fisher information of p convolved with $N(0, r^2)$. We note that their theory does not directly apply to our setting. The Fisher information of Unif$[-1, 1] \star N(0, r^2)$ diverges at a rate of $O(1/r^2)$ so that we would need $r = O(1/\sqrt{n})$ in order to obtain the optimal rate. In contrast, Theorem 1 in Gupta et al. (2023) requires $r \geq n^{-1/13}$. Moreoever, in Gupta et al. (2023), the smoothing parameter r must be chosen by the user. It is unclear whether one can choose r in a data-dependent way that leads to an estimator with optimal adaptive rate.*

Remark 13 *The proof of Theorem 10 shows that when $\mathbb{E}|Z_i|^\gamma \propto \gamma^{-\alpha}$, the midrange estimator Y_{mid} has rate $\widetilde{O}(n^{-\frac{1}{\alpha}})$. However, this does not mean we can simply use the midrange estimator – the midrange performs very poorly when the density is not compactly supported. For example, as mentioned in the introduction, if $Z_i \sim N(0,1)$, then Y_{mid} has rate $O(1/\sqrt{\log n})$. In contrast, our proposed estimator $\widehat{\gamma}_{\widehat{\gamma}}$ has rate no worse than $\widetilde{O}(1/\sqrt{n})$ regardless of the underlying distribution (so long as the variance is finite).*

Remark 14 *One reviewer pointed out that our estimator $\widehat{\theta}_{\widehat{\gamma}}$ may have a rate of convergence that is exponentially fast in n in some situations where P does not have a density, such as when P is the Rademacher distribution. In this case, we have that $\mathbb{E}|Z|^\gamma = 1$ for all $\gamma > 0$ which corresponds to $\alpha = 0$. This is not covered by Theorem 10 but we believe the analysis can be readily extended to this case to show that $\mathbb{E}|\widehat{\theta}_{\widehat{\gamma}} - \theta_0| = \widetilde{O}(2^{-n})$. Indeed, so long as not all of the Y_i's are $\theta_0 + 1$ or $\theta_0 - 1$, the midrange estimator would estimate θ_0 perfectly.*

3.1. Concrete examples

The moment condition $\frac{a_1}{\gamma^\alpha} \leq \mathbb{E}|Z|^\gamma \leq \frac{a_2}{\gamma^\alpha}$ constrains the behavior of the density $p(\cdot)$ around the boundary of the support $[-1,1]$. The following Proposition formalizes this intuition.

Proposition 15 *Let $\alpha \in (0,2)$ and suppose X is a random variable with density $p(\cdot)$ satisfying $C_{\alpha,1}(1-|x|)_+^{\alpha-1} \leq p(x) \leq C_{\alpha,2}(1-|x|)_+^{\alpha-1}, \quad \forall x \in [-1,1]$, for $C_{\alpha,1}, C_{\alpha,2} > 0$ dependent only on α. Then, there exists $C'_{\alpha,1}, C'_{\alpha,2} > 0$, dependent only on α, such that, for all $\gamma \geq 1$,*

$$\frac{C'_{\alpha,1}}{\gamma^\alpha} \leq \mathbb{E}|X|^\gamma \leq \frac{C'_{\alpha,2}}{\gamma^\alpha}.$$

We prove Proposition 15 in Section S4.2 of the Appendix.

Example 2 *Using Proposition 15, we immediately obtain examples of noise distributions where the convergence rates of our location estimator $\widehat{\theta}_{\widehat{\gamma}}$ vary over a wide range.*

1. *When Z has the semicircle density $p(x) \propto (1-|x|^2)^{1/2}$, then $\mathbb{E}|Z|^\gamma \propto \gamma^{-\frac{3}{2}}$ so that $\widehat{\theta}_{\widehat{\gamma}}$ has rate $\widetilde{O}(n^{-\frac{2}{3}})$, where we use the $\widetilde{O}(\cdot)$ notation to indicate that we have ignored polylog terms.*
2. *When $Z \sim Unif[-1,1]$, we have that $\mathbb{E}|Z|^\gamma = \frac{1}{\gamma+1}$ so that $\widehat{\theta}_{\widehat{\gamma}}$ has rate $\widetilde{O}(n^{-1})$.*
3. *More generally, let q be a symmetric continuous density on $\mathbb{R}$ and let p be a density that results from truncating q, that is, $p(x) \propto q(x)\mathbb{1}\{|x| \leq 1\}$. If $p(1) = p(-1) > 0$, then $\frac{a_1}{\gamma} \leq \mathbb{E}|Z|^\gamma \leq \frac{a_2}{\gamma}$ where a_1, a_2 depend on q. In particular, if Z is a truncated Gaussian, then $\widehat{\gamma}_{\widehat{\gamma}}$ also has $\widetilde{O}(n^{-1})$ rate.*
4. *Suppose Z has a U-shaped density of the form $p(x) \propto (1-|x|)^{-\frac{1}{2}}$, then $\mathbb{E}|Z|^\gamma \propto \gamma^{-\frac{1}{2}}$ so that $\widehat{\theta}_{\widehat{\gamma}}$ has rate $\widetilde{O}(n^{-2})$.*

Remark 16 *By Proposition S4.13 and the subsequent Remark S4.14 in Section S4.2 of the Appendix, we have that if a density p is of the form $p(x) = C_\alpha(1-|x|)^{\alpha-1}\mathbb{1}\{|x| \leq 1\}$ for $\alpha \in (0,2)$, then we have that, writing $H^2(\theta_1,\theta_2) := \int\left(\sqrt{p(x-\theta_1)} - \sqrt{p(x-\theta_2)}\right)^2 dx$, that $C_{\alpha,1}|\theta_1 - \theta_2|^\alpha \leq$*

$H^2(\theta_1, \theta_2) \le C_{\alpha,2}|\theta_1 - \theta_2|^\alpha$, for $C_{\alpha,1}$ and $C_{\alpha,2}$ dependent only on α. From Le Cam (1973, Proposition 1), any estimator $\widehat{\theta}$ has a rate lower bounded by the fact that $H^2(\widehat{\theta}, \theta_0) \gtrsim \frac{1}{n}$ so that among the class of densities

$$\mathcal{P}_{a_1,a_2} := \left\{ p \ : \ \textit{symmetric}, \ \frac{a_1}{\gamma^\alpha} \le \int |x|^\gamma p(x)dx \le \frac{a_2}{\gamma^\alpha}, \forall \gamma \ge 1, \textit{ for some } \alpha \in (0, 2] \right\}, \tag{5}$$

our proposed estimator $\widehat{\theta}_{\widehat{\gamma}}$ has a rate of convergence that is minimax optimal up to polylog factors.

4. Discussion

An immediate question is how to design an estimator whose rate adapts to any discontinuity in the underlying density p, even those in the interior of the support. A promising approach is to apply the technique of Gupta et al. (2023) and use Lepski's method to choose the smoothing parameter r. One challenge is that in other to obtain rates faster than $1/\sqrt{n}$, we may need to estimate the location of the discontinuities of p with $o(1/\sqrt{n})$. Another interesting question is how to extend the framework to nonparametric regression.

References

Yannick Baraud and Lucien Birgé. Rho-estimators revisited: General theory and applications. *The Annals of Statistics*, 46(6B):3767–3804, 2018.

Yannick Baraud, Lucien Birgé, and Mathieu Sart. A new method for estimation and model selection: ρ-estimation. *Inventiones mathematicae*, 207(2):425–517, 2017.

Rudolf Beran. An efficient and robust adaptive estimator of location. *The Annals of Statistics*, pages 292–313, 1978.

Peter J Bickel. On adaptive estimation. *The Annals of Statistics*, pages 647–671, 1982.

Peter J Bickel, Chris AJ Klaassen, Ya'acov Ritov, and Jon A Wellner. *Efficient and adaptive estimation for semiparametric models*, volume 4. Springer, 1993.

Daniel A Bloch and Joseph L Gastwirth. On a simple estimate of the reciprocal of the density function. *The Annals of Mathematical Statistics*, 39(3):1083–1085, 1968.

Flavio Chierichetti, Anirban Dasgupta, Ravi Kumar, and Silvio Lattanzi. Learning entangled single-sample gaussians. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 511–522. SIAM, 2014.

AS Dalalyan, GK Golubev, and AB Tsybakov. Penalized maximum likelihood and semiparametric second-order efficiency. *The Annals of Statistics*, 34(1):169–201, 2006.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press, 2016.

Shivam Gupta, Jasper Lee, Eric Price, and Paul Valiant. Finite-sample maximum likelihood estimation of location. *Advances in Neural Information Processing Systems*, 35:30139–30149, 2022.

Shivam Gupta, Jasper CH Lee, and Eric Price. Finite-sample symmetric mean estimation with fisher information rate. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4777–4830. PMLR, 2023.

Peter J Huber. Robust statistics. In *International encyclopedia of statistical science*, pages 1248–1251. Springer, 2011.

Ildar Abdulovich Ibragimov and Rafail Zalmanovich Has' Minskii. *Statistical estimation: asymptotic theory*, volume 16. Springer Science & Business Media, 2013.

Nilanjana Laha. Adaptive estimation in symmetric location model under log-concavity constraint. *Electronic Journal of Statistics*, 15(1):2939–3014, 2021.

TL Lai, Herbert Robbins, and KF Yu. Adaptive choice of mean or median in estimating the center of a symmetric distribution. *Proceedings of the National Academy of Sciences*, 80(18):5803–5806, 1983.

Lucien Le Cam. Convergence of estimates under dimensionality restrictions. *Annals of Statistics*, 1 (1):38–53, 1973.

OV Lepskii. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466, 1990.

OV Lepskii. Asymptotically minimax adaptive estimation. i: Upper bounds. optimally adaptive estimates. *Theory of Probability & Its Applications*, 36(4):682–697, 1991.

Enno Mammen and Byeong U Park. Optimal smoothing in adaptive location estimation. *Journal of statistical planning and inference*, 58(2):333–348, 1997.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Estimating location parameters in entangled single-sample distributions. *arXiv preprint arXiv:1907.03087*, 2019.

Anton Schick. On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, pages 1139–1151, 1986.

Charles Stein. Efficient nonparametric testing and estimation. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, volume 1, pages 187–195, 1956.

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

M Th Subbotin. On the law of frequency of error. *Matematicheskii Sbornik*, 31(2):296–301, 1923.

Aad Van Der Vaart and Jon A Wellner. A local maximal inequality under uniform entropy. *Electronic Journal of Statistics*, 5(2011):192, 2011.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Aad W Van Der Vaart and Jon Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 1996.

Constance Van Eeden. Efficiency-robust estimation of location. *The Annals of Mathematical Statistics*, 41(1):172–181, 1970.

Supplementary material to "Choosing the p in L_p loss: adaptive rates for symmetric mean estimation"

Appendix S1. Comparison with Generalized Gaussian MLE

S1.1. Comparison with the MLE

Recall from Remark 3 that for $\theta \in \mathbb{R}$ and $\sigma, \gamma > 0$, the generalized Gaussian distribution (also known as the Subbotin distribution) has a density of the form $p(x : \theta, \sigma, \gamma) = \frac{1}{2\sigma\Gamma(1+\gamma^{-1})} \exp\big(-\big|\frac{x-\theta}{\sigma}\big|^{\gamma}\big)$. We note that the uniform distribution on $[-\sigma, \sigma]$ is a limit point of the generalized Gaussian class where we let $\gamma \to \infty$.

Using univariate observations $Y_1, \ldots, Y_n$, we may then compute the MLE of γ with respect to the generalized Gaussian family:

$$\widehat{\gamma}_{\text{MLE}} = \arg\min_{\gamma} \min_{\theta,\sigma} \frac{1}{n} \sum_{i=1}^{n} \left| \frac{Y_i - \theta}{\sigma} \right|^{\gamma} + \log \sigma + \log \Gamma\left(1 + \frac{1}{\gamma}\right).$$

For any fixed γ, we may minimize over θ and σ to obtain that

$$\widehat{\gamma}_{\text{MLE}} = \arg\min_{\gamma} L_n(\gamma),$$

where

$$L_n(\gamma) := \frac{1}{\gamma} \log\left(\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} |Y_i - \theta|^{\gamma} \right) + \frac{1 + \log \gamma}{\gamma} + \log \Gamma\left(1 + \frac{1}{\gamma}\right).$$

A natural question then is how good is $\widehat{\gamma}_{\text{MLE}}$ as a selection procedure? Would the resulting location estimator $\widehat{\theta}_{\widehat{\gamma}_{\text{MLE}}}$ have good properties? If the density of Y_i belongs to the generalized Gaussian class, then we expect $\widehat{\gamma}_{\text{MLE}}$ to perform well. But when there is model misspecification, we show in this section that $\widehat{\gamma}_{\text{MLE}}$ performs suboptimally compared to the CAVS estimator that we propose in Section 2.3.

To start, let us define the population level likelihood function for every $\gamma > 0$

$$\begin{aligned} L(\gamma) &:= \min_{\theta,\sigma} \mathbb{E} \left| \frac{Y - \theta}{\sigma} \right|^{\gamma} + \log(2\sigma) + \log \Gamma\big(1 + \frac{1}{\gamma}\big) \\ &= \frac{1}{\gamma} \log\big(\min_{\theta} \mathbb{E}|Y - \theta|^{\gamma}\big) + \frac{1 + \log \gamma}{\gamma} + \log \Gamma\left(1 + \frac{1}{\gamma}\right). \end{aligned}$$

We define $L(\infty) := \lim_{\gamma \to \infty} L(\gamma)$ and $L_n(\infty) := \lim_{\gamma \to \infty} L_n(\gamma) = \log\big\{(Y_{(n)} - Y_{(1)})/2\big\}$. We note that if $\mathbb{E}|Y|^{\gamma} = \infty$, then $L(\gamma) = \infty$ and if Y is not compactly supported, then $L(\infty) = \infty$. Moreover, by Lemma S3.3 (in Section S3.2 of the appendix), we have that, for any fixed $\gamma \in \mathbb{R} \cup \{\infty\}$, we have that $L_n(\gamma) \overset{a.s.}{\to} L(\gamma)$.

Define $\gamma^*_{\text{MLE}} = \arg\min_{\gamma \geq 2} L(\gamma)$ as the minimizer of $L(\gamma)$. We show in the next Proposition that when the noise Z_i is supported on $[-1, 1]$ with a small but positive density value at the boundary, then $\gamma^*_{\text{MLE}} < \infty$ even though the optimal selection of γ is to take $\gamma \to \infty$ since the sample midrange $\widehat{\theta}_{\infty}$ would have a rate of convergence that is at least as fast as $\widetilde{O}(n^{-1})$.

Proposition S1.1 *Suppose $Y = Z + \theta_0$ where Z has a distribution symmetric around 0. Define $\gamma^*_{MLE} = \arg\min_{\gamma>0} L(\gamma)$.*

1. *If Z is supported on all of $\mathbb{R}$, then $\gamma^*_{MLE} < \infty$.*
2. *Suppose Z has a density p supported and continuous on $[-1, 1]$. Let $\gamma_E \approx 0.57721$ be the Euler–Mascheroni constant. If the density value at the boundary satisfies $p(1) < \frac{1}{2}e^{\gamma_E - 1}$, then $\gamma^*_{MLE} < \infty$.*
3. *Suppose Z has a density p supported and continuous on $[-1, 1]$. If the density value at the boundary satisfies $p(1) > \frac{1}{2}e^{\gamma_E - 1}$, then $\gamma = \infty$ is a local minimum of $L(\gamma)$.*

We relegate the proof of Proposition S1.1 to Section S4.3 of the Appendix.

If the noise density p is continuous and has boundary value $p(1) \in (0, \frac{1}{2}e^{\gamma_E - 1})$, then Proposition S1.1 suggests that we would not expect $\widehat{\gamma}_{\text{MLE}} \to \infty$. More precisely, we have that $L(\gamma^*_{\text{MLE}}) < L(\infty)$ and thus, by Lemma S3.3, when n is large enough, we also have $L_n(\gamma^*_{\text{MLE}}) < L_n(\infty)$ almost surely. Therefore, selecting γ by minimizing L_n would always favor a finite $\gamma = \gamma^*_{\text{MLE}}$ over $\gamma = \infty$. As a result, selecting γ based on MLE yields a suboptimal rate of $n^{-1/2}$.

In contrast, Theorem 10 shows that under the same setting, our proposed CAVS estimator selects a divergent $\widehat{\gamma}$ which can yield an error that is smaller than $\widetilde{O}(n^{-1/2})$ for $\widehat{\theta}_{\widehat{\gamma}}$. In fact, there are settings in which the density at the boundary is equal to zero, that is, $p(1) = 0$, where our proposed estimator can $\widehat{\theta}_{\widehat{\gamma}}$ have a rate of convergence that is faster than $n^{-1/2}$; for example, we see in that $|\widehat{\theta}_{\widehat{\gamma}} - \theta_0|$ is $\widetilde{O}(n^{-2/3})$ when the noise has the semicircle density.

We note that although Proposition S1.1 is stated for Z supported on $[-1, 1]$, by scale invariance of γ^*_{MLE}, Proposition S1.1 holds for support of the form $[-b, b]$, where the the condition on the density generalizes to $p(b) > \frac{1}{2b}e^{\gamma_0 - 1}$.

Remark S1.2 *Another drawback, one that is perhaps more alarming, of selecting γ based on the Generalized Gaussian likelihood is that the resulting location estimator may have a standard deviation (and hence error) that is larger than $O(n^{-1/2})$.*

Consider the following example: let p_1 be the density of $|W|^{\frac{1}{3}}\text{sign}(W)$, where W follows the standard Cauchy distribution, let $p_2(x) \propto \exp\big(-|x|^3\big)$, and let the noise Z have a mixture density $p = \delta p_1 + (1 - \delta)p_2$ for some $\delta \in (0, 1)$. We let $Y = Z + \theta_0$ as usual.

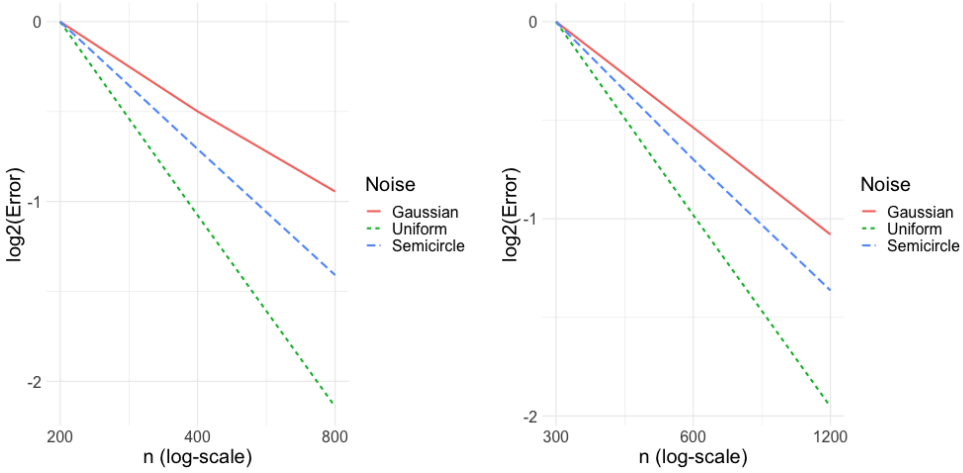
If $\delta = 0$ so that $Z \sim p_2$, then $L(\gamma)$ is minimized at $\gamma = 3$. We can thus pick a sufficiently small δ (see Lemma S5.17) such that the likelihood $L(\gamma)$ is minimized at $\gamma = 3 + \epsilon$ for some small $\epsilon > 0$. This however is a poor choice of γ since the asymptotic variance of $\widehat{\theta}_{3+\epsilon}$ is $V(3 + \epsilon) = \frac{\mathbb{E}|Z|^{4+2\epsilon}}{((2+\epsilon)\mathbb{E}|Z|^{1+\epsilon})^2} = \infty$ and moreover, we can show via a truncation argument that $\lim_{n\to\infty} \mathbb{P}(|\sqrt{n}(\widehat{\theta}_{3+\epsilon} - \theta_0)| < M) = 0$ for every $M > 0$.

In contrast, our proposed procedure would output the sample mean $\bar{Y} = \widehat{\theta}_2$, which has finite asymptotic variance. Intuitively, the CAVS procedure behaves better because it takes into account the higher moment $\mathbb{E}|Z|^{2(\gamma-1)}$ whereas the likelihood selector is based only on $\mathbb{E}|Z|^{\gamma}$.

Appendix S2. Empirical studies

We perform empirical studies on simulated data to verify our theoretical results in Section 3. We also analyze a dataset of NBA player statistics for the 2020-2021 season to show that our proposed CAVS estimator can be directly applied to real data.

S2.1. Simulations



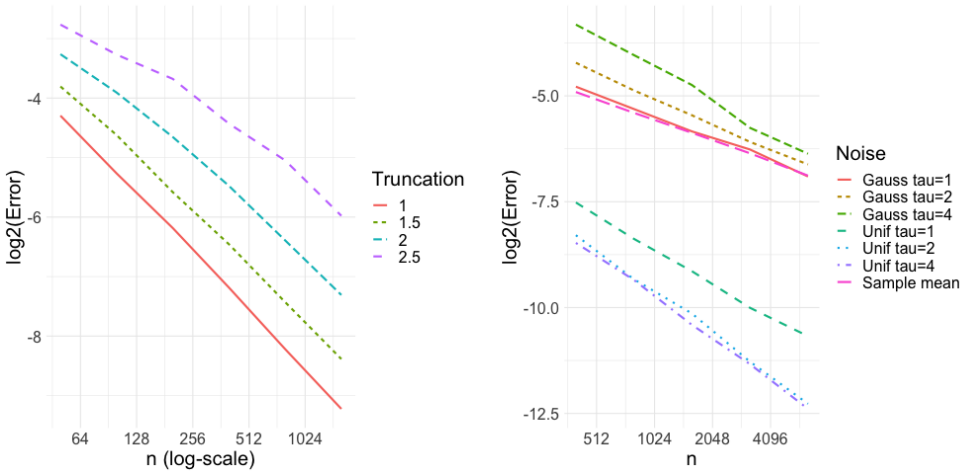
(*a*) Location estimation (*b*) Regression

Figure 3: Log-error vs. sample size plots. Sample size n is plotted on a log-scale.

Convergence rate for location estimation: Our first simulation takes the location estimation setting where $Y_i = \theta_0 + Z_i$ for $i = 1, \ldots, n$. We let the distribution of the noise Z_i be either Gaussian $N(0,1)$, uniform Unif$[-1,1]$, or semicircle (see Example 2). We let the sample size n vary between $(200, 400, 800)$. We compute our proposed CAVS estimator $\widehat{\theta}_{\widehat{\gamma}}$ (with $\tau = \sqrt{\log \frac{4n}{200}}$) and plot, in Figure 3(*a*), log-error versus the sample size n, where n is plotted on a logarithmic scale. Hence, a rate of convergence of n^{-t} would yield an error line of slope $-t$ in Figure 3(*a*). We normalize the errors so that all the lines have the same intercept. We see that error under uniform noise has a slope of -1, error under semicircle noise has a slope of $-2/3$, and error under Gaussian noise has a slope of $-1/2$ exactly as predicted by Theorem 9 and Theorem 10.

Convergence rate for regression: Then, we study the regression setting where $Y_i = X_i^\top \beta_0 + Z_i$ for $i = 1, 2, \ldots, n$. We let the distribution of the noise Z_i be either Gaussian $N(0,1)$, uniform Unif$[-1,1]$, or the semicircle density given in Example 2. We let the sample size n vary between $(200, 400, 600)$. We apply the regression version of the CAVS estimate $\widehat{\beta}_{\widehat{\gamma}}$ as described in Remark 6 (with $\tau = \sqrt{\log \frac{4n}{200}}$), and plot, in Figure 3(*a*), log-error versus the sample size n, where n is plotted on a logarithmic scale. We see that CAVS also has adaptive rate of convergence; the uniform noise yields a rate of n^{-1}, the semicircle noise yields a rate of $n^{-2/3}$, and the Gaussian noise yields a rate of $n^{-1/2}$ as n increases, as predicted by our theory.

Convergence rate for truncated Gaussian at different truncation levels: In Figure 4(*a*), we take the location model $Y_i = \theta_0 + Z_i$ where Z_i has the density $p_t(x) \propto \exp\{-\frac{1}{2}\frac{x^2}{\sigma_t^2}\}\mathbb{1}(|x| \le t/\sigma_t)$ for some $t > 0$ and where $\sigma_t > 0$ is chosen so that Z_i always has unit variance. In other words, we sample Z_i by first generating $W \sim N(0,1)$, keep W only if $|W| \le t$, and then take $Z_i = \sigma_t W$ where $\sigma_t > 0$ is chosen so that $\mathrm{Var}(Z_i) = 1$. We use four different truncation levels $t = 1, 1.5, 2, 2.5$; we let the sample size vary from $n = 50$ to $n = 1600$ and compute our CAVS



(*a*) Gaussian truncated at different levels (*b*) Different τ

Figure 4: Log-error vs. sample size plots. Sample size n is plotted on a log-scale.

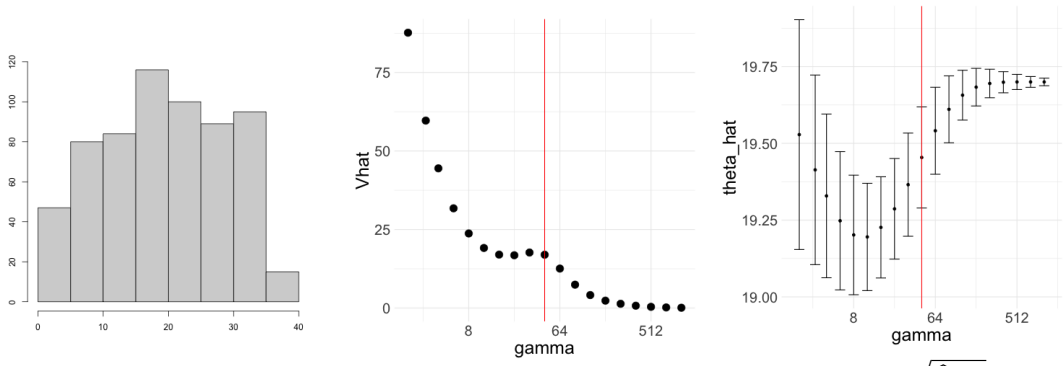
estimate $\widehat{\theta}_{\widehat{\gamma}}$ (with $\tau = \sqrt{\log \frac{4n}{50}}$). We plot in Figure 4(*a*), the log-error versus the sample size n, where n is plotted on a logarithmic scale. We observe that when the truncation level is $t = 1$ or 1.5 or 2, the error is of order n^{-1}. When the truncation level is $t = 2.5$, the error behaves like $n^{-1/2}$ for small n but transitions to n^{-1} when n becomes large. This is not surprising since, when n is small, it is difficult to know whether the Z_i's are drawn from $N(0,1)$ or drawn from truncated Gaussian with a large truncation level.

Convergence rate for different τ: In Figure 4(*b*), we take the location model $Y_i = \theta_0 + Z_i$ and take Z_i to be either Gaussian $N(0,1)$ or uniform Unif$[-M, M]$ where $M > 0$ is chosen so that Z_i has unit variance. We then apply our proposed CAVS procedure for different levels of τ, ranging from $\tau \in \{1, 2, 4\}$. We let the sample size vary from $n = 400$ to $n = 6400$ and plot the log-error versus the sample size n, where n is plotted on a logarithmic scale. For comparison, we also plot the error of the sample mean $\bar{Y}$, which does not depend on the distribution of Z_i since we scale Z_i to have unit variance in both settings. We observe in Figure 4(*b*) that when $\tau = 1$, the CAVS estimate $\widehat{\theta}_{\widehat{\gamma}}$ basically coincides with the sample mean if $Z_i \sim N(0,1)$ but has much less error when Z_i is uniform. As we increase τ, CAVS estimator has increased error under the Gaussian setting when $Z_i \sim N(0,1)$ since we select $\widehat{\gamma} > 2$ more often; under the uniform setting, it has less error. Based on these studies, we recommend $\tau = 1$ in practice as a conservative choice.

S2.2. Real data experiments

Uniform or truncated Gaussian data are not ubiquitous but they do appear in real world datasets. In this section, we use the CAVS location estimation and regression procedure to analyze a dataset of 626 NBA players in the 2020–2021 season. We consider variables AGE, MPG (average minutes played per game), and GP (games played).

Both MPG and GP variables are compactly supported. They also do not exhibit clear signs of asymmetry; MPG has an empirical skewness of -0.064 and GP has an empirical skewness of 0.013. We apply the CAVS procedure to both with $\tau = 1$ and we obtain $\widehat{\gamma} = 32$ for MPG variable and



(*a*) Histogram of MPG (*b*) $\widehat{V}(\gamma)$ (*c*) $\widehat{\theta}_\gamma \pm \sqrt{\frac{\widehat{V}(\gamma)}{n}}$

Figure 5: Analysis on MPG (average minutes played per game) in the NBA 2021 data

$\widehat{\gamma} = 2048$ for the GP variable. In contrast, the AGE variable has a skewness of 0.56 and when we apply CAVS procedure (still with $\tau = 1$), we obtain $\widehat{\gamma} = 2$. These results suggest that CAVS can be useful for practical data analysis.

Moreover, we also study the CAVS regression method by considering two regression models:

$$(\text{MODEL 1})\ \text{MPG} \sim \text{GP} + \text{AGE} + \text{W}, \qquad (\text{MODEL 2})\ \text{MPG} \sim \text{AGE} + \text{W},$$

where W is an independent Gaussian feature add so that we can assess how close the estimated coefficient $\widehat{\beta}_{\text{W}}$ is to zero to gauge the estimation error. We estimate $\widehat{\beta}_{\widehat{\gamma}}$ on 100 randomly chosen training data points and report the predictive error on the remaining test data points; we also report the average value of $|\widehat{\beta}_{\text{W}}|$, which we would like to be as close to 0 as possible. We perform 1000 trials of this experiment (choosing random training set in each trial) and report the performance of CAVS versus OLS estimator in Table 1.

	Model 1 Pred. Error	Model 1 $\lvert\widehat{\beta}_{\text{W}}\rvert$	Model 2 Pred. Error	Model 2 $\lvert\widehat{\beta}_{\text{W}}\rvert$
CAVS	**0.686**	**0.045**	**0.95**	**0.082**
OLS	0.689	0.140	1.04	0.205

Table 1: Comparison of CAVS vs. OLS on two simple regression models.

Appendix S3. Supplementary material for Section 2

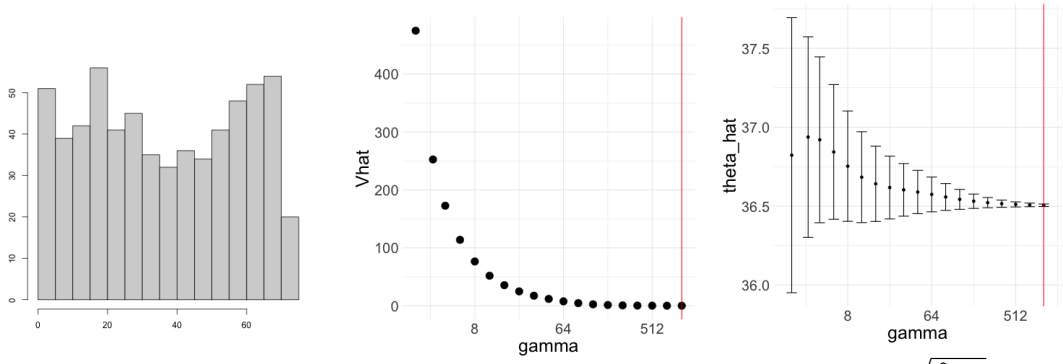
S3.1. Proof of Lemma 1

Proof (of Lemma 1)

First, we observe that if $4\frac{\log n}{\gamma} \geq 1$, then, by the fact that $\widehat{\theta}_\gamma \in [X_{(1)}, X_{(n)}]$, we have that

$$|\widehat{\theta}_\gamma - X_{\text{mid}}| \leq \frac{1}{2}(X_{(n)} - X_{(1)}) \leq 2(X_{(n)} - X_{(1)})\frac{\log n}{\gamma}.$$

Therefore, we assume that $4\frac{\log n}{\gamma} \leq 1$.



(*a*) Histogram of GP (*b*) $\widehat{V}(\gamma)$ (*c*) $\widehat{\theta}_\gamma \pm \sqrt{\frac{\widehat{V}(\gamma)}{n}}$

Figure 6: Analysis on GP (games played) in the NBA 2021 data

We apply Lemma S3.4 with $f(\theta) = \left\{\frac{1}{n}\sum_{i=1}^n |X_i - \theta|^\gamma\right\}^{1/\gamma}$ and $g(\theta) = \max_i |X_i - \theta|$ so that $\theta_g := \arg\min g(\theta) = \widehat{\theta}_{\text{mid}}$ and $\theta_f := \arg\min f(\theta) = \widehat{\theta}_\gamma$. Fix any $\delta > 0$. We observe that

$$g(\theta_g) = \max_i |X_i - X_{\text{mid}}| = \frac{X_{(n)} - X_{(1)}}{2}$$
$$g(\theta_g + \delta) = \frac{X_{(n)} - X_{(1)}}{2} + \delta, \qquad \text{and } g(\theta_g - \delta) = \frac{X_{(n)} - X_{(1)}}{2} + \delta.$$

Therefore, for $\theta \in \{\theta_g - \delta, \theta_g + \delta\}$, we have that $\frac{1}{2}(g(\theta) - g(\theta_g)) = \frac{\delta}{2}$. On the other hand, by the fact that

$$\left\{\frac{1}{n}\sum_{i=1}^n |X_i - \theta|^\gamma\right\}^{\frac{1}{\gamma}} \geq n^{-\frac{1}{\gamma}} \max_{i\in[n]} |X_i - \theta|,$$

we have that

$$g(\theta) \geq f(\theta) \geq n^{-\frac{1}{\gamma}} g(\theta) \quad \forall \theta \in \mathbb{R}. \tag{S3.1}$$

Therefore, for $\theta \in \{\theta_g - \delta, \theta_g, \theta_g + \delta\}$, we have that

$$\begin{aligned} |f(\theta) - g(\theta)| = g(\theta) - f(\theta) &\leq (1 - n^{-\frac{1}{\gamma}}) g(\theta) \\ &\leq \frac{\log n}{\gamma}\left(\frac{X_{(n)} - X_{(1)}}{2} + \delta\right). \end{aligned}$$

Using our assumption that $4\frac{\log n}{\gamma} \leq 1$, we have that for any $\delta \geq 2(X_{(n)} - X_{(1)})\frac{\log n}{\gamma}$ and any $\theta \in \{\theta_g - \delta, \theta_g, \theta_g + \delta\}$,

$$|f(\theta) - g(\theta)| \leq \frac{\log n}{\gamma}\left(\frac{X_{(n)} - X_{(1)}}{2} + \delta\right) \leq \frac{\delta}{2} = \frac{1}{2}(g(\theta) - g(\theta_g)).$$

The Lemma thus immediately follows from Lemma S3.4. ■

S3.2. Lemma S3.3 on the convergence of $\widehat{V}(\gamma)$

The following lemma implies that, for a fixed γ such that $V(\gamma)$ is well-defined, our asymptotic variance estimator $\widehat{V}(\gamma)$ is consistent. For a random variable Y, we define its essential supremum to be

$$\text{ess-sup}(Y) := \inf\big\{M \in \mathbb{R} : \mathbb{P}(Y \le M) = 1\big\},$$

where the infimum of an empty set is taken to be infinity. Note that ess-sup$(|Y|) < \infty$ if and only if Y is compacted supported and that $\lim_{\gamma\to\infty}\{\mathbb{E}|Y|^\gamma\}^{\frac{1}{\gamma}} = \text{ess-sup}(|Y|)$.

We may define ess-inf(Y) is the same way. For an infinite sequence $Y_1, Y_2, \ldots$ of independent and identically distributed random variables, it is straightforward to show that $Y_{(n),n} := \max_{i\in[n]} Y_i \overset{\text{a.s.}}{\to} \text{ess-sup}(Y)$ and $Y_{(1),n} := \min_{i\in[n]} Y_i \overset{\text{a.s.}}{\to} \text{ess-inf}(Y)$ regardless of whether the essential supremum and infimum are finite or not.

Lemma S3.3 *Let $Y_1, Y_2, \ldots$ be a sequence of independent and identically distributed random variables and let $\gamma > 1$. The following hold:*

1. *If $\mathbb{E}|Y|^\gamma < \infty$, then $\min_{\theta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n |Y_i - \theta|^\gamma \overset{a.s.}{\to} \min_{\theta\in\mathbb{R}} \mathbb{E}|Y-\theta|^\gamma$.*
2. *If $\mathbb{E}|Y|^\gamma = \infty$, then $\min_{\theta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n |Y_i - \theta|^\gamma \overset{a.s.}{\to} \infty$.*
3. *If Y is compactly supported, then we have that $\min_{\theta\in\mathbb{R}} \max_{i\le n} |Y_i - \theta| = (Y_{(n),n} - Y_{(1),n})/2 \overset{a.s.}{\to} \min_\theta$ ess-sup$(|Y-\theta|)$.*
4. *If ess-sup$(|Y|) = \infty$, then $\min_{\theta\in\mathbb{R}} \max_{i\le n} |Y_i - \theta| \overset{a.s.}{\to} \infty$.*

As a direct consequence, for any $\gamma > 1$ such that $\mathbb{E}|Y|^{\gamma-2} < \infty$, we have $\widehat{V}(\gamma) \overset{a.s.}{\to} V(\gamma)$, even when $V(\gamma) = \infty$.

Proof (of Lemma S3.3)

For the first claim, we apply Proposition S3.5 with $g(y,\theta) = |y-\theta|^\gamma$ and $\psi(\theta) = \mathbb{E}|Y-\theta|^\gamma$ and immediately obtain the desired conclusion.

We now prove the second claim by a truncation argument. Suppose $\mathbb{E}|Y|^\gamma = \infty$ so that $\min_\theta \mathbb{E}|Y-\theta|^\gamma = \infty$. Fix $M > 0$ arbitrarily. We claim there then exists $\tau > 0$ such that

$$\min_{\theta\in\mathbb{R}} \mathbb{E}\big[|Y-\theta|^\gamma \mathbb{1}\{|Y| \le \tau\}\big] > M.$$

To see this, for any $\tau > 0$, define $\theta_\tau = \arg\min_{\theta\in\mathbb{R}} \mathbb{E}\big[|Y-\theta|^\gamma \mathbb{1}\{|Y| \le \tau\}\big]$. The argmin is well-defined since $\theta \mapsto \mathbb{E}\big[|Y-\theta|^\gamma \mathbb{1}\{|Y| \le \tau\}\big]$ is strongly convex and goes to infinity as $|\theta| \to \infty$. If $\{\theta_\tau\}_{\tau=1}^\infty$ is bounded, then the claim follows because $\mathbb{E}|Y-\theta_\tau|^\gamma \mathbb{1}\{|Y| \le \tau\} \ge \{(\mathbb{E}|Y|^\gamma \mathbb{1}\{|Y| \le \tau\})^{1/\gamma} - \theta_\tau\}^\gamma$. If $\{\theta_\tau\}_{\tau=1}^\infty$ is unbounded, then there exists a sub-sequence τ_m such that $\lim_{m\to\infty} \theta_{\tau_m} \to \infty$ say. For any $a > 0$ such that $\mathbb{P}(|Y| \le a) > 0$, we have $\lim_{m\to\infty} \mathbb{E}|Y-\theta_{\tau_m}|^\gamma \mathbb{1}\{|Y| \le \tau_m\} \ge \lim_{m\to\infty} |a-\theta_{\tau_m}|^\gamma \mathbb{P}(|Y| \le a) = \infty$. Therefore, in either cases, our claim holds.

Using Proposition S3.5 again with $g(x,\theta) = |x-\theta|^\gamma \mathbb{1}\{|x| \le \tau\}$, we have that

$$\min_\theta \frac{1}{n}\sum_{i=1}^n |Y_i-\theta|^\gamma \mathbb{1}\{|Y_i| \le \tau\} \overset{a.s.}{\to} \min_\theta \mathbb{E}\big[|Y-\theta|^\gamma \mathbb{1}\{|Y| \le \tau\}\big] > M.$$

In other words, there exists an event $\widetilde{\Omega}_M$ with probability 1 such that, for any $\omega \in \widetilde{\Omega}_M$, there exists n_ω such that for all $n \geq n_\omega$,

$$\min_\theta \frac{1}{n}\sum_{i=1}^n |Y_i - \theta|^\gamma \geq \min_\theta \frac{1}{n}\sum_{i=1}^n |Y_i - \theta|^\gamma \mathbb{1}\{|Y_i| \leq \tau\} \geq M/2.$$

Thus, on $\widetilde{\Omega}_M$, we have that

$$\liminf_{n\to\infty} \min_\theta \frac{1}{n}\sum_{i=1}^n |Y_i - \theta|^\gamma > M/2.$$

Thus, on the event $\widetilde{\Omega} = \cap_{M=1}^\infty \widetilde{\Omega}_M$, we have that $\min_\theta \frac{1}{n}\sum_{i=1}^n |Y_i - \theta|^\gamma \to \infty$. Since $\widetilde{\Omega}$ has probability 1, the second claim follows. For the third claim, without loss of generality, we can assume that ess-sup$(Y) = 1$ and ess-inf$(Y) = -1$. Define $X_n = (Y_{(n),n} - Y_{(1),n})/2$, then we have $X_n \leq \min_\theta$ ess-sup$(|Y - \theta|) = 1$ and

$$\mathbb{P}\{X_n < 1 - \delta\} \leq \mathbb{P}\{Y_{(n),n} < 1 - \delta\} + \mathbb{P}\{Y_{(1),n} > -1 + \delta\},$$

where, as $n \to \infty$, the right hand side tends to 0 for every $\delta > 0$. X_n thus converges to 1 in probability. Since the collection $\{X_n\}_{n=1}^\infty$ is defined on the same infinite sequence $\{Y_1, Y_2, \ldots\}$ of independent and identically distributed random variables, we have that $1 \geq X_n \geq X_{n-1} \geq 0$ so that $X_n \overset{a.s.}{\to} 1$ by the monotone convergence theorem.

For the forth claim, suppose without loss of generality that ess-inf$(Y) \leq -1$ and that ess-sup$(Y) = \infty$. Let $X_n = (Y_{(n),n} - Y_{(1),n})/2$ as with the proof of the third claim. Then,

$$\mathbb{P}\{X_n < M\} \leq \mathbb{P}\{Y_{(n),n} < 2M\} + \mathbb{P}\{Y_{(1),n} \geq 0\}.$$

Since the right hand side tends to 0 for every $M > 0$, we have that X_n converges to infinity almost surely. The Lemma follows as desired. ■

S3.3. Bound on V

The following lower bound on $V(\gamma)$ holds regardless of whether Y is symmetric around θ_0 or not. We have

$$\begin{aligned}
V(\gamma) &= \frac{\mathbb{E}|Y - \theta_0|^{2(\gamma-1)}}{(\gamma-1)^2\{\mathbb{E}|Y - \theta_0|^{\gamma-2}\}^2} \\
&= \frac{\mathbb{E}|Y - \theta_0|^{2(\gamma-1)}}{\{\mathbb{E}|Y - \theta_0|^{\gamma-1}\}^2}\left(\frac{\mathbb{E}|Y - \theta_0|^{\gamma-1}}{\mathbb{E}|Y - \theta_0|^{\gamma-2}}\right)^2 \frac{1}{(\gamma-1)^2} \\
&\geq \frac{\mathbb{E}|Y - \theta_0|^{2(\gamma-1)}}{\{\mathbb{E}|Y - \theta_0|^{\gamma-1}\}^2}\{\mathbb{E}|Y - \theta_0|^{\gamma-2}\}^{\frac{2}{\gamma-2}} \frac{1}{(\gamma-1)^2} \\
&\geq \frac{\mathbb{E}|Y - \theta_0|^{2(\gamma-1)}}{\{\mathbb{E}|Y - \theta_0|^{\gamma-1}\}^2}\mathbb{E}|Y - \theta_0|^2 \frac{1}{(\gamma-1)^2},
\end{aligned}$$

where the first inequality follows from the fact that $\mathbb{E}|Y - \theta_0|^{\gamma-1} \geq \left(\mathbb{E}|Y - \theta_0|^{\gamma-2}\right)^{\frac{\gamma-1}{\gamma-2}}$. In particular, we have that $V(\gamma) \geq \frac{\mathbb{E}|Y-\theta_0|^2}{(\gamma-1)^2}$. Equality is attained when $Y - \theta_0$ is a Rademacher random variable.

S3.4. Optimization algorithm

We give the Newton's method algorithm for computing $\widehat{\theta}_\gamma = \arg\min_{\theta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n |Y_i - \theta|^\gamma$. It is important to note that to avoid numerical precision issues when γ is large, we have to transform the input $Y_1, \ldots, Y_n$ so that they are supported on the unit interval $[-1, 1]$.

Algorithm 2 Newton's method for location estimation

Input: observations $Y_1, \ldots, Y_n \in \mathbb{R}$ and $\gamma \geq 2$.
Output: $\widehat{\theta}_\gamma := \min_{\theta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n |Y_i - \theta|^\gamma$.
Compute $S = (Y_{(n)} - Y_{(1)})$ and $M = (Y_{(n)} + Y_{(1)})/2$ and transform $Y_i \leftarrow 2(Y_i - M)/S$.
Initialize $\theta^{(0)} = 0$.
for $t = 1, 2, 3, \ldots$ **do**
  Compute $f' = -\frac{1}{n}\sum_{i=1}^n |Y_i - \theta^{(t-1)}|^{\gamma-1}\text{sign}(Y_i - \theta^{(t-1)})$
  Compute $f'' = \frac{\gamma-1}{n}\sum_{i=1}^n |Y_i - \theta^{(t-1)}|^{\gamma-2}$.
  Set $\theta^{(t)} = \theta^{(t-1)} - \frac{f'}{f''}$.
  If $|f'| \leq \varepsilon$, break and output $S\theta^{(t)}/2 + M$.
end

To compute $\widehat{\theta}_\gamma$ for a collection of $\gamma_1 < \gamma_2 < \ldots$, we can warm start our optimization of $\widehat{\theta}_{\gamma_2}$ by initializing with $\widehat{\theta}_{\gamma_1}$. In the regression setting where γ is large, we find that it improves numerical stability to to apply a quasi-Newton's method where we add a an identity εI to the Hessian for a small $\varepsilon > 0$.

S3.5. Supporting Lemmas

Lemma S3.4 *Let $f, g : \mathbb{R}^d \to \mathbb{R}$ and suppose f is convex. Let $x_g \in \arg\min g(x)$ and $x_f \in \arg\min f(x)$. Suppose there exists $\delta > 0$ such that*

$$|f(x) - g(x)| \vee |f(x_g) - g(x_g)| < \frac{1}{2}(g(x) - g(x_g)), \quad \textit{for all } x \textit{ s.t. } \|x - x_g\| = \delta.$$

Then, we have that

$$\|x_f - x_g\| \leq \delta.$$

Proof

Let $\delta > 0$ and suppose δ satisfies the condition of the Lemma. Fix $x \in \mathbb{R}^d$ such that $\|x - x_g\| > \delta$. Define $\xi = x_g + \frac{\delta}{\|x-x_g\|}(x - x_g)$ so that $\|\xi - x_g\| = \delta$. Note by convexity of f that $f(\xi) \leq (1 - \frac{\delta}{\|x-x_g\|})f(x_g) + \frac{\delta}{\|x-x_g\|}f(x)$.

Therefore, we have that

$$\begin{aligned}\frac{\delta}{\|x - x_g\|}(f(x) - f(x_g)) &\geq f(\xi) - f(x_g) \\ &= f(\xi) - g(\xi) + g(\xi) - g(x_g) + g(x_g) - f(x_g) > 0\end{aligned}$$

under the condition of the Theorem. Therefore, we have $f(x) > f(x_g)$ for any x such that $\|x - x_g\| > \delta$. The conclusion of the Theorem follows as desired. ■

S3.5.1. LLN FOR MINIMUM OF A CONVEX FUNCTION

Proposition S3.5 *Suppose $\theta \mapsto g(y,\theta)$ is convex on $\mathbb{R}$ for all $y \in \mathcal{Y}$. Define $\psi(\theta) := \mathbb{E}g(Y,\theta)$ and suppose ψ is finite on an open subset of $\mathbb{R}$ and $\lim_{|\theta|\to\infty} \psi(\theta) = \infty$.*

Then, we have that

$$\min_{\theta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} g(Y_i,\theta) \overset{a.s.}{\to} \min_{\theta\in\mathbb{R}} \mathbb{E}g(Y,\theta),$$

and

$$\sup_{\theta_1\in\Theta_n} \min_{\theta_2\in\Theta_0} |\theta_1-\theta_2| \overset{a.s.}{\to} 0,$$

where $\Theta_n := \arg\min_{\theta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} g(Y_i,\theta)$ and $\Theta_0 := \arg\min_{\theta\in\mathbb{R}} \mathbb{E}g(Y,\theta)$

Proof Define $\widehat{\psi}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} g(Y_i,\theta)$ and observe that $\widehat{\psi}_n$ is a convex function on $\mathbb{R}$. We also observe that $\arg\min \psi$ is a closed bounded interval on $\mathbb{R}$ and we define θ_0 to be its midpoint.

Fix $\epsilon > 0$ arbitrarily. We may then choose $\theta_L \in (-\infty,\theta_0)$ and $\theta_R \in (\theta_0,\infty)$ such that

1. $\psi(\theta_L) > \psi(\theta_0)$ and $\psi(\theta_R) > \psi(\theta_0)$,
2. $(\psi(\theta_L)-\psi(\theta_0)) \vee (\psi(\theta_R)-\psi(\theta_0)) \le \epsilon$,
3. $\theta_0-\theta_L = \theta_R-\theta_0$,
4. and $\min_{\theta\in\Theta_0}|\theta_R-\theta| \vee \min_{\theta\in\Theta_0}|\theta_L-\theta| < \epsilon$.

Define $\widetilde{\epsilon} := (\psi(\theta_L)-\psi(\theta_0)) \wedge (\psi(\theta_R)-\psi(\theta_0))$ and note that $0 < \widetilde{\epsilon} < \epsilon$ by our choice of θ_L and θ_R. By LLN, there exists an event $\widetilde{\Omega}_\epsilon$ with probability 1 such that, for every $\omega \in \widetilde{\Omega}_\epsilon$, there exists $n_\omega \in \mathbb{N}$ where for all $n \ge n_\omega$,

$$|\widehat{\psi}_n(\theta_L)-\psi(\theta_L)| \vee |\widehat{\psi}_n(\theta_R)-\psi(\theta_R)| \vee |\widehat{\psi}_n(\theta_0)-\psi(\theta_0)| \le \widetilde{\epsilon}/3.$$

Fix any $\omega \in \widetilde{\Omega}_\epsilon$ and fix $n \ge n_\omega$, we have that $\widehat{\psi}_n(\theta_L) \ge \psi(\theta_L) - \widetilde{\epsilon}/3 > \psi(\theta_0)$ and likewise for $\widehat{\psi}_n(\theta_R)$. Thus, $\widehat{\psi}_n$ must attain its minimum in the interval (θ_L,θ_R), i.e., $\sup_{\theta_1\in\Theta_n}\min_{\theta_2\in\Theta_0}|\theta_1-\theta_2| < \epsilon$. We then have by Lemma S3.6 that

$$\begin{aligned}\min_{\theta\in\mathbb{R}} \widehat{\psi}(\theta) &= \min_{\theta\in(\theta_L,\theta_R)} \widehat{\psi}_n(\theta) \\ &\ge \widehat{\psi}_n(\theta_0) - |\widehat{\psi}_n(\theta_0)-\widehat{\psi}_n(\theta_R)| \vee |\widehat{\psi}_n(\theta_0)-\widehat{\psi}_n(\theta_L)| \\ &\ge \psi(\theta_0) - \widetilde{\epsilon} - \epsilon \ge \psi(\theta_0) - 2\epsilon.\end{aligned}$$

On the other hand,

$$\min_{\theta\in\mathbb{R}} \widehat{\psi}_n(\theta) \le \widehat{\psi}_n(\theta_0) \le \psi(\theta_0) + \epsilon.$$

Therefore, for all $\omega \in \widetilde{\Omega}_\epsilon$, we have that

$$\limsup_{n\to\infty} \left|\min_{\theta\in\mathbb{R}} \widehat{\psi}_n(\theta) - \psi(\theta_0)\right| \le 2\epsilon,$$

and

$$\limsup_{n\to\infty} \sup_{\theta_1\in\Theta_n} \min_{\theta_2\in\Theta_0} |\theta_1-\theta_2| < \epsilon.$$

We then define $\widetilde{\Omega} := \cap_{k=1}^{\infty} \widetilde{\Omega}_{1/k}$ and observe that $\widetilde{\Omega}$ has probability 1 and that on $\widetilde{\Omega}$,

$$\lim_{n\to\infty} \left| \min_{\theta\in\mathbb{R}} \widehat{\psi}_n(\theta) - \psi(\theta_0) \right| = 0,$$

and

$$\lim_{n\to\infty} \sup_{\theta_1\in\Theta_n} \min_{\theta_2\in\Theta_0} |\theta_1 - \theta_2| = 0.$$

The Proposition follows as desired.

■

Lemma S3.6 *Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function. For any $x_0 \in \mathbb{R}$, $x_L \in (-\infty, x_0)$ and $x_R \in (x_0, \infty)$, we have*

$$\begin{aligned} &\textit{for all } x \in (x_L, x_0),\ f(x) \geq f(x_0) + \{f(x_R) - f(x_0)\}\frac{x - x_0}{x_R - x_0} \\ &\textit{for all } x \in (x_0, x_R),\ f(x) \geq f(x_0) + \{f(x_0) - f(x_L)\}\frac{x - x_0}{x_0 - x_L}. \end{aligned}$$

As a direct consequence, if $x_0 - x_L = x_R - x_0$, then we have that for all $x \in (x_L, x_0)$,

$$f(x) \geq f(x_0) - |f(x_0) - f(x_R)|$$

and that for all $x \in (x_0, x_R)$,

$$f(x) \geq f(x_0) - |f(x_0) - f(x_L)|.$$

Proof Let $x \in (x_L, x_0)$; using the fact that $f'(x_0) \leq \frac{f(x_R)-f(x_0)}{x_R-x_0}$, we have

$$\begin{aligned} f(x) &\geq f(x_0) + f'(x_0)(x - x_0) \\ &\geq f(x_0) + \{f(x_R) - f(x_0)\}\frac{x - x_0}{x_R - x_0}. \end{aligned}$$

Likewise, for $x \in (x_0, x_R)$, we have $f'(x_0) \geq \frac{f(x_0)-f(x_L)}{x_0-x_L}$ and hence,

$$\begin{aligned} f(x) &\geq f(x_0) + f'(x_0)(x - x_0) \\ &\geq f(x_0) + \{f(x_0) - f(x_L)\}\frac{x - x_0}{x_0 - x_L}. \end{aligned}$$

■

Appendix S4. Supplementary material for Section 3

S4.1. Proof of Theorem 10

Structure of intermediate results: The proof is long and uses various intermediate technical results. The key intermediate theorems are (1) Theorem S4.7 which is essentially a corollary of Proposition S4.8 and (2) Theorem S4.9 which follow from Proposition S4.10 as well as Theorem S4.7.

Notation for constants: For all the proofs in this section, we let C indicate a generic universal constant whose value could change from instance to instance. We let C_1, C_2, C_3, C_4 be specific universal constants where C_1, C_2 are defined in the proof of Proposition S4.8 and where C_3, C_4 are defined in Theorem S4.9.

Proof (of Theorem 10)

We first prove the following: assume that n is large enough such that

$$\tau \geq \frac{C_1\sqrt{C_4}a_2^{3/2}}{a_1^{3/2}}\sqrt{\log\log n}, \quad \text{and that}$$
$$\left\{\frac{1}{C_{1\alpha}\vee C_{2\alpha}\vee C_4}\left(\frac{c_0^2 a_1^6}{a_2^3}\alpha^2\right)\frac{n}{\log n}\right\}^{\frac{1}{\alpha}} \geq e^{C_1\frac{a_2}{a_1}} \geq 2, \tag{S4.2}$$

where C_1, C_4 are universal constants and $C_{1\alpha}, C_{2\alpha}$ are constants depending only on α – the value of these are specified in Theorem S4.7 and Theorem S4.9.

We claim that

$$\mathbb{P}\left\{|\widehat{\theta}_{\widehat{\gamma}} - \theta_0| \leq C_{a_1,a_2,\alpha}\left(\frac{\log^{1+\frac{1}{\alpha}} n}{n^{\frac{1}{\alpha}}} \vee \frac{\log n}{M_n}\right)\right\} \geq 1 - \frac{4}{n^{\frac{1}{\alpha}}} - \exp(-\frac{1}{\alpha}(\tau \wedge \sqrt{\log n})\sqrt{\log n}). \tag{S4.3}$$

This immediately proves the first claim of the theorem. To see that the second claim of the theorem also holds, note that if (S4.3) holds and if $\tau \geq \sqrt{\log n}$, then, by inflating the constant $C_{a_1,a_2,\alpha}$ if necessary, we have that, for all $n \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}|\widehat{\theta}_{\widehat{\gamma}} - \theta_0| &\leq C_{a_1,a_2,\alpha}\left(\frac{\log^{1+\frac{1}{\alpha}} n}{n^{\frac{1}{\alpha}}} \vee \frac{\log n}{M_n}\right) + \frac{7}{n^{1/\alpha}} \\ &\leq C_{a_1,a_2,\alpha}\left(\frac{\log^{1+\frac{1}{\alpha}} n}{n^{\frac{1}{\alpha}}} \vee \frac{\log n}{M_n}\right),\end{aligned}$$

where the first inequality uses the fact that $|\widehat{\theta}_{\widehat{\gamma}} - \theta_0| \leq 1$. The desired conclusion would then immediately follow.

We thus prove (S4.3) under assumption (S4.2). To that end, let $c_0 = 2^{-8}$ and define $\gamma_u = \left\{\frac{1}{C_{1\alpha}\vee C_{2\alpha}\vee C_4}\left(\frac{c_0^2 a_1^6}{a_2^3}\alpha^2\right)\frac{n}{\log n}\right\}^{\frac{1}{\alpha}}$ and note that $\gamma_u \geq e^{C_1\frac{a_2}{a_1}} \geq 2$ under assumption (S4.2).

Let C_4 be a sufficiently large universal constant as defined in Theorem S4.9 and define the event

$$\mathcal{E}_1 := \left\{\frac{1}{C_4}\frac{a_1}{a_2^2}\gamma^{\alpha-2} \leq \widehat{V}(\gamma) \leq C_4\frac{a_2}{a_1^2}\gamma^{\alpha-2}, \quad \text{for all } \gamma \in [2, (\gamma_u+1)/2]\right\}, \tag{S4.4}$$

It holds by Theorem S4.9 that $\mathbb{P}(\mathcal{E}_1) \geq 1 - 2n^{-\frac{1}{\alpha}}$.

Now define $\tau' = \frac{1}{\sqrt{C_4}} \frac{\sqrt{a_1}}{a_2} \tau$ and note that $\tau' \geq \frac{C_1 \sqrt{a_2}}{a_1} \sqrt{\log \log n}$ under assumption (S4.2). Define the event

$$\mathcal{E}_2 := \left\{ |\widehat{\theta}_\gamma - \theta_0| \leq \tau' \sqrt{\frac{\gamma^{\alpha-2}}{n}}, \text{ for all } \gamma \in [2, \gamma_u] \right\}. \tag{S4.5}$$

Then we have by Theorem S4.7 that

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_2^c) &\leq \exp\Big\{ -\frac{a_1^2}{C_2 a_2} (\tau' \wedge \sqrt{\log n}) \sqrt{\frac{n}{\gamma_u^\alpha}} \Big\} \\
&\leq \exp\Big\{ -\frac{1}{\alpha} \frac{\sqrt{C_4} a_2}{\sqrt{a_1}} (\tau' \wedge \sqrt{\log n}) \sqrt{\log n} \Big\} \\
&\leq \exp\Big\{ -\frac{1}{\alpha} (\tau \wedge \sqrt{\log n}) \sqrt{\log n} \Big\}.
\end{aligned}$$

On the event $\mathcal{E}_1 \cap \mathcal{E}_2$, we have that, for all $\gamma \in [2, (\gamma_u + 1)/2]$,

$$|\widehat{\theta}_\gamma - \theta_0| \leq \tau' \sqrt{\frac{\gamma^{\alpha-2}}{n}} \leq \tau' \sqrt{C_4} \frac{a_2}{\sqrt{a_1}} \sqrt{\frac{\widehat{V}(\gamma)}{n}} \leq \tau \sqrt{\frac{\widehat{V}(\gamma)}{n}}.$$

Therefore, we have that

$$\theta_0 \in \bigcap_{\gamma \in \mathcal{N}_n, \gamma \leq (\gamma_u+1)/2} \left[\widehat{\theta}_\gamma - \tau \sqrt{\frac{\widehat{V}(\gamma)}{n}}, \widehat{\theta}_\gamma + \tau \sqrt{\frac{\widehat{V}(\gamma)}{n}} \right].$$

Since $\mathcal{N}_n$ contains $\{2^k : k \leq \log_2 M_n\}$, either $\frac{\gamma_u+1}{2} \geq M_n$ or there exists $\gamma \in \mathcal{N}_n$ such that $\gamma \geq \frac{\gamma_u+1}{4}$. In either case, it holds by the definition of $\gamma_{\max}$ that $\gamma_{\max} \geq \frac{\gamma_u+1}{4} \wedge M_n$. Write $\widetilde{\gamma} := \frac{\gamma_u+1}{4} \wedge M_n$. For any $\gamma < \frac{1}{C_4^2} \big(\frac{a_1}{a_2}\big)^{\frac{3}{2-\alpha}} \widetilde{\gamma}$, we have

$$\widehat{V}(\gamma) \geq \frac{1}{C_4} \frac{a_1}{a_2^2} \gamma^{\alpha-2} > C_4 \frac{a_2}{a_1^2} \widetilde{\gamma}^{\alpha-2} \geq \widehat{V}(\widetilde{\gamma}).$$

Since $\widehat{\gamma} = \arg\min_{\gamma \in \mathcal{N}_n, \gamma \leq \gamma_{\max}} \widehat{V}(\gamma)$ and since $\frac{1}{C_4^2} \big(\frac{a_1}{a_2}\big)^{\frac{3}{2-\alpha}} \leq 1$ so that there exists $\gamma \in \mathcal{N}_n$ such that $\gamma_{\max} \geq \gamma \geq \frac{1}{C_4^2} \big(\frac{a_1}{a_2}\big)^{\frac{3}{2-\alpha}} \widetilde{\gamma}$, it must be that

$$\widehat{\gamma} \geq \frac{1}{C_4^2} \big(\frac{a_1}{a_2}\big)^{\frac{3}{2-\alpha}} \widetilde{\gamma} \geq \frac{1}{C_4^2} \big(\frac{a_1}{a_2}\big)^{\frac{3}{2-\alpha}} \left(\frac{\gamma_u + 1}{2} \wedge M_n \right) \geq \widetilde{C}_{a_1,a_2,\alpha}^{-1} \left\{ \left(\frac{n}{\log n} \right)^{\frac{1}{\alpha}} \wedge M_n \right\},$$

where we define $\widetilde{C}_{a_1,a_2,\alpha}^{-1} := \frac{1}{4C_4^2} \big(\frac{a_1}{a_2}\big)^{\frac{3}{2-\alpha}} \big(\frac{1}{C_{1\alpha} \vee C_{2\alpha} \vee C_4} \frac{c_0^2 a_1^4}{a_2} \alpha^2 \big)^{\frac{1}{\alpha}}$.

Now define $\mathcal{E}_3$ as the event that $|\widehat{\theta}_{\text{mid}} - \theta_0| \leq 2^{2+\frac{2}{\alpha}} a_1^{-\frac{1}{\alpha}} \frac{1}{\alpha^{\frac{1}{\alpha}+1}} \frac{\log^{1+\frac{1}{\alpha}} n}{n^{\frac{1}{\alpha}}}$. We have by Corollary S4.12 that $\mathbb{P}(\mathcal{E}_3) \geq 1 - \frac{2}{n^{1/\alpha}}$. Therefore, on the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, we have by Lemma 1

that

$$\begin{aligned}|\widehat{\theta}_{\widehat{\gamma}} - \theta_0| &\le |\widehat{\theta}_{\widehat{\gamma}} - \widehat{\theta}_{\text{mid}}| + |\widehat{\theta}_{\text{mid}} - \theta_0| \\ &\le 4\frac{\log n}{\widehat{\gamma}} + 2^{2+\frac{2}{\alpha}} a_1^{-\frac{1}{\alpha}} \frac{1}{\alpha^{\frac{1}{\alpha}+1}} \frac{\log^{1+\frac{1}{\alpha}} n}{n^{\frac{1}{\alpha}}} \\ &\le \widetilde{C}_{a_1,a_2,\alpha}\left\{\frac{\log^{1+\frac{1}{\alpha}} n}{n^{\frac{1}{\alpha}}} \vee \frac{\log n}{M_n}\right\} + 2^{2+\frac{2}{\alpha}} a_1^{-\frac{1}{\alpha}} \frac{1}{\alpha^{\frac{1}{\alpha}+1}} \frac{\log^{1+\frac{1}{\alpha}} n}{n^{\frac{1}{\alpha}}} \\ &\le C_{a_1,a_2,\alpha}\left\{\frac{\log^{1+\frac{1}{\alpha}} n}{n^{\frac{1}{\alpha}}} \vee \frac{\log n}{M_n}\right\},\end{aligned}$$

where, in the final inequality, we define $C_{a_1,a_2,\alpha} := \widetilde{C}_{a_1,a_2,\alpha} + 2^{2+\frac{2}{\alpha}} a_1^{-\frac{1}{\alpha}} \frac{1}{\alpha^{\frac{1}{\alpha}+1}}$.

Since $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3) \ge 1 - \frac{4}{n^{1/\alpha}} - \exp\big(-\frac{1}{\alpha}(\tau \wedge \sqrt{\log n})\sqrt{\log n}\big)$, the desired conclusion (S4.3) follows. Hence, the Theorem follows as well.

■

Theorem S4.7 *Let $Z_1, \dots, Z_n$ be independent and identically distributed random variables on $\mathbb{R}$ with a distribution P symmetric around 0 and write $\nu_\gamma := \mathbb{E}|Z|^\gamma$. Suppose there exists $\alpha \in (0,2)$ and $a_1 \in (0,1]$ and $a_2 \ge 1$ such that $\frac{a_1}{\gamma^\alpha} \le \nu_\gamma \le \frac{a_2}{\gamma^\alpha}$ for all $\gamma \ge 1$.*

Let $C_1, C_2 > 0$ be universal constants and $C_{1\alpha} > 0$ be a constant depending only on α, as defined in Proposition S4.8. Let $c_0 \in (0, 2^{-8})$, let $\gamma_u^\alpha = \frac{1}{C_{1\alpha} \vee C_{2,\alpha}} \big(\frac{c_0^2 a_1^6}{a_2^3} \alpha^2\big) \frac{n}{\log n}$, and let $\tau' \ge \frac{C_1\sqrt{a_2}}{a_1}\sqrt{\log\log n}$.

Suppose n is large enough so that $\gamma_u \ge 2$. Then, we have that

$$\mathbb{P}\left\{\sup_{\gamma \in [2,\gamma_u]} \frac{4\gamma}{a_1 c_0} |\widehat{\theta}_\gamma - \theta_0| \ge 1\right\} \le n^{-\frac{1}{\alpha}}. \tag{S4.6}$$

Moreover, if n is large enough such that $\gamma_u \ge e^{C_1 \frac{a_2}{a_1}} \ge 2$ and that $\sqrt{\log n} \ge \frac{C_1\sqrt{a_2}}{a_1}\sqrt{\log\log n}$. Then, we also have

$$\mathbb{P}\left\{\sup_{\gamma \in [2,\gamma_u]} \frac{|\widehat{\theta}_\gamma - \theta_0|}{\tau' \sqrt{\frac{\gamma^{\alpha-2}}{n}}} \ge 1\right\} \le \exp\left\{-\frac{a_1^2}{C_2 a_2}\left(\tau' \wedge \sqrt{\log n}\right)\sqrt{\frac{n}{\gamma_u^\alpha}}\right\} \tag{S4.7}$$

Proof

Since $\widehat{\theta}_\gamma$ for any $\gamma \ge 2$ is location equivariant, we assume without loss of generality that $\theta_0 = 0$ so that $Y_i = Z_i$.

Define $\widetilde{\tau} = \tau' \wedge \sqrt{\log n}$ and note that $\frac{C_1\sqrt{a_2}}{a_1}\sqrt{\log\log n} \le \widetilde{\tau} \le \sqrt{\log n} \le \frac{1}{4}\sqrt{\frac{n}{\gamma_u^\alpha}}$ since $a_1 \le 1$, $a_2 \ge 1$, and $c_0 \le 2^{-8}$. We further note that with our definition of and assumptions, the conditions in Proposition S4.8 (i) and (ii) are all satisfied.

Let $\{\Delta_\gamma\}_{\gamma \ge 2}$ be a collection of positive numbers. For any $\gamma \ge 2$, we have by the second claim of Lemma S5.16 that, for $t \in \{-\Delta_\gamma, \Delta_\gamma\}$,

$$\big|\mathbb{E}\big[-\operatorname{sgn}(Z-t)|Z-t|^{\gamma-1}\big]\big| \ge \frac{a_1}{2}\Delta_\gamma \gamma^{1-\alpha}. \tag{S4.8}$$

To prove the first claim of the theorem, we let $\Delta_\gamma = \frac{a_1 c_0}{4}$. We use Proposition S4.8 (noting that the probability bound in (S4.9) is less than $\exp\{-\frac{1}{\alpha}\log n\}$ under our definition of γ_u) and (S4.8) to obtain that, with probability at least $1 - n^{-\frac{1}{\alpha}}$, the following holds simultaneously for all $\gamma \in [2, \gamma_u]$:

$$\frac{1}{n}\sum_{i=1}^{n}\{-\text{sgn}(Y_i - \Delta_\gamma)|Y_i - \Delta_\gamma|^\gamma\} \geq \frac{1}{2}\mathbb{E}\big[-\text{sgn}(Y - \Delta_\gamma)|Y - \Delta_\gamma|^{\gamma-1}\big] > 0,$$

where, in the last inequality, we use the fact that the function $\theta \mapsto \mathbb{E}|Y - \theta|^\gamma$ is strongly convex for all $\gamma > 1$ and minimized at $\theta = \theta_0 = 0$.

Likewise, we have that

$$\frac{1}{n}\sum_{i=1}^{n}\{-\text{sgn}(Y_i + \Delta_\gamma)|Y_i + \Delta_\gamma|^\gamma\} \geq \frac{1}{2}\mathbb{E}\big[-\text{sgn}(Y + \Delta_\gamma)|Y + \Delta_\gamma|^{\gamma-1}\big] < 0.$$

By the strong convexity of the function $\theta \mapsto \frac{1}{n}\sum_{i=1}^n |Y_i - \theta|^\gamma$ therefore, we have that $|\widehat{\theta}_\gamma - \theta_0| = |\widehat{\theta}_\gamma| \leq \Delta_\gamma$. The first claim thus follows as desired.

To prove the second claim, we let $\Delta_\gamma = \widetilde{\tau}\sqrt{\frac{\gamma^\alpha}{n}}$ and follow exactly the same argument. The only difference is that the probability bound of Proposition S4.8 in this case becomes, under our assumptions on $\widetilde{\tau}$,

$$\exp\left\{-\frac{a_1^2}{C_2 a_2}\left(\frac{\widetilde{\tau}^2}{\sqrt{\frac{\gamma_u^\alpha}{n}\log\log\gamma_u}} \wedge \frac{\widetilde{\tau}}{\sqrt{\frac{\gamma_u^\alpha}{n}}}\right\}\right. \leq \exp\left\{-\frac{a_1^2}{C_2 a_2}\widetilde{\tau}\sqrt{\frac{n}{\gamma_u^\alpha}}\right\}.$$

The entire theorem then follows. ■

Proposition S4.8 *Let $Z_1, \ldots, Z_n$ be independent and identically distributed random variables on $\mathbb{R}$ with a distribution P symmetric around 0 and write $\nu_\gamma := \mathbb{E}|Z|^\gamma$. Suppose there exists $\alpha \in (0, 2)$ and $a_1 \in (0, 1]$ and $a_2 \geq 1$ such that $\frac{a_1}{\gamma^\alpha} \leq \nu_\gamma \leq \frac{a_2}{\gamma^\alpha}$ for all $\gamma \geq 1$.*

For $\gamma \geq 1$ and $x \in \mathbb{R}$, define $\psi_\gamma(x) := -sgn(x)|x|^{\gamma-1}$. Let $\gamma_u > 2$ and let $\{\Delta_\gamma\}_{\gamma\in[2,\gamma_u]}$ be a collection of positive numbers; define the event

$$\mathcal{E} \equiv \mathcal{E}_{\gamma_u,\{\Delta_\gamma\},\alpha} := \left\{\sup_{\substack{\gamma\in[2,\gamma_u]\\ |t|=\Delta_\gamma}}\left|\frac{\frac{1}{n}\sum_{i=1}^n \psi_\gamma(Z_i - t) - \mathbb{E}\psi_\gamma(Z - t)}{\frac{a_1}{2}\Delta_\gamma\gamma^{1-\alpha}}\right| \geq \frac{1}{2}\right\}$$

Let $C_1, C_2 > 0$ be universal constants and $C_{1\alpha}$ be a constant depending only on α (the values of these are specified in the proof). Then, the following holds:

(i) Suppose $\Delta_\gamma\gamma = \frac{a_1 c_0}{4}$ for some $c_0 \in (0, 1)$ and suppose that $\gamma_u \geq 2$ and $\sqrt{\frac{\gamma_u^\alpha}{n}} \leq \frac{1}{4}\frac{c_0 a_1}{C_1\sqrt{a_2}}$, then, we have that

$$\mathbb{P}(\mathcal{E}^c) \leq \exp\left\{-\frac{c_0^2 a_1^4}{C_{1\alpha} a_2}\left(\frac{n}{\gamma_u^\alpha}\right)\right\}, \tag{S4.9}$$

(ii) Let $\widetilde{\tau} \geq \frac{C_1\sqrt{a_2}}{a_1}\sqrt{\log\log n}$ *and suppose* $\Delta_\gamma \gamma = \widetilde{\tau}\sqrt{\frac{\gamma^\alpha}{n}}$. *Suppose also* $\gamma_u \geq e^{C_1 \frac{a_2}{a_1}}$ *and* $\sqrt{\frac{\gamma_u^\alpha}{n}} \leq \frac{1}{4\widetilde{\tau}}$. *Then, we have that*

$$\mathbb{P}(\mathcal{E}^c) \leq \exp\left\{-\frac{a_1^2}{C_2 a_2}\left(\frac{\widetilde{\tau}^2}{\sqrt{\frac{\gamma_u^\alpha}{n}\log\log\gamma_u}} \wedge \frac{\widetilde{\tau}}{\sqrt{\frac{\gamma_u^\alpha}{n}}}\right\}\right.\right.$$

Proof

We define the function class

$$\mathcal{F} \equiv \mathcal{F}_{\gamma_u,\{\Delta_\gamma\},\alpha} := \left\{ \frac{\psi_\gamma(z-t)}{\frac{a_1}{2}\Delta_\gamma \gamma^{1-\alpha}} \,:\, t \in \{-\Delta_\gamma, \Delta_\gamma\}, \gamma \in [2,\gamma_u] \right\}. \tag{S4.10}$$

We now use Talagrand's inequality (Theorem S5.18) to prove the Proposition. To this end, we derive upper bounds on various quantities involved in Talagrand's inequality.

Step 1: bounding $\sup_{f\in\mathcal{F}} \|f(Z)\|_{\text{ess-sup}}$ and $\widetilde{\sigma}^2 := \sup_{f\in\mathcal{F}} \mathbb{E}f(Z)^2$.

Using the fact $\Delta_\gamma \leq \frac{1}{4\gamma}$ in both cases, we observe that for any $\gamma \geq 2$, if $|t| = \Delta_\gamma$ and $|z| \leq 1$, then $|\psi_\gamma(z-t)| \leq (1+\Delta_\gamma)^{\gamma-1} \leq e$. Therefore, we have that,

$$U := \sup_{\gamma\in[2,\gamma_u]} \left| \frac{\frac{1}{n}\sum_{i=1}^n \psi_\gamma(Z_i - t)}{\frac{a_1}{2}\Delta_\gamma\gamma^{1-\alpha}} \right| \leq \frac{2e}{a_1} \sup_{\gamma\in[2,\gamma_u]} \frac{\gamma^\alpha}{\Delta_\gamma\gamma}.$$

Thus, it follows that

$$U \leq \begin{cases} \frac{C}{c_0 a_1^2}\gamma_u^\alpha & \text{if } \Delta_\gamma\gamma = \frac{a_1 c_0}{4} \\ \frac{C}{a_1}\frac{\sqrt{\gamma_u^\alpha n}}{\widetilde{\tau}} & \text{if } \Delta_\gamma\gamma = \widetilde{\tau}\sqrt{\frac{\gamma^\alpha}{n}} \end{cases}. \tag{S4.11}$$

Next, we have that, writing $\widetilde{\sigma}^2 := \sup_{f\in\mathcal{F}} \mathbb{E}f(Z)^2$,

$$\begin{aligned} \widetilde{\sigma}^2 &\leq \frac{1}{4a_1^2} \sup_{\substack{\gamma\in[2,\gamma_u] \\ |t|=\Delta_\gamma}} \frac{\gamma^{2\alpha}\mathbb{E}|Z-t|^{2(\gamma-1)}}{\Delta_\gamma^2\gamma^2} \\ &= \frac{1}{4a_1^2} \sup_{\gamma\in[2,\gamma_u]} \frac{\gamma^{2\alpha}\mathbb{E}|Z-\Delta_\gamma|^{2(\gamma-1)}}{\Delta_\gamma^2\gamma^2} \\ &\leq \frac{Ca_2}{a_1^2} \sup_{\gamma\in[2,\gamma_u]} \frac{\gamma^\alpha}{\Delta_\gamma^2\gamma^2}, \end{aligned}$$

where the last inequality follows from Lemma S5.16.

Therefore, we have that

$$\widetilde{\sigma}^2 \leq \begin{cases} \frac{Ca_2}{c_0^2 a_1^4}\gamma_u^\alpha & \text{if } \Delta_\gamma\gamma = \frac{a_1 c_0}{4} \\ \frac{Ca_2}{a_1^2}\frac{n}{\widetilde{\tau}^2} & \text{if } \Delta_\gamma\gamma = \widetilde{\tau}\sqrt{\frac{\gamma^\alpha}{n}} \end{cases}. \tag{S4.12}$$

When $\Delta_\gamma\gamma = \widetilde{\tau}\sqrt{\frac{\gamma^\alpha}{n}}$, we also see that

$$\widetilde{\sigma}^2 \geq \frac{C}{a_1}\frac{1}{4\Delta_2^2} \geq \frac{C}{a_1}\frac{n}{\widetilde{\tau}^2}. \tag{S4.13}$$

Step 2: bounding the envelope function.

Define $F(z) := \sup_{f\in\mathcal{F}} |f(z)|$. Since, for any $z \in \mathbb{R}$,

$$\sup_{\gamma\in[2,\gamma_u],|t|=\Delta_\gamma} |\psi_\gamma(z-t)| = |(|z|+\Delta_\gamma)^{\gamma-1}|,$$

we have that

$$F(z) = \frac{4}{a_1} \sup_{\gamma\in[2,\gamma_u]} \frac{\gamma^\alpha |(|z|+\Delta_\gamma)^{\gamma-1}|}{\Delta_\gamma \gamma}. \tag{S4.14}$$

Using the fact that the distribution of Z is symmetric around 0, and defining $K := \lceil \log_2 \gamma_u \rceil$,

$$\begin{aligned}
\mathbb{E}F^2(Z) &= \frac{16}{a_1^2} \int_0^1 \sup_{\gamma\in[2,\gamma_u]} \frac{\gamma^{2\alpha}(z+\Delta_\gamma)^{2(\gamma-1)}}{\Delta_\gamma^2\gamma^2}\, dP(z) \\
&= \leq \frac{16}{a_1^2} \sum_{k=1}^K \int_0^1 \sup_{\gamma\in[2^k,2^{k+1}]} \frac{\gamma^{2\alpha}(z+\Delta_\gamma)^{2(\gamma-1)}}{\Delta_\gamma^2\gamma^2}\, dP(z)
\end{aligned} \tag{S4.15}$$

Case 1: suppose $\Delta_\gamma\gamma = \frac{a_1 c_0}{4}$. In this case, we have that

$$\begin{aligned}
\mathbb{E}F^2(Z) &= \frac{16}{c_0^2 a_1^4} \sum_{k=1}^K \int_0^1 \sup_{\gamma\in[2^k,2^{k+1}]} \gamma^{2\alpha}(z+\Delta_\gamma)^{2(\gamma-1)}\, dP(z) \\
&\leq \frac{C}{c_0^2 a_1^4} \sum_{k=1}^K 2^{2k\alpha} \int_0^1 \sup_{\gamma\in[2^k,2^{k+1}]} (z+\Delta_\gamma)^{2(\gamma-1)}\, dP(z) \\
&\leq \frac{C}{c_0^2 a_1^4} \sum_{k=1}^K 2^{2k\alpha} \cdot a_2 2^{-k\alpha} \\
&\leq \frac{C_\alpha a_2}{c_0^2 a_1^4} \gamma_u^\alpha,
\end{aligned}$$

where the second inequality follows from the third claim of Lemma S5.16.
Case 2: suppose $\Delta_\gamma\gamma = \widetilde{\tau}\sqrt{\frac{\gamma^\alpha}{n}}$. In this case,

$$\begin{aligned}
\mathbb{E}F^2(Z) &= \frac{C}{a_1^2}\frac{n}{\widetilde{\tau}^2} \sum_{k=1}^K \int_0^1 \sup_{\gamma\in[2^k,2^{k+1}]} \gamma^{\alpha}(z+\Delta_\gamma)^{2(\gamma-1)}\, dP(z) \\
&\leq \frac{C}{a_1^2}\frac{n}{\widetilde{\tau}^2} \sum_{k=1}^K 2^{k\alpha} \int_0^1 \sup_{\gamma\in[2^k,2^{k+1}]} (z+\Delta_\gamma)^{2(\gamma-1)}\, dP(z) \\
&\leq \frac{C}{a_1^2}\frac{n}{\widetilde{\tau}^2} \sum_{k=1}^K 2^{k\alpha} \cdot C a_2 2^{-k\alpha} \leq \frac{C a_2}{a_1^2}\frac{n}{\widetilde{\tau}^2} \log \gamma_u,
\end{aligned} \tag{S4.16}$$

where, in the second inequality, we use Lemma S5.16 again.

Step 3: bounding the VC-dimension of $\mathcal{F}$.

We first note that the class of univariate functions $\mathcal{G} := \left\{\frac{|\cdot|^{\gamma-1}}{\Delta_\gamma \gamma} : \gamma \geq 2\right\}$ has VC dimension at most 4. This holds because $\log \mathcal{G}$ consists of functions of the form

$$(\gamma - 1)\log|\cdot| + \log(\Delta_\gamma \gamma)$$

and thus lies in a subspace of dimension 2. It then follows from Lemma 2.6.15 and 2.6.18 (viii) of Van Der Vaart and Wellner (1996) that $\mathcal{G}$ has VC-dimension at most 4.

It then follows from Lemma 2.6.18 (vi) that $\mathcal{F}$ has VC-dimension at most 8.

Step 4: bounding the expected supremum.

Let us define

$$\widetilde{S}_n := \sup_{\substack{\gamma\in[2,\gamma_u]\\ |t|=\Delta_\gamma}} \left| \frac{1}{n}\sum_{i=1}^n \frac{\psi_\gamma(Z_i - t) - \mathbb{E}\psi_\gamma(Z-t)}{\frac{a_1}{2}\Delta_\gamma \gamma^{1-\alpha}} \right|. \tag{S4.17}$$

Case 1: suppose $\Delta_\gamma \gamma = \frac{a_1 c_0}{4}$. Then, using the second claim of Theorem S5.21, we have that

$$\mathbb{E}\widetilde{S}_n \leq \frac{C_\alpha \sqrt{a_2}}{c_0 a_1^2}\sqrt{\frac{\gamma_u^\alpha}{n}}. \tag{S4.18}$$

Case 2: suppose now that $\Delta_\gamma \gamma = \widetilde{\tau}\sqrt{\frac{\gamma^\alpha}{n}}$.

We first note that, by (S4.13) and (S4.16),

$$\frac{\widetilde{\sigma}}{\|F\|_{L_2(P)}} \geq \frac{C a_1}{a_2}\frac{1}{\sqrt{\log \gamma_u}}. \tag{S4.19}$$

Define the entropy integral $J(\delta)$ as (S5.29) and note that $\frac{1}{\delta}J(\delta)$ is decreasing for $\delta \in (0,1]$. By Corollary S5.20 and our bound on the VC-dimension of $\mathcal{F}$, we have that

$$\frac{\|F\|_{L_2(P)}}{\widetilde{\sigma}} J\left(\frac{\widetilde{\sigma}}{\|F\|_{L_2(P)}}\right) \leq \sqrt{1 \vee \log\left(\frac{a_2}{C a_1}\sqrt{\log \gamma_u}\right)} \leq \sqrt{\log\log \gamma_u + \log\left(\frac{a_2}{a_1}\right) + C}.$$

Therefore, using our upper and lower bounds on $\widetilde{\sigma}$, upper bound on U and upper bound on $\|F\|_{L_2(P)}$, we have, by the first claim of Theorem S5.21, that

$$\begin{aligned}
\mathbb{E}\widetilde{S}_n &\leq C\frac{\widetilde{\sigma}}{\sqrt{n}}\left(\sqrt{\log\log \gamma_u + \log\left(\frac{a_2}{a_1}\right) + C}\right)\left(1 + \frac{U}{\sqrt{n}\widetilde{\sigma}}\sqrt{\log\log \gamma_u + \log\left(\frac{a_2}{a_1}\right) + C}\right) \\
&\leq \frac{C\sqrt{a_2}}{a_1}\frac{1}{\widetilde{\tau}}\sqrt{\log\log \gamma_u + \log\left(\frac{a_2}{a_1}\right) + C} \leq \frac{C\sqrt{a_2}}{a_1}\frac{1}{\widetilde{\tau}}\sqrt{\log\log \gamma_u}.
\end{aligned}$$

where, in the second inequality, we used the fact that $\frac{U}{\sqrt{n}\widetilde{\sigma}} \leq C\sqrt{\frac{\gamma_u^\alpha}{n}} \leq C$, and in the last inequality, we used the hypothesis that $\gamma_u \geq e^{C_1 \frac{a_2}{a_1}}$ (with C_1 as a sufficiently large universal constant).

Step 5: bounding the tail probability.

Using our assumption that $\frac{C_1\sqrt{a_2}}{c_0 a_1^2}\sqrt{\frac{\gamma_u^\alpha}{n}} \le \frac{1}{4}$ and $\widetilde{\tau} \ge \frac{C_1\sqrt{a_2}}{a_1}\sqrt{\log\log n}$ (with C_1 as a sufficiently large universal constant), we have that $\mathbb{E}\widetilde{S}_n \le \frac{1}{4}$ in both the case where $\Delta_\gamma \gamma = \frac{a_1 c_0}{4}$ and the case where $\Delta_\gamma \gamma = \widetilde{\tau}\sqrt{\frac{\gamma^\alpha}{n}}$.
Case 1: when $\Delta_\gamma \gamma = \frac{a_1 c_0}{4}$, we have that, writing $t = 3/4$,

$$\begin{aligned}\mathbb{P}(\mathcal{E}^c) &\le \mathbb{P}(\widetilde{S}_n - \mathbb{E}\widetilde{S}_n \ge \frac{3}{4}) \\ &\le \exp\left\{-\frac{nt^2}{U\mathbb{E}\widetilde{S}_n + \widetilde{\sigma}^2} \wedge \frac{nt}{\frac{2}{3}U}\right\} \\ &\le \exp\left\{-\frac{c_0^2 a_1^4}{C_\alpha a_2}\left(\left(\frac{n}{\gamma_u^\alpha}\right)^{3/2} \wedge \frac{n}{\gamma_u^\alpha}\right)\right\}.\end{aligned}$$

Case 2: when $\Delta_\gamma \gamma = \widetilde{\tau}\sqrt{\frac{\gamma^\alpha}{n}}$, we have

$$\begin{aligned}\mathbb{P}(\mathcal{E}^c) &\le \mathbb{P}(\widetilde{S}_n - \mathbb{E}\widetilde{S}_n \ge \frac{3}{4}) \\ &\le \exp\left\{-\frac{nt^2}{U\mathbb{E}\widetilde{S}_n + \widetilde{\sigma}^2} \wedge \frac{nt}{\frac{2}{3}U}\right\} \\ &\le \exp\left\{-\frac{a_1^2}{C a_2}\left(\frac{\widetilde{\tau}^2}{\sqrt{\frac{\gamma_u^\alpha}{n}}\sqrt{\log\log\gamma_u}} \wedge \frac{\widetilde{\tau}}{\sqrt{\frac{\gamma_u^\alpha}{n}}}\right)\right\}.\end{aligned}$$

■

Theorem S4.9 *Let $Z_1, \dots, Z_n$ be independent and identically distributed random variables on $\mathbb{R}$ with a distribution P symmetric around 0 and write $\nu_\gamma := \mathbb{E}|Z|^\gamma$. Suppose there exists $\alpha \in (0,2)$ and $a_1 \in (0,1]$ and $a_2 \ge 1$ such that $\frac{a_1}{\gamma^\alpha} \le \nu_\gamma \le \frac{a_2}{\gamma^\alpha}$ for all $\gamma \ge 1$.*

Let $c_0 \in (0, 2^{-8}]$ and let $C_{1\alpha}, C_{2\alpha} > 0$ be constants depending only on α defined in Theorem S4.7 and Proposition S4.10. Define $\gamma_u^\alpha = \frac{1}{C_{1\alpha}\vee C_{2\alpha}}\left(\frac{c_0^2 a_1^6}{a_2^3}\alpha^2\right)\frac{n}{\log n}$ and suppose n is large enough so that $\gamma_u \ge 2$. Then, with probability at least $1 - 2n^{-\frac{1}{\alpha}}$, there exists a constant $C_3 \le 2$ such that

$$C_3 V(\gamma) \ge \widehat{V}(\gamma) \ge \frac{1}{C_3} V(\gamma), \quad \text{for all } \gamma \in [2, (\gamma_u + 1)/2].$$

Moreover, on the same event, there exists a universal constant $C_4 \ge 1$ such that

$$C_4 \frac{a_2}{a_1^2}\gamma^{\alpha-2} \ge \widehat{V}(\gamma) \ge \frac{1}{C_4}\frac{a_1}{a_2^2}\gamma^{\alpha-2}, \quad \text{for all } \gamma \in [2, (\gamma_u + 1)/2].$$

We note that, in Theorem S4.9, by choosing c_0 arbitrarily close to 0, we can have C_3 be arbitrarily close to 1.

Proof

By Theorem S4.7, with probability at least $1 - n^{-\frac{1}{\alpha}}$, we have that, simultaneously for all $\gamma \in [2, \gamma_u]$,

$$|\widehat{\theta}_\gamma - \theta_0| \leq \frac{a_1 c_0}{4\gamma}.$$

On this event, we have that

$$\widehat{\nu}_\gamma := \inf_{\theta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n |Y_i - \theta|^\gamma = \inf_{|\theta - \theta_0| \leq \frac{a_1 c_0}{4}} \frac{1}{n} \sum_{i=1}^n |Y_i - \theta|^\gamma.$$

Then, by Proposition S4.10, with probability at least $1 - n^{-\frac{1}{\alpha}}$, simultaneously for all $\gamma \in [2, \gamma_u]$,

$$1 - 3\sqrt{c_0} \leq \frac{\widehat{\nu}_\gamma}{\nu_\gamma} \leq 1 + 3\sqrt{c_0}.$$

Therefore,

$$\widehat{V}(\gamma) = \frac{\widehat{\nu}_{2(\gamma-1)}}{(\gamma-1)^2 \widehat{\nu}_{\gamma-2}^2} \geq \frac{1 - 3\sqrt{c_0}}{(1 + 3\sqrt{c_0})^2} \frac{\nu_{2(\gamma-1)}}{(\gamma-1)^2 \nu_{\gamma-2}^2} = \frac{1 - 3\sqrt{c_0}}{(1 + 3\sqrt{c_0})^2} V(\gamma).$$

Likewise, we have that $\widehat{V}(\gamma) \leq \frac{1+3\sqrt{c_0}}{(1-3\sqrt{c_0})^2} V(\gamma)$. Using our assumption that $c_0 \leq 2^{-8}$, the first claim of the theorem directly follows.

The second claim of the theorem follows then from Lemma S5.15.

■

Proposition S4.10 *Let $Z_1, \ldots, Z_n$ be independent and identically distributed random variables on $\mathbb{R}$ with a distribution P symmetric around 0 and write $\nu_\gamma := \mathbb{E}|Z|^\gamma$. Suppose there exists $\alpha \in (0, 2)$ and $a_1 \in (0, 1]$ and $a_2 \geq 1$ such that $\frac{a_1}{\gamma^\alpha} \leq \nu_\gamma \leq \frac{a_2}{\gamma^\alpha}$ for all $\gamma \geq 1$.*

Let $\gamma_u \geq 2$ and $c_0 \in (0, 1)$. Define the event

$$\mathcal{A}_{\gamma_u, c_0} := \left\{ \sup_{\gamma \in [2, \gamma_u]} \left| \frac{\inf_{|\theta - \theta_0| \leq \frac{a_1 c_0}{4\gamma}} \frac{1}{n} \sum_{i=1}^n |Y_i - \theta|^\gamma - \nu_\gamma}{\nu_\gamma} \right| \leq 3\sqrt{c_0} \right\}. \tag{S4.20}$$

Let $C_{2\alpha} > 0$ be a constant depending only on α (its value is specified in the proof). Suppose $\frac{\gamma_u^\alpha}{n} \leq c_0^2 \frac{a_1^2}{C_{2\alpha} a_2}$. Then, we have that

$$\mathbb{P}(\mathcal{A}_{\gamma_u, c_0}^c) \leq \exp\left\{ -\frac{a_1^2}{C_{2\alpha} a_2} c_0^2 \left(\frac{n}{\gamma_u^\alpha} \right) \right\}.$$

Proof

First, we claim that, for all $\gamma \geq 1$, $z \in \mathbb{R}$, $t \geq 0$, and $\kappa > 0$, it holds that

$$\begin{aligned} |z - t|^\gamma \geq (|z| - t)_+^\gamma &= |z|^\gamma \left(1 - \frac{t}{|z|} \right)_+^\gamma \\ &= |z|^\gamma \left(1 - \frac{t}{\kappa} \frac{\kappa}{|z|} \right)_+^\gamma \geq |z|^\gamma \left(1 - \frac{t}{\kappa} \right)_+^\gamma - \kappa^\gamma \\ &\geq |z|^\gamma \left(1 - \gamma \frac{t}{\kappa} \right) - \kappa^\gamma. \end{aligned} \tag{S4.21}$$

Now define $\Delta_\gamma = \frac{a_1 c_0}{4\gamma}$ and $\widetilde{\theta} := \arg\min_{|\theta-\theta_0|\le\Delta_\gamma} \frac{1}{n}\sum_{i=1}^n |Y_i-\theta|^\gamma$ and observe that $\widetilde{t} := \widetilde{\theta}-\theta_0 = \arg\min_{|t|\le\Delta_\gamma} \frac{1}{n}\sum_{i=1}^n |Z_i - t|^\gamma$. Suppose without loss of generality that $\widetilde{t}\ge 0$. Then, using (S4.21), we have that, for any $\kappa>0$,

$$\begin{aligned}\inf_{|\theta-\theta_0|\le\Delta_\gamma} \frac{1}{n}\sum_{i=1}^n |Y_i-\theta|^\gamma &= \frac{1}{n}\sum_{i=1}^n |Z_i-\widetilde{t}|^\gamma \\ &\ge \left(1-\frac{\gamma\Delta_\gamma}{\kappa}\right)\left(\frac{1}{n}\sum_{i=1}^n |Z_i|^\gamma\right) - \kappa^\gamma.\end{aligned}$$

We also trivially have that $\inf_{|\theta-\theta_0|\le\Delta_\gamma} \frac{1}{n}\sum_{i=1}^n |Y_i-\theta|^\gamma \le \frac{1}{n}\sum_{i=1}^n |Z_i|^\gamma$. Therefore, writing $\mathbb{E}_n|z|^\gamma := \frac{1}{n}\sum_{i=1}^n |Z_i|^\gamma$ and $\mathbb{E}_n|y-\theta|^\gamma := \frac{1}{n}\sum_{i=1}^n |Y_i-\theta|^\gamma$, we have that, for any $\kappa>0$,

$$\frac{\mathbb{E}_n|z|^\gamma-\nu_\gamma}{\nu_\gamma} \ge \frac{\min_{|\theta-\theta_0|\le\Delta_\gamma}\mathbb{E}_n|y-\theta|^\gamma-\nu_\gamma}{\nu_\gamma} \ge \frac{\left(1-\frac{\gamma\Delta_\gamma}{\kappa}\right)\mathbb{E}_n|z|^\gamma-\nu_\gamma-\kappa^\gamma}{\nu_\gamma}. \tag{S4.22}$$

Therefore, we have that

$$\left|\frac{\min_{|\theta-\theta_0|\le\Delta_\gamma}\mathbb{E}_n|y-\theta|^\gamma-\nu_\gamma}{\nu_\gamma}\right| \le \underbrace{\left|\frac{\mathbb{E}_n|z|^\gamma-\nu_\gamma}{\nu_\gamma}\right|}_{\text{Term 1}} + \underbrace{\inf_{\kappa>0}\left(\frac{\gamma\Delta_\gamma}{\kappa}+\frac{\kappa^\gamma}{\nu_\gamma}\right)}_{\text{Term 2}}. \tag{S4.23}$$

Bounding Term 2:

Since $\Delta_\gamma\gamma = \frac{a_1c_0}{4}$, by setting $\kappa = \left(\frac{a_1^2c_0}{4\gamma^\alpha}\right)^{\frac{1}{\gamma+1}}$, we have

$$\begin{aligned}\text{Term 2} &\le \inf_{\kappa>0}\left(\frac{a_1c_0}{4\kappa}+\frac{\kappa^\gamma}{a_1\gamma^\alpha}\right) \\ &\le \frac{a_1c_0}{4}\left(\frac{4}{a_1^2c_0}\right)^{\frac{1}{\gamma+1}}\gamma^{\frac{\alpha}{\gamma+1}} \le 2\sqrt{c_0}.\end{aligned}$$

Bounding Term 1:

To bound Term 1, we define the function class

$$\mathcal{F}_{\gamma_u} := \left\{z\mapsto\frac{|z|^\gamma}{\nu_\gamma} \,:\, \gamma\in[2,\gamma_u]\right\},$$

so that we have $\sup_{\gamma\in[2,\gamma_u]}\left|\frac{\mathbb{E}_n|z|^\gamma-\nu_\gamma}{\nu_\gamma}\right| = \sup_{f\in\mathcal{F}_{\gamma_u}}\left|\mathbb{E}_nf(z)-\mathbb{E}f(Z)\right|$.

We observe that

$$\begin{aligned}\widetilde{\sigma}^2 &:= \sup_{\gamma\in[2,\gamma_u]}\mathbb{E}\frac{|Z|^\gamma}{\nu_\gamma} \le \frac{a_2}{a_1} \\ U &:= \sup_{\gamma\in[2,\gamma_u]}\frac{\|Z\|^\gamma_{\text{ess-inf}}}{\nu_\gamma} \le \frac{1}{a_1}\gamma_u^\alpha.\end{aligned}$$

Moreover, defining $F(z) := \sup_{\gamma\in[2,\gamma_u]} \frac{|z|^\gamma}{\nu_\gamma}$ and $K = \lceil \log \gamma_u \rceil$, we have that

$$\mathbb{E}F^2(Z) \le \mathbb{E}\left(\sup_{\gamma\in[2,\gamma_u]} \frac{|Z|^{2\gamma}}{\nu_\gamma^2} \right) \tag{S4.24}$$

$$\le \frac{1}{a_1^2} \sum_{k=1}^{K} \int_0^1 \sup_{\gamma\in[2^k,2^{k+1}]} \gamma^{2\alpha} |z|^\gamma \, dP(z) \tag{S4.25}$$

$$\le \frac{1}{a_1^2} \sum_{k=1}^{K} 2^{2k\alpha+1} \nu_{2^k} \le \frac{a_2}{a_1^2} \sum_{k=1}^{K} 2^{k\alpha+1} \le C_\alpha \frac{a_2}{a_1^2} \gamma_u^\alpha. \tag{S4.26}$$

We note that $\log \mathcal{F}_{\gamma_u}$ is a subset of a linear subspace of dimension 2 (see Step 3 in the proof of Proposition S4.8). By Lemma 2.6.15 and 2.6.18 (viii) of Van Der Vaart and Wellner (1996), we know that the VC dimension of $\mathcal{F}_{\gamma_u}$ is at most 4.

Write $\widetilde{S}_n = \sup_{\gamma\in[2,\gamma_u]} \left| \frac{1}{n} \sum_{i=1}^n |Z_i|^\gamma - \nu_\gamma \right\}$. Then, by Corollary S5.20 and the second claim of Theorem S5.21, we have that

$$\mathbb{E}\widetilde{S}_n \le \frac{C_\alpha \sqrt{a_2}}{a_1} \sqrt{\frac{\gamma_u^\alpha}{n}}.$$

Therefore, using our hypothesis that $\frac{\gamma_u^\alpha}{n} \le c_0^2 \frac{a_1^2}{C_{2\alpha} a_2}$ where $C_{2\alpha}$ is chosen to be sufficiently large, then $\mathbb{E}\widetilde{S}_n \le \frac{1}{2} a_0$. Then,

$$\mathbb{P}\left\{ \sup_{\gamma\in[2,\gamma_u]} \left| \frac{\frac{1}{n}\sum_{i=1}^n |Z_i|^\gamma - \nu_\gamma}{\nu_\gamma} \right| \ge c_0 \right\} \le \mathbb{P}\left(\widetilde{S}_n - \mathbb{E}\widetilde{S}_n \ge \frac{c_0}{2} \right)$$

$$\le \exp\left\{ -\frac{a_1^2}{C_\alpha a_2} \left(c_0^2 \left(\frac{n}{\gamma_u^\alpha} \right)^{3/2} \wedge \frac{c_0 n}{\gamma_u^\alpha} \right) \right\}.$$

Therefore, by (S4.23), it holds that

$$\mathbb{P}\left\{ \sup_{\gamma\in[2,\gamma_u]} \left| \frac{\inf_{|\theta-\theta_0|\le\Delta_\gamma} \frac{1}{n}\sum_{i=1}^n |Y_i-\theta|^\gamma - \nu_\gamma}{\nu_\gamma} \right| \ge 3\sqrt{c_0} \right\}$$

$$\le \exp\left\{ -\frac{a_1^2}{C_\alpha a_2} \left(c_0^2 \left(\frac{n}{\gamma_u^\alpha} \right)^{3/2} \wedge c_0 \frac{n}{\gamma_u^\alpha} \right) \right\}.$$

By inflating the value of $C_{2\alpha}$ if necessary, the Proposition follows as desired. ■

Lemma S4.11 *Let X be a random variable on $[-1,1]$ with a distribution P symmetric around 0. If there exists $a_1 > 0$ and $\alpha \ge 0$ such that $\mathbb{E}|X|^\gamma \ge \frac{a_1}{\gamma^\alpha}$ for all $\gamma \ge 1$, then we have that*

$$\mathbb{P}\left(X \ge 1 - 2^{1+\frac{2}{\alpha}} a_1^{-\frac{1}{\alpha}} \frac{1}{\alpha^{\frac{1}{\alpha}+1}} \frac{\log^{1+\frac{1}{\alpha}} n}{n^{\frac{1}{\alpha}}} \right) \ge \alpha^{-1} \frac{\log n}{n}.$$

Proof

As a short-hand, write $\delta := 2^{1+2/\alpha} a_1^{-1/\alpha} \frac{1}{\alpha^{\frac{1}{\alpha}+1}} \frac{\log^{1+1/\alpha} n}{n^{1/\alpha}}$ and $p = \left(\alpha \frac{a_1 n}{4 \log n}\right)^{1/\alpha}$; note that $\delta p = 2\alpha^{-1} \log n$. Then,

$$\begin{aligned}
\mathbb{P}(X \geq 1-\delta) &= \int_{1-\delta}^{1} dP(x) \geq \int_{1-\delta}^{1} x^p dP(x) \\
&= \frac{1}{2}\mathbb{E}|X|^p - \int_0^{1-\delta} x^p dP(x) \\
&\geq \frac{a_1}{2p^\alpha} - (1-\delta)^p \geq \frac{a_1}{2p^\alpha} - e^{-\delta p} \\
&= 2\frac{1}{\alpha}\frac{\log n}{n} - \frac{1}{\alpha}\frac{1}{n^2} \geq \frac{1}{\alpha}\frac{\log n}{n}.
\end{aligned}$$

■

Corollary S4.12 *Let $X_1, \ldots, X_n$ be independent and identically distributed random variables on $[-1, 1]$ with a distribution P symmetric around 0 and let $X_{mid} = \frac{X_{(n)}+X_{(1)}}{2}$. If there exists $a_1 > 0$ and $\alpha \geq 0$ such that $\mathbb{E}|X|^\gamma \geq \frac{a_1}{\gamma^\alpha}$ for all $\gamma \in \mathbb{N}$, then we have that*

$$\mathbb{P}\left(|X_{mid}| \leq 2^{2+\frac{2}{\alpha}} a_1^{-\frac{1}{\alpha}} \frac{1}{\alpha^{\frac{1}{\alpha}+1}} \frac{\log^{1+\frac{1}{\alpha}} n}{n^{\frac{1}{\alpha}}}\right) \geq 1 - \frac{2}{n^{1/\alpha}}.$$

Proof

As a short-hand, write $\delta = 2^{1+\frac{2}{\alpha}} a_1^{-\frac{1}{\alpha}} \frac{1}{\alpha^{\frac{1}{\alpha}+1}} \frac{\log^{1+\frac{1}{\alpha}} n}{n^{\frac{1}{\alpha}}}$. By the fact that P is symmetric around 0 and Lemma S4.11, we have

$$\begin{aligned}
\mathbb{P}(|X_{\text{mid}}| \geq 2\delta) &\leq \mathbb{P}(X_{(n)} \leq 1-\delta \text{ or } X_{(1)} \geq -1+\delta) \\
&\leq 2\mathbb{P}(X_{(n)} \leq 1-\delta) \\
&\leq 2\left\{\mathbb{P}(X_1 \leq 1-\delta)\right\}^n \\
&\leq 2\left(1 - \frac{1}{\alpha}\frac{\log n}{n}\right)^n \leq 2e^{-\frac{1}{\alpha}\log n} \leq \frac{2}{n^{1/\alpha}}.
\end{aligned}$$

The desired conclusion thus follows. ■

S4.2. Proof of Examples

Proof (of Proposition 15)

It suffices to show that there exists constants $C''_{\alpha,1}, C''_{\alpha,2} > 0$ such that

$$\frac{C''_{\alpha,1}}{\gamma^\alpha} \leq \int_{-1}^{1} |x|^\gamma (1-|x|)^{\alpha-1} dx \leq \frac{C''_{\alpha,2}}{\gamma^\alpha}.$$

Indeed, we have by Stirling's approximation that

$$\begin{aligned}\int_{-1}^{1}|x|^{\gamma}(1-|x|)^{\alpha-1}dx &= 2\int_{0}^{1}x^{\gamma}(1-x)^{\alpha-1}dx\\ &= 2\frac{\Gamma(\alpha)\Gamma(\gamma+1)}{\Gamma(\gamma+\alpha+1)}\\ &\asymp 2\Gamma(\alpha)\left(\frac{\gamma}{e}\right)^{\gamma}\left(\frac{\gamma+\alpha}{e}\right)^{-(\gamma+\alpha)}\left(\frac{\gamma}{\gamma+\alpha}\right)^{1/2}\\ &= 2\Gamma(\alpha)\left(\frac{\gamma}{\gamma+\alpha}\right)^{\gamma+1/2}e^{\alpha}\left(\frac{\gamma}{\gamma+\alpha}\right)^{\alpha}\gamma^{-\alpha}.\end{aligned}$$

The conclusion of the Proposition then directly follows from the fact that $1 \geq (\frac{\gamma}{\gamma+\alpha})^{\gamma+1/2} \geq e^{-3}$ for all $\gamma \geq 2$.

■

For a given density $p(\cdot)$, we define

$$H^2(\theta_1,\theta_2) := \frac{1}{2}\int_{\mathbb{R}}\left(p(x-\theta_1)^{1/2}-p(x-\theta_2)^{1/2}\right)^2 dx$$

for any $\theta_1, \theta_2 \in \mathbb{R}$.

Proposition S4.13 *Let $\alpha \in (0,2)$ and suppose X is a random variable with density $p(\cdot)$ satisfying*

$$C_{\alpha,1}(1-|x|)_+^{\alpha-1} \leq p(x) \leq C_{\alpha,2}(1-|x|)_+^{\alpha-1}$$

for $C_{\alpha,1}, C_{\alpha,2} > 0$ dependent only on α. Suppose also that $\left|\frac{p'(x)}{p(x)}\right| \leq \frac{C}{1-|x|}$ for some $C > 0$.

Suppose $p(\cdot)$ is symmetric around 0. Then, there exist $C'_{\alpha,1}, C'_{\alpha,2}$ dependent only on α and C such that

$$C'_{\alpha,1}|\theta_1-\theta_2|^{\alpha} \leq H^2(\theta_1,\theta_2) \leq C'_{\alpha,2}|\theta_1-\theta_2|^{\alpha}$$

for all $\theta_1, \theta_2 \in \mathbb{R}$.

Proof Since $H^2(\theta_1,\theta_2) = H^2(0,\theta_1-\theta_2)$, it suffices to bound $H^2(0,\theta)$ for $\theta \geq 0$.

For the lower bound, we observe that

$$\begin{aligned}H^2(0,\theta) &= \int_{-1}^{1+\theta}\left\{p(x)^{1/2}-p(x-\theta)^{1/2}\right\}^2 dx\\ &\geq \int_{1}^{1+\theta}p(x-\theta)dx = \int_{1-\theta}^{1}f(t)dt\\ &\geq C_{\alpha,1}\int_{1-\theta}^{1}(1-t)^{\alpha-1}dt\\ &= C_{\alpha,1}\left[-\frac{(1-t)^{\alpha}}{\alpha}\right]_{1-\theta}^{1} = \frac{C_{\alpha,1}}{\alpha}\theta^{\alpha}.\end{aligned}$$

To establish the upper bound, observe that, by symmetry of $p(\cdot)$,

$$\begin{aligned} H^2(0,\theta) &= 2\int_{\theta/2}^{1+\theta} \{p(x)^{1/2} - p(x-\theta)^{1/2}\}^2 dx \\ &= 2\int_{1-\theta}^{1+\theta} \{p(x)^{1/2} - p(x-\theta)^{1/2}\}^2 dx + 2\int_{\theta/2}^{1-\theta} \{p(x)^{1/2} - p(x-\theta)^{1/2}\}^2 dx. \end{aligned} \tag{S4.27}$$

We upper bound the two terms of (S4.27) separately. To bound the first term,

$$\begin{aligned} &\int_{1-\theta}^{1+\theta} \{p(x)^{1/2} - p(x-\theta)^{1/2}\}^2 dx \\ &\le \int_{1}^{1+\theta} p(x-\theta)dx + \int_{1-\theta}^{1} p(x) \vee p(x-\theta)dx \\ &\le \frac{C_{\alpha,2}}{\alpha}\theta^\alpha + C_{\alpha,2}\int_{1-\theta}^{1} \{(1-x)^{\alpha-1} \vee (1-(x-\theta))^{\alpha-1}\}dx \end{aligned}$$

If $\alpha \ge 1$, then $(1-x)^{\alpha-1} \vee (1-(x-\theta))^{\alpha-1} = (1-(x-\theta))^{\alpha-1}$ and

$$\int_{1-\theta}^{1} (1-(x-\theta))^{\alpha-1}dx = \int_{1-2\theta}^{1-\theta} (1-x)^{\alpha-1}dx = (2^\alpha - 1)\frac{\theta^\alpha}{\alpha} \ge \frac{\theta^\alpha}{\alpha}.$$

On the other hand, if $\alpha < 1$, then $(1-x)^{\alpha-1} \vee (1-(x-\theta))^{\alpha-1} = (1-x)^{\alpha-1}$ and $\int_{1-\theta}^{1}(1-x)^{\alpha-1}dx = \frac{\theta^\alpha}{\alpha}$. Hence, we have that

$$\int_{1-\theta}^{1+\theta} \{p(x)^{1/2} - p(x-\theta)^{1/2}\}^2 dx \le \frac{2C_{\alpha,2}}{\alpha}\theta^\alpha.$$

We now turn to the second term of (S4.27). Write $\phi(x) = \log p(x)$ and note that $\phi'(x) = \frac{p'(x)}{p(x)}$. Then, by mean value theorem, there exists $\theta_x \in (0,\theta)$ depending on x such that

$$\begin{aligned} 2\int_{\theta/2}^{1-\theta} \{p(x)^{1/2} - p(x-\theta)^{1/2}\}^2 dx &= 2\int_{\theta/2}^{1-\theta} \frac{\theta^2}{4}\phi'(x-\theta_x)^2 e^{\phi(x-\theta_x)}dx \\ &\le \frac{CC_{\alpha,2}}{2}\theta^2 \int_{\theta/2}^{1-\theta} \left(\frac{1}{1-|x-\theta_x|}\right)^2 (1-|x-\theta_x|)^{\alpha-1}dx \\ &= \frac{CC_{\alpha,2}}{2}\theta^2 \int_{\theta/2}^{1-\theta} (1-|x-\theta_x|)^{\alpha-3}dx \\ &\le \frac{CC_{\alpha,2}}{2}\theta^2 \int_{0}^{1-\theta} (1-x)^{\alpha-3}dx \\ &= \frac{CC_{\alpha,2}}{2}\theta^2 \frac{\theta^{\alpha-2}}{2-\alpha} = \frac{CC_{\alpha,2}}{2(2-\alpha)}\theta^\alpha, \end{aligned}$$

where the second inequality follows because $\alpha - 3 < 0$. The desired conclusion immediately follows. ∎

Remark S4.14 *We observe that if a density p is of the form*

$$p(x) = C_\alpha(1-|x|)^{\alpha-1}\mathbb{1}\{|x|\le 1\},$$

for a normalization constant $C_\alpha > 0$, then $\frac{p'(x)}{p(x)} \lesssim \frac{1}{1+|x|}$ as required in Proposition S4.13. Therefore, we immediately see that for such a density, it holds that $H^2(\theta_1,\theta_2) \propto_\alpha |\theta_1-\theta_2|^\alpha$.

S4.3. Proof of Proposition S1.1

Proof

We first note that if $Y = \theta_0 + Z$ where Z has a density $p(\cdot)$ symmetric around 0, then, for $\gamma > 0$,

$$L(\gamma) = \frac{1}{\gamma}\log\big(\mathbb{E}|Z|^\gamma\big) + \frac{1+\log\gamma}{\gamma} + \log\Gamma\big(1+\frac{1}{\gamma}\big).$$

To prove the first claim, suppose that Z is supported on all of $\mathbb{R}$. We observe that

$$\lim_{\gamma\to\infty} L(\gamma) = \log\bigg(\lim_{\gamma\to\infty}\big\{\mathbb{E}|Z|^\gamma\big\}^{\frac{1}{\gamma}}\bigg).$$

We thus need only show that $\lim_{\gamma\to\infty}\big\{\mathbb{E}|Z|^\gamma\big\}^{\frac{1}{\gamma}} = \infty$. Let $M > 0$ be arbitrary, then, for any $\gamma > 0$,

$$\begin{aligned}\{\mathbb{E}|Z|^\gamma\}^{\frac{1}{\gamma}} &\ge \{\mathbb{E}\big[|Z|^\gamma\mathbb{1}\{|Z|\ge M\}\big]\}^{\frac{1}{\gamma}}\\ &\ge M\cdot\mathbb{P}(|Z|\ge M)^{\frac{1}{\gamma}}.\end{aligned}$$

Since $\mathbb{P}(|Z|\ge M) > 0$ for all $M > 0$ by assumption, we see that $\lim_{\gamma\to\infty}\{\mathbb{E}|Z|^\gamma\}^{\frac{1}{\gamma}} \ge M$. Since M is arbitrary, the claim follows.

Now consider the second claim of the Proposition and assume that $\|Z\|_\infty = 1$; write $g(\cdot)$ as the density of $|Z|$. Writing $\eta = \frac{1}{\gamma}$, we have that

$$L(1/\eta) = \eta\log\big(\mathbb{E}|Z|^{\frac{1}{\eta}}\big) + \eta(1-\log\eta) + \log\Gamma(1+\eta).$$

Differentiating with respect to η, we have

$$\begin{aligned}\frac{dL(1/\eta)}{d\eta} &= \log\frac{\mathbb{E}|Z|^{\frac{1}{\eta}}}{\eta} - \frac{\mathbb{E}\{|Z|^{\frac{1}{\eta}}\log|Z|\}}{\eta\mathbb{E}|Z|^{\frac{1}{\eta}}} + \frac{\Gamma'(1+\eta)}{\Gamma(1+\eta)}\\ &= \log\frac{\int_0^1 u^{\frac{1}{\eta}}g(u)du}{\eta} - \frac{\int_0^1 u^{\frac{1}{\eta}}\log(u)g(u)du}{\eta\int_0^1 u^{\frac{1}{\eta}}g(u)du} + \frac{\Gamma'(1+\eta)}{\Gamma(1+\eta)}.\end{aligned}$$

We make a change of variable by letting $t = -\frac{1}{\eta}\log u$ to obtain

$$\begin{aligned}\frac{dL(1/\eta)}{d\eta} &= \log\frac{\int_0^\infty e^{-t}g(e^{-\eta t})e^{-\eta t}\eta dt}{\eta} - \frac{\int_0^\infty e^{-t}(-\eta t)g(e^{-\eta t})e^{-\eta t}\eta dt}{\eta\int_0^\infty e^{-t}g(e^{-\eta t})e^{-\eta t}\eta dt} + \frac{\Gamma'(1+\eta)}{\Gamma(1+\eta)}\\ &= \log\bigg\{\int_0^\infty e^{-t}g(e^{-\eta t})e^{-\eta t}dt\bigg\} + \frac{\int_0^\infty te^{-t}g(e^{-\eta t})e^{-\eta t}dt}{\int_0^\infty e^{-t}g(e^{-\eta t})e^{-\eta t}dt} + \frac{\Gamma'(1+\eta)}{\Gamma(1+\eta)}.\end{aligned}$$

Therefore, using the fact that $\lim_{\eta\to 0}\frac{\Gamma'(1+\eta)}{\Gamma(1+\eta)} = -\gamma_{\mathrm{E}}$, we have that

$$\begin{aligned}\lim_{\eta\to 0}\frac{dL(1/\eta)}{d\eta} &= \log\left\{g(1)\int_0^\infty e^{-t}dt\right\} + \frac{\int_0^\infty te^{-t}dt}{\int_0^\infty e^{-t}dt} - \gamma_{\mathrm{E}} \\ &= \log g(1) + 1 - \gamma_{\mathrm{E}}.\end{aligned}$$

Therefore, if $g(1) > e^{\gamma_{\mathrm{E}}-1}$, then $\lim_{\eta\to 0}\frac{dL(1/\eta)}{d\eta} > 0$ and hence, $\eta = 0$ is a local minimum of $L(1/\eta)$. On the other hand, if $g(1) < e^{\gamma_{\mathrm{E}}-1}$, then $\lim_{\eta\to 0}\frac{dL(1/\eta)}{d\eta} < 0$ and $\eta = 0$ is not a local minimum. The Proposition follows as desired.

■

Appendix S5. Other material

S5.1. Technical Lemmas

Lemma S5.15 *Let Z be a random variable supported on $[-1,1]$. For $\gamma \geq 1$, define $\nu_\gamma := \mathbb{E}|Z|^\gamma$ and suppose there exists $\alpha \in (0,2]$, $a_1 \in (0,1]$, and $a_2 \in [1,\infty)$ such that $a_1\gamma^{-\alpha} \leq \nu_\gamma \leq a_2\gamma^{-\alpha}$ for all $\gamma \geq 1$.*

Define $V(\gamma) := \frac{\mathbb{E}|Z|^{2(\gamma-1)}}{(\gamma-1)^2\{\mathbb{E}|Z|^{\gamma-2}\}^2}$. Then, for some universal constant $C \geq 1$, for all $\gamma \geq 2$,

$$C\frac{a_2}{a_1^2}\gamma^{\alpha-2} \geq V(\gamma) \geq \frac{1}{C}\frac{a_1}{a_2^2}\gamma^{\alpha-2}.$$

Proof First suppose $\gamma \in [2,3]$. Then we have that

$$a_1 \leq \mathbb{E}|Z| \leq \mathbb{E}|Z|^{\gamma-2} \leq \{\mathbb{E}|Z|\}^{\gamma-2} \leq a_2,$$

where the second inequality follows because $|Z| \leq 1$, the third inequality follows from Jensen's inequality. Therefore, we have that

$$V(\gamma) \geq \frac{a_1\{2(\gamma-1)\}^{-\alpha}}{(\gamma-1)^2a_2^2} \geq \frac{1}{4^{\alpha+1}3^{\alpha-2}}\frac{a_1}{a_2^2}\gamma^{\alpha-2}.$$

The upper bound on $V(\gamma)$ follows similarly.

Now suppose $\gamma \geq 3$, then,

$$\begin{aligned}V(\gamma) &= \frac{\nu_{2(\gamma-1)}}{(\gamma-1)^2\nu_{\gamma-2}^2} \geq \frac{a_1\{2(\gamma-1)\}^{-\alpha}}{(\gamma-1)^2a_2^2(\gamma-2)^{-2\alpha}} \\ &= \frac{a_1}{a_2^2}2^{-\alpha}\left(\frac{\gamma-2}{\gamma-1}\right)^\alpha\left(\frac{\gamma-2}{\gamma}\right)^\alpha\left(\frac{\gamma}{\gamma-1}\right)^2\gamma^{\alpha-2} \geq \frac{1}{C}\frac{a_1}{a_2^2}\gamma^{\alpha-2}.\end{aligned}$$

The upper bound on $V(\gamma)$ follows in an identical manner. The conclusion of the Lemma then follows as desired.

■

Lemma S5.16 *Let Z be a random variable on $[-1,1]$ with a distribution symmetric around 0 and write $\nu_\gamma := \mathbb{E}|Z|^\gamma$ for $\gamma \geq 1$. Suppose $a_1\gamma^{-\alpha} \leq \nu_\gamma \leq a_2\gamma^{-\alpha}$ for all $\gamma \geq 1$ and for some $\alpha \in (0,2]$, $a_1 \in [0,1]$ and $a_2 \in [1,\infty)$. Then, for any $\gamma \geq 1$ and any $0 \leq \Delta \leq \frac{1}{4\gamma}$, we have*

$$\mathbb{E}|Z-\Delta|^\gamma \leq Ca_2\gamma^{-\alpha}.$$

Moreover, we have that for any $\gamma \geq 2$ and any $\Delta \in \mathbb{R}$,

$$\mathbb{E}\big[-|Z-\Delta|^{\gamma-1}sgn(Z-\Delta)\big] \geq \frac{a_1}{2}|\Delta|\gamma^{1-\alpha}.$$

Lastly, for any $k \in \mathbb{N}$ and any Δ_γ (allowed to depend on γ) such that $0 \leq \Delta_\gamma \leq \frac{1}{4\gamma}$, we have

$$\mathbb{E}\left[\sup_{\gamma\in[2^k,2^{k+1}]}(|Z|+\Delta_\gamma)^{2(\gamma-1)}\right] \leq Ca_2 2^{-k\alpha}.$$

Proof Consider the first claim. Observe that

$$\mathbb{E}|Z-\Delta|^\gamma = \underbrace{\mathbb{E}\Big[|Z-\Delta|^\gamma \mathbb{1}\{|Z|\leq 1/4\}\Big]}_{\text{Term 1}} + \underbrace{\mathbb{E}\Big[|Z-\Delta|^\gamma \mathbb{1}\{|Z|> 1/4\}\Big]}_{\text{Term 2}}.$$

To bound Term 1, we have that

$$|Z-\Delta|^\gamma \mathbb{1}\{|Z|\leq 1/4\} \leq 2^{-\gamma} \leq 2\gamma^{-\alpha},$$

where, in the last inequality, we use the fact that $\alpha \in (0,2]$ and that $2^{-x} \leq 2x^{-2}$ for all $x \geq 1$. It is clear then that Term 1 is bounded by $2\gamma^{-\alpha}$. To bound Term 2, we have that

$$\begin{aligned}|Z-\Delta|^\gamma \mathbb{1}\{|Z|>1/4\} &= |Z|^\gamma\left|1-\frac{\Delta}{Z}\right|^\gamma \mathbb{1}\{|Z|>1/4\} \\ &\leq |Z|^\gamma\left|1+\frac{1}{\gamma}\right|^\gamma \mathbb{1}\{|Z|>1/4\} \leq e|Z|^\gamma,\end{aligned}$$

where in the second inequality, we use the fact that $\Delta \leq \frac{1}{4\gamma}$. Therefore, we have that

$$\mathbb{E}\big[|Z-\Delta|^\gamma \mathbb{1}\{|Z|<1/4\}\big] \leq e\mathbb{E}|Z|^\gamma \leq Ca_2\gamma^{-\alpha}.$$

Combining the bounds on the two terms, we have that $\mathbb{E}|Z-\Delta|^\gamma \leq Ca_2\gamma^{-\alpha}$ as desired.

We now turn to the second claim. Without loss of generality, assume that $\Delta \geq 0$ so that, by symmetry of the distribution of Z, we have $\mathbb{E}\big[-|Z-\Delta|^{\gamma-1}\text{sgn}(Z-\Delta)\big] \geq 0$.

Since $\mathbb{E}\big[-|Z|^{\gamma-1}\text{sgn}(Z)\big] = 0$,

$$\begin{aligned}\mathbb{E}\big[-|Z-\Delta|^{\gamma-1}\text{sgn}(Z-\Delta)\big] &= \int_0^\Delta (\gamma-1)\mathbb{E}\big[|Z-t|^{\gamma-2}\big]\,dt \\ &\geq |\Delta|(\gamma-1)\mathbb{E}\big[|Z|^{\gamma-2}\big]\end{aligned}$$

For $\gamma \in [2,3)$, it holds that $\mathbb{E}\big[|Z|^{\gamma-2}\big] \geq \mathbb{E}|Z| \geq a_1$ since Z is supported on $[-1,1]$. For $\gamma \geq 3$, it holds that $\mathbb{E}\big[|Z|^{\gamma-2}\big] = \nu_{\gamma-2} \geq a_1(\gamma-2)^{-\alpha}$. Therefore, we have that

$$\mathbb{E}\big[-|Z-\Delta|^{\gamma-1}\mathrm{sgn}(Z-\Delta)\big] \geq \begin{cases} a_1|\Delta|(\gamma-1) & \text{if } \gamma \in [2,3), \\ a_1|\Delta|(\gamma-1)(\gamma-2)^{-\alpha} & \text{else.} \end{cases}$$

Thus, for all $\gamma \geq 2$, we have that

$$\mathbb{E}\big[-|Z-\Delta|^{\gamma-1}\mathrm{sgn}(Z-\Delta)\big] \geq \frac{a_1}{2}|\Delta|^{1-\alpha}.$$

Finally, we consider the third claim. The argument is similar to that of the first claim. We observe that

$$\begin{aligned}\mathbb{E}\left[\sup_{\gamma\in[2^k,2^{k+1}]}(|Z|+\Delta_\gamma)^{2(\gamma-1)}\right] &= \int_0^{\frac{1}{4}} \sup_{\gamma\in[2^k,2^{k+1}]}(z+\Delta_\gamma)^{2(\gamma-1)}\,dP(z) \\ &\quad + \int_{\frac{1}{4}}^1 \sup_{\gamma\in[2^k,2^{k+1}]}(z+\Delta_\gamma)^{2(\gamma-1)}\,dP(z). \qquad \text{(S5.28)}\end{aligned}$$

To bound the first term of (S5.28), we use the fact that $\Delta_\gamma \leq \frac{1}{4\gamma} \leq \frac{1}{4}$ and that $\alpha \in (0,2]$ to obtain

$$\int_0^{\frac{1}{4}} \sup_{\gamma\in[2^k,2^{k+1}]}(z+\Delta_\gamma)^{2(\gamma-1)}dP(z) \leq 2^{-2(2^k-1)} \leq 2^{-k\alpha}.$$

To bound the second term of (S5.28), we have

$$\begin{aligned}\int_{\frac{1}{4}}^1 \sup_{\gamma\in[2^k,2^{k+1}]}(z+\Delta_\gamma)^{2(\gamma-1)}dP(z) &\leq \int_{\frac{1}{4}}^1 \sup_{\gamma\in[2^k,2^{k+1}]} z^{2(\gamma-1)}(1+4\Delta_\gamma)^{2(\gamma-1)}dP(z) \\ &\leq e^2\mathbb{E}|Z|^{2(2^k-1)} \leq Ca_2 2^{-k\alpha}.\end{aligned}$$

The third claim of the lemma thus follows as desired. ■

Lemma S5.17 *Define* $L(\gamma,\mathbf{P}) := \frac{1}{\gamma}\min_\theta \log\big(\int|y-\theta|^\gamma\mathbf{P}(dy)\big) + \frac{1+\log\gamma}{\gamma} + \log\Gamma\left(1+\frac{1}{\gamma}\right)$ *for every* $\gamma \geq 2$. *Given* $\lim_{\gamma\to\infty}L(\gamma,\mathbf{P}_1) = \lim_{\gamma\to\infty}L(\gamma,\mathbf{P}_2) = \infty$, γ_1^* *being the unique minimizer of* $L(\gamma,\mathbf{P}_1)$, *and* $L(\gamma_1^*,\mathbf{P}_2) < \infty$, *we have that* γ_1^* *is the unique minimizer of* $L(\gamma,(1-\delta)\mathbf{P}_1+\delta\mathbf{P}_2)$ *for all small positive* δ.

Proof We first show that $\lim_{\gamma\to\infty}\inf_{0\leq\delta\leq1}L(\gamma,(1-\delta)\mathbf{P}_1+\delta\mathbf{P}_2) = \infty$. Given $M > 0$, there exists a $N \in \mathbb{N}$ such that $\frac{1}{\gamma}\min_\theta\log\big(\int|y-\theta|^\gamma\mathbf{P}_1(dy)\big) \vee \frac{1}{\gamma}\min_\theta\log\big(\int|y-\theta|^\gamma\mathbf{P}_2(dy)\big) > M$ for every $\gamma > N$, and thus

$$\begin{aligned}L(\gamma,(1-\delta)\mathbf{P}_1+\delta\mathbf{P}_2) &\geq \frac{1}{\gamma}\min_\theta\log\left[\int|y-\theta|^\gamma((1-\delta)\mathbf{P}_1+\delta\mathbf{P}_2)(dy)\right] \\ &\geq \frac{1}{\gamma}\min_\theta\left[(1-\delta)\log\int|y-\theta|^\gamma\mathbf{P}_1(dy) + \delta\log\int|y-\theta|^\gamma\mathbf{P}_2(dy)\right] \\ &\geq (1-\delta)\frac{1}{\gamma}\min_\theta\log\left(\int|y-\theta|^\gamma\mathbf{P}_1(dy)\right) + \delta\frac{1}{\gamma}\min_\theta\log\left(\int|y-\theta|^\gamma\mathbf{P}_2(dy)\right) \\ &\geq M, \text{for every } \gamma > N.\end{aligned}$$

For a fixed $\gamma \geq 2$, we have

$$\lim_{\delta \to 0^+} L(\gamma, (1-\delta)\mathbf{P}_1 + \delta \mathbf{P}_2) = \begin{cases} L(\gamma, \mathbf{P}_1), & \text{if } L(\gamma, \mathbf{P}_2) < \infty \\ \infty, & \text{otherwise.} \end{cases}$$

■

S5.2. Reference results

We use the following statement of Talagrand's inequality:

Theorem S5.18 *(Talagrand's Inequality; see e.g. Giné and Nickl (2016, Theorem 3.3.9)) Let $Z_1, \ldots, Z_n$ be independent and identically distributed random objects taking values in some measurable space $\mathcal{Z}$. Let $\mathcal{F}$ be a class of real-valued Borel measurable functions on $\mathcal{Z}$.*

Define $S_n = \sup_{f \in \mathcal{F}} \sum_{i=1}^n \{f(Z_i) - \mathbb{E}f(Z)\}$. Let $U > 0$ be a scalar that $\sup_{f \in \mathcal{F}} |f(Z)| \leq U$ almost surely; let $\sigma^2 := \sup_{f \in \mathcal{F}} \mathbb{E}f^2(Z)$. Then, for any $t > 0$,

$$\mathbb{P}(S_n - \mathbb{E}S_n \geq t) \leq \exp\left\{ -\frac{t^2}{2U \cdot \mathbb{E}S_n + n\sigma^2} \wedge \frac{t}{\frac{2}{3}U} \right\}.$$

We use the following bound on the expected supremum of the empirical process. For a class of real-valued functions $\mathcal{F}$ on some measurable domain $\mathcal{Z}$, we write $F(z) := \sup_{f \in \mathcal{F}} |f(z)|$ as its envelope function. For $\delta \in [0, 1)$, define the entropy integral

$$J(\delta) \equiv J(\delta, \mathcal{F}) := \int_0^\delta \sup_Q \sqrt{\log \mathcal{N}(\epsilon \|F\|_{L_2(Q)}, \mathcal{F}, L_2(Q))}\, d\epsilon, \tag{S5.29}$$

where the supremum is taken over all finitely discrete probability measures.

Lemma S5.19 *(Van Der Vaart and Wellner, 1996, Theorem 2.6.7) If $\mathcal{F}$ has finite VC dimension $V(\mathcal{F}) \geq 2$, then, for any $\epsilon \in (0, 1)$,*

$$N(\epsilon \|F\|_{L_2(Q)}, \mathcal{F}, L_2(Q)) \leq CV(\mathcal{F})(16e)^{V(\mathcal{F})} \left(\frac{1}{\epsilon}\right)^{2(V(\mathcal{F})-1)}.$$

Corollary S5.20 *If $\mathcal{F}$ has finite VC dimension $V(\mathcal{F})$, then, for any $\delta \in (0, 1]$,*

$$J(\delta) \leq C\sqrt{V(\mathcal{F})}\delta\sqrt{\log \frac{1}{\delta} \wedge \vee 1}.$$

Proof Using Lemma S5.19, we have that

$$\begin{aligned} J(\delta) &\leq \int_0^\delta \sqrt{CV(\mathcal{F}) + 2(V(\mathcal{F}) - 1)\log \frac{1}{\epsilon}}\, d\epsilon \\ &\leq C\sqrt{V(\mathcal{F})}\left\{\delta + \int_0^\delta \sqrt{\log \frac{1}{\epsilon}}\, d\epsilon\right\} \leq C\sqrt{V(\mathcal{F})}\left(\delta\sqrt{\log \frac{1}{\delta} \wedge \vee 1}\right). \end{aligned}$$

■

Theorem S5.21 (Van Der Vaart and Wellner (2011)) *Let* $F(x) := \sup_{f\in\mathcal{F}} |f(x)|$, $M := \max_{1\le i\le n} F(Z_i)$, *and* $\sigma^2 := \sup_{f\in\mathcal{F}} \mathbb{E}f(Z)^2$. *Then the following two bounds hold:*

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbb{E}f(Z)\right| \lesssim \frac{\|F\|_{L_2(P)}}{\sqrt{n}} J\left(\frac{\sigma}{\|F\|_{L_2(P)}}\right)\left[1 + \frac{\|M\|_{L_2(P)}\|F\|_{L_2(P)} J\left(\frac{\sigma}{\|F\|_{L_2(P)}}\right)}{\sqrt{n}\sigma^2}\right]$$

as well as

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbb{E}f(Z)\right| \lesssim \frac{\|F\|_{L_2(P)} J(1)}{\sqrt{n}}.$$