Robust regression using biased objectives

Matthew J. Holland* Kazushi Ikeda

Nara Institute of Science and Technology
Ikoma, Nara, Japan

Abstract

For the regression task in a non-parametric setting, designing the objective function to be minimized by the learner is a critical task. In this paper we propose a principled method for constructing and minimizing robust losses, which are resilient to errant observations even under small samples. Existing proposals typically utilize very strong estimates of the true risk, but in doing so require *a priori* information that is not available in practice. As we abandon direct approximation of the risk, this lets us enjoy substantial gains in stability at a tolerable price in terms of bias, all while circumventing the computational issues of existing procedures. We analyze existence and convergence conditions, provide practical computational routines, and also show empirically that the proposed method realizes superior robustness over wide data classes with no prior knowledge assumptions.

1 Introduction

Accurate prediction of response $y \in \mathbb{R}$ from novel pattern $\boldsymbol{x} \in \mathbb{R}^d$, based on an observed sample sequence of pattern-response pairs $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$, $\boldsymbol{z} \coloneqq (\boldsymbol{x}, y)$, is one of the most fundamental of statistical estimation tasks. Under particular assumptions such as bounded losses or sub-Gaussian residuals, a rich theory has developed in recent decades [29, 5, 2, 6, 40, 7], with variants of empirical risk minimization (ERM) routines playing a central role. The principle underlying such procedures is the use of the sample mean to approximate the risk (expected loss), which in turn functions as a location parameter of the unknown loss distribution. When the loss is concentrated around this value, this approximation is accurate, and ERM procedures perform well with appealing optimality properties [39].

Unfortunately, these assumptions are stringent, and in general, without *a priori* evidence of the contrary, our data cannot reasonably be expected to satisfy them. The fundamental problem manifests itself clearly in the simple setting of heavy-tailed real observations, in which the sub-optimality of the empirical mean is well-known [13]. A simple solution when using ERM is to leverage slower-growing loss functions (e.g., ℓ_1 instead of ℓ_2), but making this decision is inherently *ad hoc* and requires substantial prior information. Another option is model regularization [43, 7, 23], potentially combined with quantile regression [30, 42], though both methods introduce new parameters and we are faced with a difficult model selection problem [14], whose optimal solution is in practice often very sensitive to the empirical distribution. Put simply, in a non-parametric setting, one incurs a major risk of bias in the form of minimizing an impractical location parameter (e.g., the median under asymmetric losses), in order to ensure estimates are stable.

*Supported by the Grant-in-Aid for JSPS Research Fellows. Email: matthew-h@is.naist.jp.

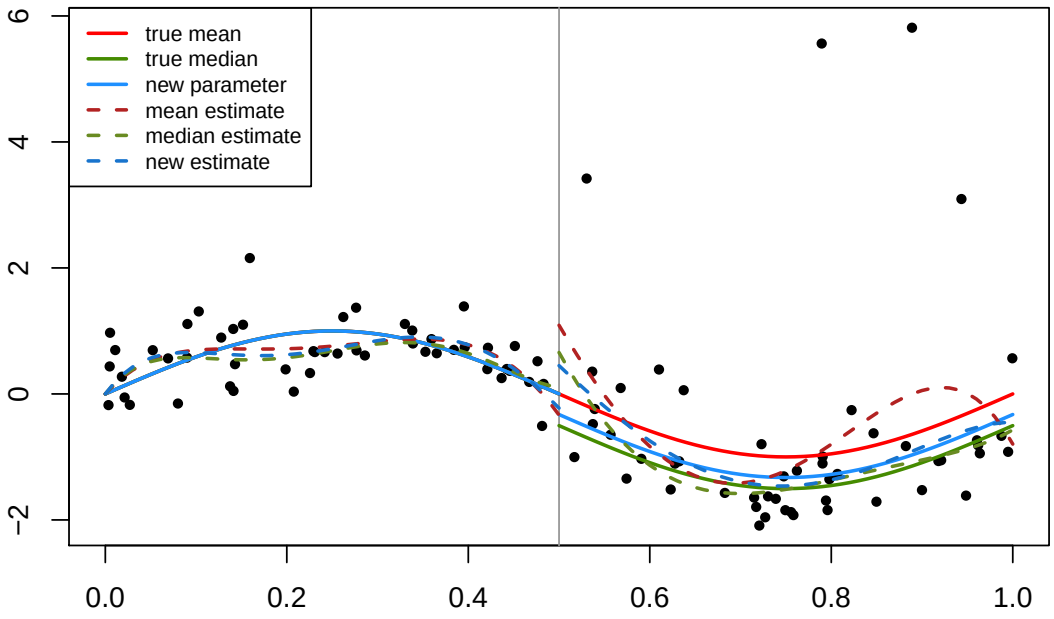

Figure 1: A one-dimension regression example (see Appendix C). When additive noise is heavy-tailed (the right half), estimating $\mathbf{E}(y;\boldsymbol{x})$ via least squares is difficult under small samples. On the other hand, estimating $\mathrm{med}(y;\boldsymbol{x})$ often introduces an unacceptable bias. In this paper we investigate "robust objectives" which act as all-purpose parameters to be estimated under diverse settings.

Considering these issues, it would be desirable to design an objective function which achieves the desired stability, but pays a smaller price in terms of bias, and therefore has minimal *a priori* requirements (Fig. 1). It is the objective of this paper to derive a regression algorithm which utilizes such a mechanism at tolerable computational cost. In section 2 we review the technical literature, giving our contributions against this backdrop. Section 3 introduces the core routine and important ideas underlying its construction in an intuitive manner, with formal justification and convergence analysis following in 4. Numerical performance tests are given in section 5, with key take-aways summarized in section 6.

2 Background and contributions

In this section we review the technical literature which is closely related to our work, and then within this context establish the main contributions made in this paper.

Related work Many tasks involve minimizing a function, say $L(\cdot)$, as a function of candidate $h \in \mathcal{H}$, which depends on the underlying distribution and is thus unknown. One line of work explicitly looks at refining the approximate objective function used. A key theme is to down-weight errant observations automatically, and to construct a new estimate $\widehat{L}(h) \approx L(h)$ of the risk, re-coding the algorithm as $\widehat{h} := \arg\min_{h\in\mathcal{H}} \widehat{L}(h)$. The now-classic work of Rousseeuw and Yohai [37] on S-estimators highlights important concepts in our work. They use the M-estimator of scale of the residual $h(\boldsymbol{x}) - y$, written $\widehat{s}(h)$, directly as objective function, setting $\widehat{L}(h) = \widehat{s}(h)$. The idea is appealing and has (classical) robustness properties, though serious

issues of stability and computational cost have been raised [28], and indeed even the fast modern routines are designed only for the rather special parametric setting where errant data can be discarded [38], which severely limits utility in our setting.

Re-weighting of extreme observations using M-estimators of the mean has been recently revisited by Catoni [12], later revised and published as Catoni [13]. A multi-dimensional extension of this theory appears in Audibert and Catoni [4], where they propose a function of the form

$$d(h, h') \coloneqq \lambda(\|h\|^2 - \|h'\|^2) + \mathbf{E}\, \psi_C \left(l(h; \boldsymbol{z}) - l(h'; \boldsymbol{z}) \right),$$

where $\lambda > 0$ is a user-set parameter, $l(h; \boldsymbol{z})$ is a penalty assigned to h on the event of observing $\boldsymbol{z}$, and ψ_C is a sigmoidal truncation function

$$\psi_C(u) \coloneqq \begin{cases} -\log(1 - u + u^2/2), & 0 \leq u \leq 1 \\ \log(2), & u \geq 1 \\ -\psi(-u), & u \leq 0. \end{cases}$$

The refined loss is then $\widehat{L}(h) = \sup\{d(h, h') : h' \in \mathcal{H}\}$, and is effectively a robust proxy of the "ridge risk" $\mathbf{E}\, l(h; \boldsymbol{z}) + \lambda \|h\|^2$. Many novel results are given, but it is not established whether an algorithm realizing the desired performance actually exists or not. More precisely, they show that one requires $\widehat{L}(\widehat{h}) = \inf_{h \in \mathcal{H}} \widehat{L}(h) + O(d/n)$ where d is model dimension. Unfortunately, construction of such a $\widehat{h}$ is left as future work, though a sophisticated iterative attempt is proposed by the authors. Another natural extension is given by Brownlees et al. [11], who directly apply these foundational results by using the Catoni class of M-estimators of risk, generalizing ψ_C above, to build $\widehat{L}$, which amounts to minimizing the root of the sample mean of $\{\psi_C(l(h; \boldsymbol{z}_i) - \theta)\}_{i=1}^n$ in θ. Novel bounds on excess risk are given, but this depends on an "optimal" scaling procedure which requires knowledge of the true variance. In addition, as this "robust loss" is defined implicitly, actually minimizing it is a non-trivial and expensive computational task.

Another interesting line of recent work revisits the merits of aggregation, a well-known notion from, for example, the bagging and boosting literature [9, 19]. The idea is to construct k candidates $\widehat{h}_{(1)}, \ldots, \widehat{h}_{(k)}$, typically by partitioning the data $D = \cup_{j=1}^k D_j$, and to aggregate them such that estimates derived from errant or uncharacteristic sub-samples are downweighted. One lucid example is the work of Minsker [33], who uses

$$\widehat{h} = \operatorname*{arg\,min}_{h \in \mathcal{H}} \sum_{i=1}^{k} \|h - \widehat{h}_{(i)}\|, \quad \widehat{h}_{(i)} \coloneqq \operatorname*{arg\,min}_{h \in \mathcal{H}} \frac{1}{|D_i|} \sum_{j \in D_i} l(h; \boldsymbol{z}_j)$$

namely the geometric median of the candidates (in norm $\|\cdot\|$), where each $\widehat{h}_{(i)}$ is the ERM estimate on the ith partition. The key notion here is that as long as most of the candidates are not overly poor, the aggregrate will be strong. This same notion was explored by Lerasle and Oliveira [31], where the "not overly poor" notion was made concrete with margin type conditions (section 5.1, page 14). As well, the work of Hsu and Sabato [24, 25] generalizes the formulation of these two works, casting the aggregation task as a "robust distance approximation," which is highly intuitive, is suggestive of algorithm design techniques, and yields tools applicable to many other problems [16, 32]. One major issue is that when sample sizes are small, very few partitions can be made. The key concern then is that when samples are large enough that k can be taken large, a less sophisticated method might already perform equally well on the full sample.

Our contributions In this work, the key idea is to use an approximate minimization technique to efficiently make use of powerful but computationally unwieldy robust losses. We propose a novel routine which is rooted in theoretical principles, but makes enough concessions to be useful in practice. Our main contributions can be summarized as follows:

- A fast minimizer of robust losses for general regression tasks, which is easily implemented, inexpensive, and requires no knowledge of higher-order moments of the data.
- Analysis of conditions for existence and convergence of the core routine.
- Comprehensive empirical performance testing, illustrating dominant robustness in both simulated settings and on real-world benchmark data sets.

Taken together, the theoretical and empirical insights suggest that we have a routine which behaves as we would expect statistically, converges quickly in practice, and which achieves a superior balance between cost and performance in the non-parametric setting standard to machine learning problems.

3 Fast minimization of robust objectives

In this section, we introduce the learning task of interest and give an intuitive derivation of our proposed algorithm. More formal analysis of the convergence properties of this procedure, from both statistical and computational viewpoints, is carried out in section 4.

A general learning task Given "candidate" $h \in \mathcal{H}$, member of a class of vectors or functions, and particular input/output instance $\boldsymbol{z} = (\boldsymbol{x}, y)$, we assign a penalty, $l(h; \boldsymbol{z}) \geq 0$ via loss function l—smaller is better—and evaluate the quality of h. Assuredly, doing this for a single observation $\boldsymbol{z}$ is insufficient; as this is a *learning* task, given incomplete prior information, we must choose h such that when we draw $\boldsymbol{z}$ randomly from an unknown probability distribution μ, representing unknown physical or social processes in our system of interest, the (random) quantity $l(h; \boldsymbol{z})$ is small. If the expected value $L_\mu(h) \coloneqq \mathbf{E}_\mu\, l(h; \boldsymbol{z})$, also called the *risk*, is small, then we expect the penalty $l(h; \boldsymbol{z})$ to be small on average. As such, a natural strategy is to choose a "best" candidate by the following program:

$$\min L_\mu(h), \quad \text{s.t. } h \in \mathcal{H}.$$

At this point, we run into a problem: μ is unknown, and thus L_μ is unknown. All we have access to is n independent draws of $\boldsymbol{z}$, namely the sample $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, and from this we must *approximate* the true objective, and then *minimize* this approximation as a proxy of L_μ.

Example 1 (Typical formulations). The pattern recognition problem has generic input space $\mathcal{X}$ and discrete labels, namely $\boldsymbol{x} \in \mathcal{X}$ and $y \in \{1, \ldots, C\}$. Here the "zero-one" loss $l(h; \boldsymbol{z}) = I\{h(\boldsymbol{x}) \neq y\}$ makes for a natural penalty to classifier h. More generally, the regression problem task has response $y \in \mathbb{R}$, and the classic metric for evaluating the quality of predictor $h : \mathcal{X} \to \mathbb{R}$ is the quadratic loss $l(h; \boldsymbol{z}) = (y - h(\boldsymbol{x}))^2$.

Issues to overcome Intuitively, if our approximation, say $\widehat{L}$, of L_μ, is not very accurate, then any minima of $\widehat{L}$ will likely be useless. Thus the first item to deal with is making sure the approximation $\widehat{L} \approx L_\mu$ is sharp. Perhaps the most typical approach is to set $\widehat{L}(h)$ to the sample mean, $\sum_{i=1}^n l(h; \boldsymbol{z}_i)/n$. In this case, the estimate is "unbiased" as $\mathbf{E}\, \widehat{L}(h) = L_\mu(h)$, but unfortunately the variance can be highly undesirable [12, 13]. There is no need to constrain

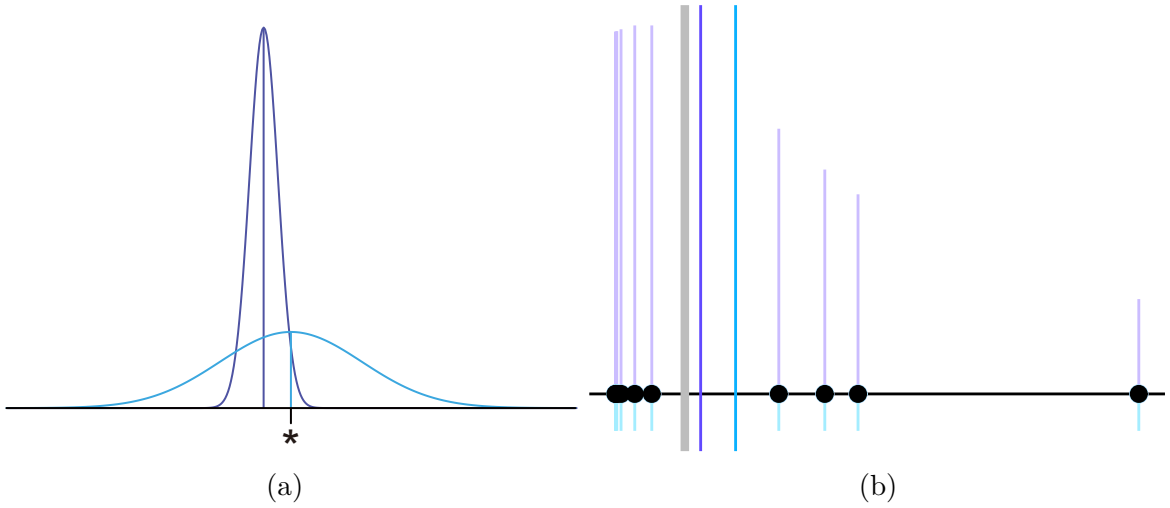

Figure 2: (a) Schematic of two estimators of L_μ (their density in n-sample space), one unbiased but with high variance (turquoise), another biased but concentrated (purple). (b) Points along the black line are observations $x_1, \ldots, x_n \in \mathbb{R}$ sampled from a heavy-tailed distribution ($n = 7$). The three vertical rules are: true mean (thick grey), sample mean (turquoise), and the M-estimate of location (purple). Vertical ranges associated with each point denote weight sizes, computed by $1/n$ (pale turquoise) and $\rho'(x_i - \gamma)/(x_i - \gamma)$ (pale purple). Down-weighting errant observations has a clear positive impact on estimates.

ourselves to unbiased estimators, as Figure 2(a) illustrates; paying a small cost in term of bias (allowing $\mathbf{E}\,\widehat{L}(h) \neq L_\mu(h)$) for much stabler output (large reduction in variance of $\widehat{L}$) is an appealing route.

One strategy to do this is as follows. Consider a "re-weighted" average approximation, namely $\widehat{L}(h; \boldsymbol{\alpha})$ given as

$$\widehat{L}(h; \boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i l(h; \boldsymbol{z}_i)$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ with $0 \leq \alpha_i \leq 1$ are our weights. In the sample mean case, $\alpha_i = 1/n$ for all observation points. However, since n is finite, one often runs into "errant" points which, when given the same amount of weight as all other points, do not accurately reflect the true underlying distribution. Thus, down-weighting these errant points by assigning them small weights (α_i near 0), and subsequently treating all the "typical" points as equals, should in principle allow us to overcome this issue. A mechanism which effectively does this for us is to use the M-estimate of location [26]; that is, to set

$$\widehat{L}(h; \rho, s) = \arg\min_{\theta} \sum_{i=1}^{n} \rho\left(\frac{l(h; \boldsymbol{z}_i) - \theta}{s}\right) \tag{1}$$

for each $h \in \mathcal{H}$. Here ρ is a convex function which is effectively quadratic around the origin, but grows much more slowly (Figure 3), and $s > 0$ is a scaling parameter. The re-weighting is implicit here, enacted via a "soft" truncation of errant points. Data points which are fairly close to the bulk of the sample are taken as-is (in the region where ρ is quadratic), while the impact of outlying points is attenuated (in the region where ρ is linear). We remark that such an estimator is assuredly biased in the sense that $\mathbf{E}\,\widehat{L}(h; \rho, s) \neq L_\mu(h)$ in most cases, but the desired impact is readily confirmed via simple tests, as in Figure 2(b).

Following such a strategy, the algorithm to run is

$$\widehat{h} = \arg\min_{h \in \mathcal{H}} \widehat{L}(h; \rho, s).$$

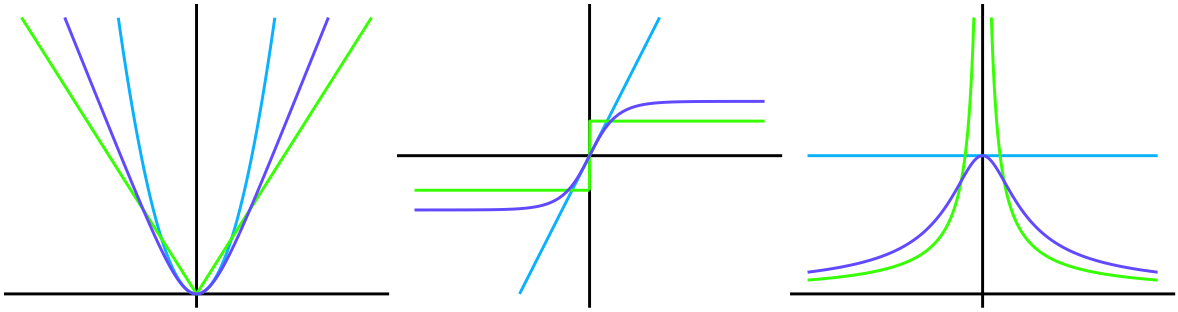

Figure 3: From left to right, each figure houses the graphs of $\rho(u)$, $\rho'(u)$, and $\rho(u)/u$ respectively. Colours denote different choices for ρ, namely the ℓ_2 loss (turquoise), the ℓ_1 loss (green), and the Gudermannian function (purple) from Example 3.

Given knowledge of the true variance, the utility of this approach from a statistical perspective has been elegantly analyzed by Brownlees et al. [11]. That we do not know the true variance is one issue; another critical issue is that this new "robust loss" $\widehat{L}(h;\rho,s)$ is defined *implicitly*, and is thus computationally quite uncongenial. Derivatives are not available in closed form, and every call to $\widehat{L}(h;\rho,s)$ requires an iterative sub-routine, a major potential roadblock. In what follows, we propose a principled, practical solution to these problems.

Deriving a fast minimizer Here we pursue an efficient routine for approximately minimizing the robust loss $\widehat{L}(h;\rho,s)$, in the context of the general regression task ($\boldsymbol{z} = (\boldsymbol{x}, y)$, with $y \in \mathbb{R}$). A useful heuristic strategy follows from noting that given any candidate $h \in \mathcal{H}$, and computing a central tendency metric γ (e.g., the median or average of $\{l(h;\boldsymbol{z}_i)\}_{i=1}^n$), since $l \geq 0$, in order for $\widehat{L}(h;\rho,s)$ to be small, it is *necessary* that the deviations $|l(h;\boldsymbol{z}_i) - \gamma|$ be small for most i. To see this, note that if most deviations are say larger than A, then there must be some points where $\widehat{L}$ is far to the right, that is i where

$$\widehat{L}(h;\rho,s) - l(h;\boldsymbol{z}_i) > A, \text{ which implies } \widehat{L}(h;\rho,s) > A.$$

With this condition in hand, note that the quantity

$$q(h) := \sum_{i=1}^{n} \rho\left(\frac{l(h;\boldsymbol{z}_i) - \gamma}{s}\right)$$

in fact directly measures these deviations. If most points are far away from γ, then $q(h)$ will be large; if most points are close to γ, then $q(h)$ will be small.

Our new task then, is to minimize $q(\cdot)$ in h. Fortunately, this can be done efficiently, using the re-weighting idea (see $\widehat{L}(h;\boldsymbol{u})$) discussed earlier. More precisely, let us set the weights to

$$\boxed{\alpha_i(h) = \rho'\left(\frac{(l(h;\boldsymbol{z}_i) - \gamma)}{s}\right) / \left(\frac{l(h;\boldsymbol{z}_i) - \gamma}{s}\right)}$$

For proper ρ (see section 4), we can ensure $0 \leq \alpha_i \leq 1$, and intuitively α_i will be very small when $l(h;\boldsymbol{z}_i)$ is inordinately far away from γ. Solving a re-weighted least squares problem, namely

$$\min \sum_{i=1}^{n} \alpha_i (y_i - g(\boldsymbol{x}_i))^2, \quad \text{s.t. } g \in \mathcal{H}$$

can typically be done very quickly, as Example 2 illustrates. What does this re-weighted least squares solution have to do with minimizing $q(\cdot)$? Fortunately, fixing any h, if we set update F as

$$\boxed{F(h) \coloneqq \arg\min_{g \in \mathcal{H}} \sum_{i=1}^{n} \alpha_i(h)(y_i - g(\boldsymbol{x}_i))^2}$$

then using classic results from the robust statistics literature [27, Ch. 7], we have that

$$q(F(h)) \leq q(h)$$

meaning the update from h to $F(h)$ is guaranteed to move us "in the right direction." That said, as our motivating condition was necessary, but not sufficient, the simplest approach is to *check* if this update actually monotonically improves the objective $\widehat{L}(\cdot; \rho, s)$, namely:

$$\boxed{\text{Update to } F(h) \text{ if and only if } \widehat{L}(F(h)) < \widehat{L}(h).}$$

The merits that this technique offers are clear: if we limit the number of iterations to T, then over $t = 1, 2, \ldots, T$ we need only compute $\widehat{L}$ *once* per iteration, meaning that the sub-routine for acquiring $\widehat{L}$ will only be called upon at most T times total. Initializing some $h_{(0)}$ and following the procedure just given, with re-centred (via the term γ) and re-scaled (via the factor s) observations at each step, we get Algorithm 1 below.

Algorithm 1 Fast robust loss minimizer (`fRLM`)

for $t \in [T]$ **do**
  $u_i \leftarrow \left(l(h_{(t-1)}; \boldsymbol{z}_i) - \gamma(D_{(t-1)})\right) / s(D_{(t-1)})$
  $\alpha_i \leftarrow \rho'(u_i)/u_i$ ▷ Downweight errant points; $i \in [n]$.
  $\widetilde{h} \leftarrow \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{n} \alpha_i(y_i - h(\boldsymbol{x}_i))^2$ ▷ Fast approximate update.
  $D_{(t)} \leftarrow \{l(\widetilde{h}; \boldsymbol{z}_i)\}_{i=1}^{n}$ ▷ Compute loss for new candidate.
  $\widehat{L}_{(t)} \leftarrow \arg\min_{\theta \in \mathbb{R}} \sum_{i=1}^{n} \rho\left(\frac{l(\widetilde{h}; \boldsymbol{z}_i) - \theta}{s(D_{(t)})}\right)$ ▷ Evaluate using robust loss.
  if $\widehat{L}_{(t)} < \widehat{L}_{(t-1)}$ **then** ▷ Check for monotonic improvement.
    $h_{(t)} \leftarrow \widetilde{h}$
  else
    return $h_{(t-1)}$
  end if
end for

Example 2 (Update under linear model). In the special case of a linear model where $h(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$ for some vector $\boldsymbol{w} \in \mathbb{R}^d$, then inverting a $d \times d$ matrix and then some matrix multiplication is all that is required. Writing $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T$ for the $n \times d$ design matrix, $\boldsymbol{y} = (y_1, \ldots, y_n)$, $h(X) = (h(\boldsymbol{x}_1), \ldots, h(\boldsymbol{x}_n))$, and $U = \operatorname{diag}(u_1, \ldots, u_n)$, then the solution is $(X^TUX)^\dagger X^TU(\boldsymbol{y} - h(X))$, where $(\cdot)^\dagger$ denotes the Moore-Penrose inverse. ■

Example 3 (Choice of ρ function). Extreme examples of ρ, the convex function used in (1), are the ℓ_2 and ℓ_1 losses, namely $\rho(u) = u^2$ and $\rho(u) = |u|$. These result in estimates of the sample mean and median respectively. A more balanced choice might be $\rho(u) = \log\cosh(u)$. We can also define ρ in terms of its derivative; for example, one useful choice is

$$\rho(u) = \int_0^u \psi(x)\,dx, \quad \psi(u) = 2\operatorname{atan}(\exp(u)) - \pi/2,$$

where ψ here is the function of Gudermann [1, Ch. 4], though there are numerous alternatives (see Appendix A). ■

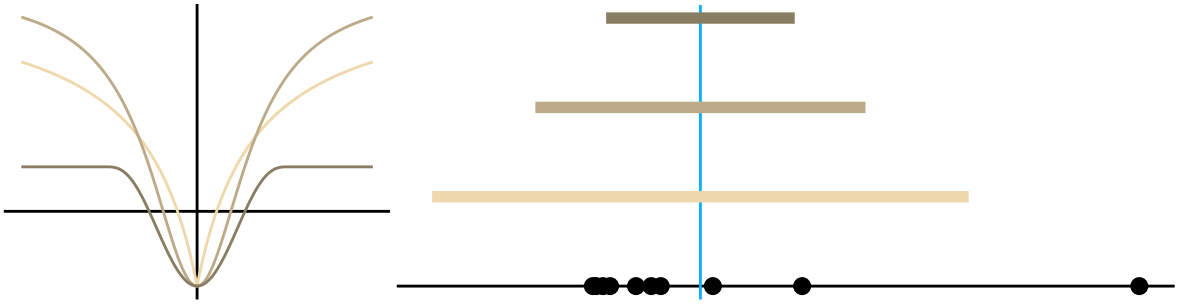

Figure 4: In the left plot, we have the graphs of three χ choices with common value $\chi(0)$. From light to dark brown, χ is respectively the absolute Geman-type, quadratic Geman-type, and Tukey function (see Example 4). In the right plot, we have randomly generated data D, and solved (2) using the three χ functions in the left plot (colours correspond), with $\gamma(D)$ as the sample mean (turquoise rule). Coloured horizontal rules in $\pm$ direction from $\gamma(D)$ represent $s(D)$ for each choice of χ.

Actual computation of key quantities Here we discuss precisely how we carry out the various sub-routines required in Algorithm 1, namely the tasks of initialization, re-centring, re-scaling, and finding robust loss estimates. Initialization is the first and the easiest: $h_{(0)}$ is initialized to the ℓ_2 empirical risk minimizer. When this value is optimal, it should be difficult to improve $\widehat{L}$, and thus the algorithm should finish quickly; when it is highly sub-optimal, this should result in a large value for $\widehat{L}(h_{(0)}; \rho, s)$, upon which subsequent steps of the algorithm seek to improve.

The "pivot" term γ is computed given a set of losses $D = \{l(h; \boldsymbol{z}_i)\}_{i=1}^n$ evaluated at some h; in particular, the losses are computed for $h_{(t-1)}$ at iteration t of Algorithm 1. This $\gamma(D)$ is used to centre the data; terms $l(h; \boldsymbol{z}_i)$ which are inordinately far away from $\gamma(D)$, either above or below, are treated as errant. One natural choice that requires sorting the data is the median D. A rough but fast choice is the arithmetic mean of D, which we have used throughout our tests.

As with γ, we carry out the re-scaling of our observations using D, denoting a set of losses. While there exist theoretically optimal scaling strategies [13], these require knowledge of $\mathrm{var}_\mu\, l(h; \boldsymbol{z})$ and setting of an additional confidence parameter. Since estimating second-order moments in order to estimate first-order moments is highly inefficient, we take the natural approach of using γ to centre the data, seeking a measure of how dispersed these losses are about this pivot, which will be our scale estimate. More concretely, for D induced by $h \in \mathcal{H}$, we seek any s satisfying

$$\sum_{i=1}^{n} \chi\left(\frac{l(h; \boldsymbol{z}_i) - \gamma(D)}{s}\right) = 0, \quad s > 0. \tag{2}$$

as our choice for $s(D)$. Here χ is an even function, assumed to satisfy $\chi(0) < 0$ and $\chi(u) > 0$ as $u \to \pm\infty$, ensuring that the scale is neither too big nor too small when compared with the deviations; see Figure 4 and Hampel et al. [22] for both theory and applications of this technique.

Our definition of $s(D)$ in (2) is implicit, as indeed is the robust loss computation $\widehat{L}$ in (1). We thus require iterative procedures to acquire sufficiently good approximations to these desired quantities. Updates taking a fixed-point form are typical for this sort of exercise, and we use the following two routines. Starting with the location estimate for h and given $s > 0$,

we run

$$\widehat{\theta}_{(k+1)} \leftarrow \widehat{\theta}_{(k)} + \frac{s}{n}\sum_{i=1}^{n}\rho'\left(\frac{l(h;\boldsymbol{z}_i)-\widehat{\theta}_{(k)}}{s}\right) \tag{3}$$

noting that this has the desired fixed point, namely a stationary point of the function in (1) to be minimized in θ. For the scale updates, centred by $\gamma \in \mathbb{R}$, we run

$$s_{(k+1)} \leftarrow s_{(k)}\left(1-\frac{1}{\chi(0)n}\sum_{i=1}^{n}\chi\left(\frac{l(h;\boldsymbol{z}_i)-\gamma}{s_{(k)}}\right)\right)^{1/2} \tag{4}$$

which has a fixed point at the desired root sought in (2).

Intuitively, for h and D, we expect that as $k \to \infty$

$$\widehat{\theta}_{(k)} \to \widehat{L}(h;\rho,s) \text{ and } s_{(k)} \to s(D),$$

and indeed such properties can be both formally and empirically established (see section 4.4).

Example 4 (Role of scale, choice of χ). Take the simple choice of $\chi(u) := u^2 - \beta$ for any fixed $\beta > 0$. If we have data set D with $|D| = n$, and let $\gamma(D)$ be the sample mean, then it immediately follows from (2) that $s(D) = (n-1)\,\mathrm{sd}(D)/(n\beta)$, namely a re-scaled sample standard deviation. Countless alternatives exist; one simple and useful choice is the Geman type function

$$\chi(u) = \frac{|u|^p}{1+|u|^p} - \beta, \quad p \in \{1,2\}$$

which originate in widely-cited image processing literature [21, 20] and also appear in machine learning work [46]. More classical choices include the bi-weight antiderivative of Tukey (see `tuk` in Appendix A), which has seen much use in robust statistics over the past half-century [22, Section 2.6]. ■

Summary of `fRLM` algorithm To recapitulate, we have put forward a procedure for minimizing the robust loss $\widehat{L}(h;\rho,s)$ in h, by using a fast re-weighted least squares technique that is guaranteed to improve a quantity (q above) very closely related to the actual unwieldy objective $\widehat{L}$. Using the iterative nature of this routine, we can perform the re-scaling and location estimates sequentially (rather than simultaneously), making for simple and fast updates. All together, this allows us to leverage the ability of ρ to truncate errant observations, while utilizing the fast approximate minimization program to alleviate issues with $\widehat{L}$ being implicit, all without using moment oracles for scaling as in the analysis of Catoni [13] and Brownlees et al. [11], which are notable merits of our proposed approach.

This algorithm makes use of statistical quantities that are defined as the minimizer of a class of estimators. As discussed in our literature review of section 2, the properties of learning algorithms that leverage these statistics have been analyzed by Brownlees et al. [11]. This does not, however, capture the properties of the resulting estimator itself: how does it behave as a function of sample size? Does it converge to a readily-interpreted parameter? We address these questions in the following section.

4 Analysis of convergence

In this section, we formulate the problem of interest with a bit more rigour in 4.1, give some fundamental existence results in 4.2, and then show that robust loss minimizers converges in a manner analogous to classical M-estimators in 4.3, using computationally convergent subroutines examined in 4.4. All proofs are relegated to Appendix B.

4.1 Preliminaries

Data model The learning problem, as discussed in the previous sections, is that of predicting response $y \in \mathbb{R}$, given an instance $\boldsymbol{x} \in \mathbb{R}^d$, based on a sequence of pairs $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ generated independently from an unknown distribution, $\boldsymbol{z} := (\boldsymbol{x}, y) \sim \mu$. Denote by $\mathcal{H}$ a collection of candidates $h : \mathcal{X} \to \mathbb{R}$ from which the learning algorithm will select an appropriate member. The task is of an "agnostic" nature, in that we do not know or assume knowledge of the relation between y and $\boldsymbol{x}$, all we want is to find an $h \in \mathcal{H}$ which reliably approximates $h(\boldsymbol{x}) \approx y$, without concern of identifying any true underlying model.

Evaluation mechanism To facilitate both formal analysis and the learning decision process, a loss function $l(h; \boldsymbol{z}) \geq 0$ will be utilized, which evaluates candidate h upon the random draw of $\boldsymbol{z}$, with smaller values being interpreted as more desirable, or a "better fit." We shall frequently use $\widehat{h}$ to denote the output of an algorithm, typically as $\widehat{h}_n(\boldsymbol{x}) := \widehat{h}(\boldsymbol{x}; \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$, a process which takes the n-sized data sample and returns a function $\widehat{h}_n \in \mathcal{H}$ to be used for prediction. A standard metric of generalization ability is the risk

$$L_\mu(h) := \mathbf{E}_\mu \, l(h; \boldsymbol{z}) = \int l(h; \cdot) \, d\mu.$$

One considers the performance of an algorithm $\widehat{h}_n$ to be good if the risk is sufficiently small, up to computational cost. Since μ is unknown, this can either be estimated formally, using inequalities that provide high-probability confidence intervals for this error over the random draw of the sample, or via controlled simulations where the performance metrics are computed over many independent trials.

Example 5. As a concrete case, the classical linear regression model with quadratic risk has $\boldsymbol{z} = (\boldsymbol{x}, y)$ with $h(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}$ for some $\boldsymbol{w} \in \mathbb{R}^d$, and $l(h; \boldsymbol{z}) = (y - \boldsymbol{w}^T \boldsymbol{x})^2$. When the model is correctly specified, i.e., when we have $y = \boldsymbol{w}_0^T + \epsilon$ for an unknown $\boldsymbol{w}_0 \in \mathbb{R}^d$, and noise $\mathbf{E}_\mu \, \epsilon = 0$, the loss takes on a convenient form, making additional results easy to obtain, though our general approach does not require such assumptions.

Additional notation We shall denote by μ a probability on $\mathbb{R}^{d+1}$, equipped with some appropriate σ-field, say the Borel sets $\mathcal{B}_{d+1}$. Let μ_n denote the empirical measure supported on the sample, namely $\mu_n(B) := n^{-1} \sum_{i=1}^n I\{\boldsymbol{z}_i \in B\}$, $B \in \mathcal{B}_{d+1}$. Expectation of vectors is naturally element-wise, namely $\mathbf{E}_\mu(\boldsymbol{x}, y) = (\mathbf{E}_\mu \, x_1, \ldots, \mathbf{E}_\mu \, x_d, \mathbf{E}_\mu \, y)$, and we shall use $\mathrm{var}_\mu \, \boldsymbol{z}$ to denote the $(d+1) \times (d+1)$ covariance matrix of $\boldsymbol{z}$, and so forth. $\mathbf{P}$ will be used to denote a generic probability measure, though in almost all cases it will be over the n-sized data sample, and thus correspond to the product measure μ^n. Let $[k] := \{1, \ldots, k\}$ for integer k.

4.2 Existence of valid estimates

Generalization performance is completely captured by the *distribution* of $l(h; \boldsymbol{z})$. Unfortunately, inferring this distribution from a finite sample is exceedingly difficult, and so we estimate parameters of this distribution to gain insight into performance; the expected value $L_\mu(h)$ is a case in point. In pursuit of a routine for estimating the risk, with low variance and controllable risk, the basic strategy ideas in section 3 seem intuitively promising. Here we show that following the strategy outlined, one can create a procedure which is valid in a statistical sense, under very weak assumptions.

Our starting point is to introduce new parameters, distinct from the risk, which have controllable bias, and can be approximated more reliably than the expected value, using a finite sample. The following definition specifies such a parameter class.

Definition 6 (General target parameters)**.** For $\rho : \mathbb{R} \to [0, \infty)$ and scale $s > 0$, define

$$\theta^*(h) \in \arg\min_{\theta \in \mathbb{R}} \mathbf{E}_\mu \, \rho\left(\frac{l(h; \boldsymbol{z}) - \theta}{s}\right) \tag{5}$$

where s may depend on h. We require that ρ be symmetric about 0, with $\rho(0) = 0$, and further that

$$\rho(u) = O(u), \text{ as } u \to \pm\infty$$
$$\frac{\rho(u)}{u^2} \to K < \infty, \text{ as } u \to 0.$$

For clean notation, normalize such that $K = 1/2$. If ρ is differentiable, denote $\psi := \rho'$. If twice-differentiable and $\psi' > 0$, say that ρ *specifies a robust objective*, namely $\theta^*(\cdot)$.

Remark 7. The logic here is as follows: the mean $L_\mu(h)$ can be considered a good target if the data are approximately symmetric, or if (regardless of symmetry) they are tightly concentrated about the mean. In both of these cases, we have $\theta^*(h) \approx L_\mu(h)$. To see this, If $l(h; \boldsymbol{z})$ is symmetric about some l_0, that is to say for all $\varepsilon > 0$,

$$\mathbf{P}\{l(h; \boldsymbol{z}) - l_0 \geq \varepsilon\} = \mathbf{P}\{-(l(h; \boldsymbol{z}) - l_0) \geq \varepsilon\},$$

it is sufficient to minimize

$$\int_{\{l(h;\boldsymbol{z}) \geq l_0\}} \rho\left(\frac{l(h; \boldsymbol{z}) - \theta}{s}\right) d\mu$$

on $[l_0, \infty)$, where $\theta = l_0 = L_\mu(h)$ is a solution. Thus in the symmetric case, we end up with $\theta^*(h) = L_\mu(h)$, irrespective of scaling and truncating mechanisms. Here "tightly concentrated" is relative, in the sense that

$$|l(h; \boldsymbol{z}) - L_\mu(h)| < s$$

with high probability. Since we have required $\rho(u) \sim u^2$, tight concentration would imply $\theta^*(h) \approx L_\mu(h)$. As for the linear growth requirement, $\rho(u) = o(u^2)$ as $u \to \pm\infty$ is necessary if we are to reduce dependence on the tails, but making the much stronger requirement of $\rho(u) = O(u)$ is very useful as it implies that ψ is bounded. Note that of the functions ρ given in Example 3, the ℓ_p choices do not meet our criteria, but the Gudermannian and $\log\cosh$ choices both satisfy all conditions. ■

This $\theta^*(\cdot)$, a new parameter of the loss $l(\cdot; \boldsymbol{z})$, can be readily interpreted as an alternative performance metric to the risk $L_\mu(\cdot)$. Denote optimal performance in this metric on $\mathcal{H}$ by

$$\theta^*(\mathcal{H}) := \inf_{h \in \mathcal{H}} \theta^*(h) \geq 0 \tag{6}$$

and the empirical estimate of these parameters by

$$\widehat{\theta}(h) \in \arg\min_{\theta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \rho\left(\frac{l(h; \boldsymbol{z}_i) - \theta}{s}\right). \tag{7}$$

Note that we call this the empirical estimate as we have simply replaced μ by μ_n in the definition of θ^* to derive $\widehat{\theta}$. The procedure of Algorithm 1 outputs an approximation of

$$\widehat{h}_n \in \arg\min_{h \in \mathcal{H}} \widehat{\theta}(h) \tag{8}$$

which is none other than a minimizer of the robust loss $\widehat{\theta}$, an empirical estimate of the alternative performance metric θ^*.

First, we show that these new "objectives" are indeed well-defined objective functions, which is important since our algorithm seeks to minimize them.

Lemma 8 (Existence of parameter and its estimate)**.** *Let ρ specify a robust objective $\theta^*(h)$. This function is well-defined in h, in that for each $h \in \mathcal{H}$, the value of $\theta^*(h)$ is uniquely determined, characterized by*

$$\mathbf{E}_\mu \, \psi\left(\frac{l(h;\boldsymbol{z}) - \theta^*(h)}{s}\right) = 0. \tag{9}$$

Analogously, the empirical estimate is uniquely defined, and almost surely given by

$$\sum_{i=1}^{n} \psi\left(\frac{l(h;\boldsymbol{z}_i) - \widehat{\theta}(h)}{s}\right) = 0. \tag{10}$$

With a well-defined objective function, next we consider the existence of the minimizer of this new objective. While measurability is by no means our chief concern here, for completeness we include a technical result useful for proving the existence of a valid minimizer of the proxy objective.

Lemma 9. *Let ρ be even and continuously differentiable with ρ' non-decreasing on $\mathbb{R}$. Let $s_h : \mathbb{R}^{d+1} \to \mathbb{R}_+$ be measurable for all $h \in \mathcal{H}$. For any $n \in \mathbb{N}$, denote sequence space $\mathcal{Z} \coloneqq (\mathbb{R}^{d+1})^n$. Then defining*

$$\widehat{\theta}(h) \coloneqq \inf\left(\operatorname*{arg\,min}_{u\in\mathbb{R}} \sum_{i=1}^{n} \rho\left(\frac{l(h;\boldsymbol{z}_i) - u}{s_h(\boldsymbol{z}_i)}\right)\right), \tag{11}$$

we have that $\widehat{\theta}$ is measurable as a function on $\mathcal{H} \times \mathcal{Z}$.

This gives us a formal definition of $\widehat{\theta}(h)$ which has the desired property specified by (7). It simply remains to show that we can always minimize this objective in h.

Theorem 10 (Existence of minimizer)**.** *Let $h \mapsto s_h$ be continuous and $s_h > 0$, $h \in \mathcal{H}$. Using $\widehat{\theta}$ from Lemma 9, define*

$$\widehat{\theta}(\mathcal{H}) \coloneqq \inf_{h\in\mathcal{H}} \widehat{\theta}(h). \tag{12}$$

For any ρ specifying a robust objective (Defn. 6), and any sample $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$,

$$\exists\, \widehat{h} \in \mathcal{H}, \quad \widehat{\theta}(\widehat{h}) = \widehat{\theta}(\mathcal{H}),$$

and there exists a random variable $\widehat{h}_n$ such that $\mathbf{P}\{\widehat{\theta}(\widehat{h}_n) = \widehat{\theta}(\mathcal{H})\} = 1$.

There are many potential methods for carrying out the scaling in practice. Here we verify that the simple method proposed in section 3 does not disrupt the assurances above. First a definition.

Definition 11 (General-purpose scale)**.** For random variable $x \sim \nu$, introduce even function $\chi : \mathbb{R} \to \mathbb{R}$, non-decreasing on $\mathbb{R}_+$, which satisfies

$$0 < \lim_{|u|\to\infty} \chi(u), \quad \chi(0) < 0.$$

Let $\beta \geq 0$ be the value such that $\chi(0) = -\beta$. With the help of χ and pivot term γ_ν which may depend on ν, define

$$\sigma_\nu \coloneqq \inf\left\{\sigma > 0 : \mathbf{E}\,\chi\left(\frac{x - \gamma_\nu}{\sigma}\right) = 0\right\}. \tag{13}$$

With this definition in place, substituting $\nu = \mu_n$ yields an empirical scale estimate

$$s_h = \inf\left\{\sigma > 0 : \sum_{i=1}^{n} \chi\left(\frac{l(h; \boldsymbol{z}_i) - \gamma_{\mu_n}(h)}{\sigma}\right) = 0\right\} \tag{14}$$

with $\sum_{i=1}^{n} l(h; \boldsymbol{z}_i)/n$ a natural pivot value, though we certainly have more more freedom in constructing $\gamma_{\mu_n}(h)$, as the following result shows.

Proposition 12 (Validity of scaling mechanism)**.** *If $\gamma_{\mu_n}(h) < \infty$ almost surely for all $h \in \mathcal{H}$, and χ (Defn. 11) is increasing on $\mathbb{R}_+$, then the minimizer $\widehat{h}_n$ (8) as constructed in Theorem 10 satisfies*

$$\widehat{\theta}(\widehat{h}_n) = \widehat{\theta}(\mathcal{H})$$

almost surely when scaling with $s = s_h$ as in (14).

Note that $\gamma_{\mu_n}(h)$ here corresponds directly to $\gamma(D)$ in Algorithm 1, where $D = \{l(h; \boldsymbol{z}_i)\}_{i=1}^{n}$. With basic facts related to existence and measurability in place, we proceed to look at some convergence properties of the estimators and computational procedures concerned in the sections 4.3–4.4.

4.3 Statistical convergence

For some context, we start with a well-known consistency property of M-estimators, adapted to our setting.

Theorem 13 (Pointwise consistency under known scale)**.** *For any ρ specifying a robust objective, fixing any $h \in \mathcal{H}$ and $s > 0$, then*

$$\mathbf{P}\left\{\lim_{n\to\infty} \widehat{\theta}(h) = \theta^*(h)\right\} = 1.$$

Note that this strong consistency result is "pointwise" in the sense that the event of probability 1 is dependent on the choice of $h \in \mathcal{H}$. Were we to take a different $h' \in \mathcal{H}$, while the probability would still be one, the events certainly need not coincide. This becomes troublesome since $\widehat{h}_n$ will in all likelihood take a different h value for distinct samples $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$. Intuitively, we do expect that as n grows, the estimate $\widehat{h}_n$ should get progressively better and in the limit we should have

$$\widehat{\theta}(\widehat{h}_n) \to \theta^*(\mathcal{H}), \quad n \to \infty.$$

Here we show that such a property does indeed hold, focusing on the case where $\mathcal{H}$ is a linear model, though the assumptions on $\boldsymbol{x}$ and y are still completely general (agnostic). More precisely, we assume that $\mathcal{H}$ is defined by a collection of real-valued functions $\varphi_1, \ldots, \varphi_k$ on $\mathbb{R}^d$, and a bounded parameter space $\mathcal{W} \subset \mathbb{R}^k$. The model is thus of the form

$$\mathcal{H} = \left\{h = \sum_{j=1}^{k} w_j \varphi_j : (w_1, \ldots, w_k) \in \mathcal{W}\right\}. \tag{15}$$

Under this model, the class of parameters given in Defn. 6 and the corresponding estimators (7) are such that convenient uniform convergence results are available using standard combinatorial arguments. First a general lemma of a technical nature.

Lemma 14 (Uniform strong convergence)**.** *Let $\mathcal{H}$ satisfy (15), and ρ specify a robust objective (Defn. 6). Denoting $\Lambda \coloneqq \mathcal{H} \times \mathbb{R} \times \mathbb{R}_+$, $\lambda \coloneqq (h, u, s) \in \Lambda$, and*

$$\psi(\boldsymbol{z}; \lambda) \coloneqq \psi\left(\frac{l(h; \boldsymbol{z}) - u}{s}\right),$$

we have that

$$\lim_{n\to\infty} \sup_{\lambda\in\Lambda} \left| \frac{1}{n} \sum_{i=1}^{n} \psi(\boldsymbol{z}_i; \lambda) - \mathbf{E}_\mu \, \psi(\boldsymbol{z}; \lambda) \right| = 0$$

almost surely.

A corollary of this general result will be particularly useful.

Corollary 15. *The robust objective minimizer $\widehat{h}_n$ defined in (8), equipped with any scaling mechanism s depending on $\widehat{h}_n$ (and thus potentially random), satisfies*

$$\lim_{n\to\infty} \mathbf{E}_\mu \, \psi\left(\frac{l(\widehat{h}_n; \boldsymbol{z}) - \widehat{\theta}(\widehat{h}_n)}{s}\right) = 0$$

almost surely.

These facts are sufficient for showing that a very natural analogue of the strong pointwise consistency of M-estimators (Theorem 13) holds in a uniform fashion for our robust objective minimizer $\widehat{h}_n$.

Theorem 16 (Consistency analogue)**.** *Let $\widehat{h}_n$ be determined by (8) equipped with any fixed scaling mechanism $s_h : \mathbb{R}^{d+1} \to \mathbb{R}_+$. Let ρ specify a robust objective, with ρ' concave on $\mathbb{R}_+$. If there exists constants s_1, s_2, ϵ such that*

$$0 < s_1 \leq s_h(\boldsymbol{z}) \leq s_2 < \infty$$
$$0 < \epsilon \leq \inf_{h\in\mathcal{H}} \mathbf{E}_\mu \, \psi'(l(h; \boldsymbol{z})/s_1)$$

then it follows that

$$\mathbf{P}\left\{ \lim_{n\to\infty} \widehat{\theta}(\widehat{h}_n) = \theta^*(\mathcal{H}) \right\} = 1.$$

That is, $\widehat{\theta}(\widehat{h}_n)$ is a strongly consistent estimator of the optimal value $\theta^*(\mathcal{H})$.

With these rather natural statistical properties understood, we shift our focus to the behaviour of the computational routines used.

4.4 Computational convergence

As regards computational convergence, since Algorithm 1 is meant to be a fast approximation to a minimizer of $\widehat{L}(\cdot)$ on $\mathcal{H}$, we should not expect the $\widehat{h}$ produced after $t \to \infty$ iterations to actually converge to the true $\widehat{h}_n$ in (8). What we should expect, however, is that the subroutines (3) and (4), used to compute $\widehat{L}_{(t)}$ and $s(D_{(t)})$ for *each* t, should converge to the true values specified by (1) and (2) respectively. We show that this convergence holds.

Proposition 17 (Convergence of updates)**.** *Let ρ specify a robust objective (Defn. 6). Fixing $s > 0$, and any initial value $\widehat{\theta}_{(0)}$, the iterative update $(\widehat{\theta}_{(k)})$ specified in (3) satisfies*

$$\lim_{k\to\infty} \widehat{\theta}_{(k)} = \widehat{\theta}(h),$$

recalling that $\widehat{\theta}(h) = \widehat{L}(h; \rho, s)$ from section 3. Similarly, for χ as specified by Defn. 11, under some additional regularity conditions on χ, (see proof) we have that for any initialization $s_{(0)} > 0$, the update $(s_{(k)})$ in (4) satisfies

$$\sum_{i=1}^{n} \chi\left(\frac{l(h; \boldsymbol{z}_i) - \gamma}{\lim_{k\to\infty} s_{(k)}}\right) = 0.$$

Using ρ as in Defn. 6 and χ as in Prop. 12, note that the above convergence guarantees will not be ambiguous, since the location and scale estimates are uniquely determined.

Efficiency of iterative sub-routines As a complement to the formal convergence properties just examined, we conduct numerical tests in which we run (3) and (4) until they respectively compute the true $\widehat{\theta}$ and s values up to a specified degree of precision. It is of practical importance to answer the following questions: Do the iterative routines reliably converge to the correct optimal value? How many iterations does this take on average? How does this depend on factors such as the data distribution, sample size, and our choice of ρ and χ?

To investigate these points, we carry out the following procedure. Generating $x_1, \ldots, x_n \in \mathbb{R}$ from some distribution, denote

$$f_1(u) := \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{x_i - u}{s}\right), \quad f_2(u) := \frac{1}{n}\sum_{i=1}^{n} \chi\left(\frac{x_i - \gamma}{u}\right).$$

The location task is to minimize f_1 on $\mathbb{R}$, and the scale task is to seek a root of f_2 on $\mathbb{R}_+$. Two choices of distribution were used. First is $x \sim N(0, 3)$, i.e., centred Normal random variables with variance of nine. The second is asymmetric and heavy-tailed, generated as $\exp(x)$ where x is again $N(0, 3)$; this is the log-Normal distribution. For f_1, the s value is a parameter; this is set to the standard deviation of the x_i. For f_2, the γ value is a parameter; this is set to the sample mean of the x_i. As for ρ and χ, we examine five choices of each, all defined in Appendix A. Initial values are $\widehat{\theta}_{(0)} = n^{-1}\sum_{i=1}^{n} x_i$ and $s_{(0)} = \mathrm{sd}\{x_i\}_{i=1}^{n}$.

In Figure 5, we show the average *iterations to converge*, as a function of sample size n, computed as follows. The terminating iteration for these tasks, at accuracy level ε, is defined

$$K_\varepsilon(\widehat{\theta}) := \min\{k : |\widehat{\theta}_{(k)} - \widehat{\theta}_{OR}| \le \varepsilon\}, \quad K_\varepsilon(s) = \min\{k : |s_{(k)} - s_{OR}| \le \varepsilon\}$$

where $\widehat{\theta}_{OR}$ and s_{OR} are "oracle" values of the minimum/root of f_1/f_2. These are obtained via `uniroot` in R [36], an implementation of Brent's univariate root finder [10], recalling the ρ minimization can be cast as a root-finding problem (Lemma 8). These K_ε values are thus the number of iterations required; 100 independent trials are carried out, and the arithmetic mean of these values is taken. Updates $\widehat{\theta}_{(k)}$ and $s_{(k)}$ are precisely as in (3) and (4). Accuracy level is $\varepsilon = 10^{-4}$ for all trials.

We have convergence at a high level of precision, requiring very few iterations, and this holds uniformly across the conditions observed. As such, the convergence of the routines is just as expected (Proposition 17), and the speed is encouraging. In general, convergence tends to speed up for larger n, and the relative difference in speed is very minor across distinct ρ choices, though slightly more pronounced in the case of χ, but even the slowest choice seems tolerable. Finally, location estimation is slightly slower in the Normal case than in the log-Normal case, while the opposite holds for scale estimation.

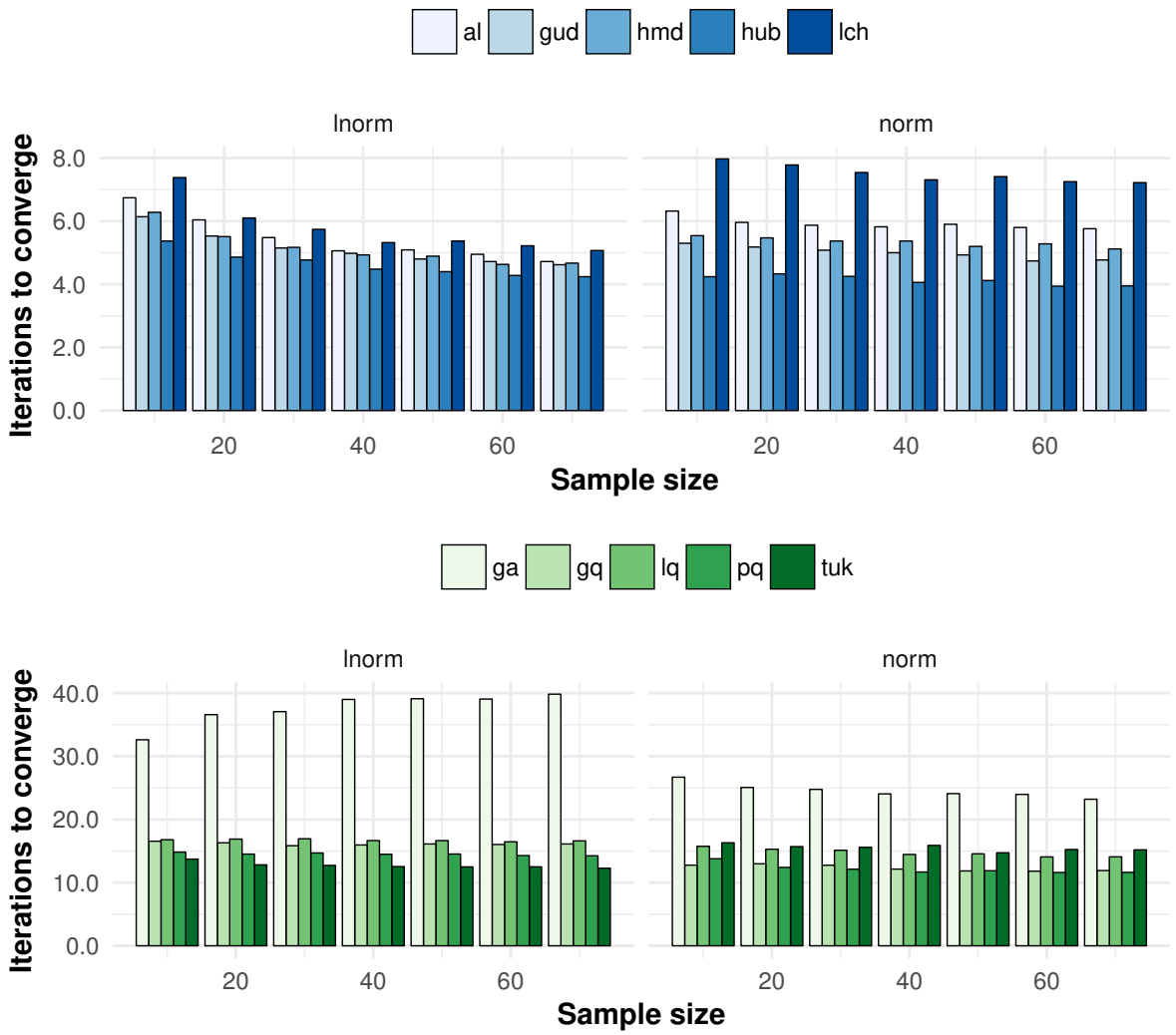

Figure 5: Iterations required to reach ε-accurate estimates given n sample, under Normal/log-Normal observations. Top row: average $K_\varepsilon(\widehat{\theta})$. Bottom row: average $K_\varepsilon(s)$. See appendix A for ρ and χ definitions.

5 Numerical performance tests

We derived a new algorithm in 3, formally investigated statistical properties in 4.2–4.3, and computational properties in 4.4. Here we evaluate the actual performance of this algorithm against standard competitive algorithms in a variety of situations, including both tightly controlled numerical simulations and real-world benchmark data sets. We seek to answer the following questions.

1. How well does `fRLM` (Algorithm 1) generalize off-sample?
2. Fixing ρ, can we still succeed under both light- and heavy-tailed noise?
3. How does performance depend on n and d?

Our experimental setup and competing algorithms used are described in 5.1–5.2, and results follow in 5.3–5.4 where we give concrete responses to all the questions posed above. All

experimental parameters, as well as source code for all methods used, are included in the supplementary source code.[1]

5.1 Experimental setup

Every experimental condition and trial has us generating n training observations, of the form $y_i = \boldsymbol{w}_0^T \boldsymbol{x} + \epsilon_i, i \in [n]$. Distinct experimental conditions are specified by the setting of (n, d) and μ. Inputs $\boldsymbol{x}$ are assumed to follow a d-dimensional isotropic Gaussian distribution, and thus to determine μ is to specify the distribution of noise ϵ. In particular, we look at several families of distributions, and within each family look at 15 distinct *noise levels*. Each noise level is simply a particular parameter setting, designed such that $\mathrm{sd}_\mu(\epsilon)$ monotonically increases over the range 0.3–20.0, approximately linearly over the levels.

To ensure a wide range of signal/noise ratios is spanned, for each trial, $\boldsymbol{w}_0 \in \mathbb{R}^d$ is randomly generated as follows. Defining the sequence $w_k := \pi/4 + (-1)^{k-1}(k-1)\pi/8, k = 1, 2, \ldots$ and uniformly sampling $i_1, \ldots, i_d \in [d_0]$ with $d_0 = 500$, we set $\boldsymbol{w}_0 = (w_{i_1}, \ldots, w_{i_d})$. As such, given our control of noise standard deviation, and noting that the signal to noise ratio in this setting is computed as $\mathrm{SN}_\mu = \|\boldsymbol{w}_0\|_2^2 / \mathrm{var}_\mu(\epsilon)$, the ratio ranges between $0.2 \leq \mathrm{SN}_\mu \leq 1460.6$.

Regarding the noise distribution families, the tests described above were run for 27 different families, but as space is limited, here we provide results for four representative families: log-Normal (denoted `lnorm` in figures), Normal (`norm`), Pareto (`pareto`), and Weibull (`weibull`). Even with just these four, we capture both symmetric and asymmetric families, sub-Gaussian families, as well as heavy-tailed families both with and without finite higher-order moments.

Our chief performance indicator is *prediction error*, computed as follows. For each condition and each trial, an independent test set of m observations is generated identically to the corresponding n-sized training set. All competing methods use common sample sets for training and are evaluated on the same test data, for all conditions/trials. For each method, in the kth trial, some estimate $\widehat{\boldsymbol{w}}$ is determined. To approximate the ℓ_2-risk, compute root mean squared error $e_k(\widehat{\boldsymbol{w}}) := (m^{-1} \sum_{i=1}^m (\widehat{\boldsymbol{w}}^T \boldsymbol{x}_{k,i} - y_{k,i})^2)^{1/2}$, and output prediction error as the average of normalized errors $e_k(\widehat{\boldsymbol{w}}(k)) - e_k(\boldsymbol{w}_0(k))$ taken over all trials. While n and d values vary, in all experiments the number of trials is fixed at 250, and test size $m = 1000$.

5.2 Competing methods

Benchmark routines used in these experiments are as follows. Ordinary least squares, denoted `ols` and least absolute deviations, denoted `lad`, represent classic methods. In addition, we look at three very modern alternatives, namely three routines directly from the references papers of Minsker [33] (`geomed`), Brownlees et al. [11] (`bjl`), and Hsu and Sabato [25] (`hs`). The `hs` routine used here is a faithful R translation of the MATLAB code published by the authors. Our implementation of `geomed` uses the geometric median algorithm of [45, Eqn. 2.6], and all partitioning conditions as given in the original paper are satisfied. Regarding `bjl`, scaling is done using a sample-based estimate of the true variance bound used in their analysis, with optimization carried out using the Nelder-Mead gradient-free method implemented in the R function `optim`.

For our `fRLM` (Algorithm 1, section 3), we tried several different choices of ρ and χ, including those in Appendix A, and overall trends were almost identical. Thus as a representative, we use the Gudermannian for ρ and $\chi(u) = \mathrm{sign}(|u| - 1)$ as a particularly simple and illustrative example implementation. Estimates of location and scale were carried out by (3) and (4).

[1] All materials available at `https://github.com/feedbackward/rtm_code`.

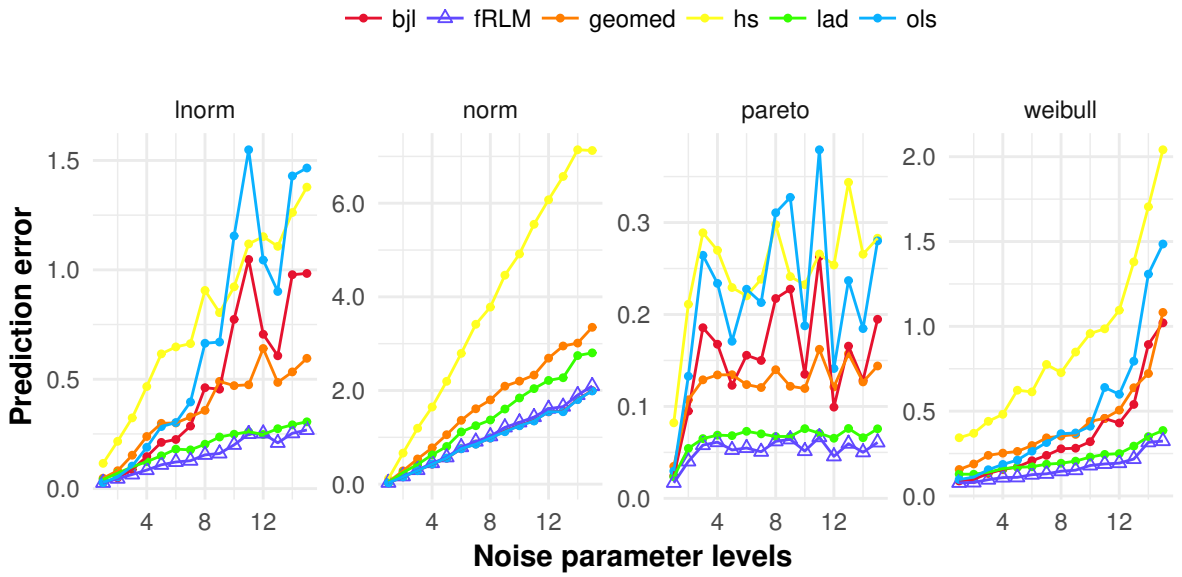

Figure 6: Prediction error as a function of noise level, with $n = 15$ and $d = 5$. Moving from left to right on the horizontal axis corresponds to larger noise magnitude.

5.3 Test results: simulation

Here we assemble the results of distinct experiments which highlight different facets of the statistical procedures being evaluated.

Performance over noise levels Fig. 6 shows how predictive performance deteriorates as the noise magnitude (described in 5.1) grows larger, under fixed (n, d) setting. We see that our method closely follow the performance of `ols` only when it is strong (the Normal case), but critically remain stable under settings in which `ols` deteriorates rapidly (all other cases). Our method, much like the other robust methods, incurs a bias by designing objective functions using estimators for target parameters other than the true risk. It is clear, however, that the bias in the case of our method is orders of magnitude smaller than that of competing routines, suggesting that the proposed procedure for minimizing a robust loss is effective. Note that `bjl` needs an off-the-shelf non-linear optimizer and directly requires variance estimates; our routine circumvents these steps, and is seen to be better for it.

Impact of sample size (n grows, d fixed) In Fig. 7 we look at prediction error, at the middle noise level, for different settings of n under a fixed d. We have fixed $d = 5$ and the sample size ranges between 12–122. Once again we see that in the Normal case where `ols` is optimal, our routine closely mimics its behaviour and converge in the same way. On the other hand for the heavier-tailed settings, we find that the performance is once against extremely strong, with far better performance under small sample sizes, and a uniformly dominant rate of convergence as n gets large.

Impact of dimension The role played by model dimension is also of interest, and can highlight weaknesses in optimization routines that do not appear when only a few parameters are being determined. Such issues are captured most effectively by keeping the d/n ratio fixed and increasing the model dimension.

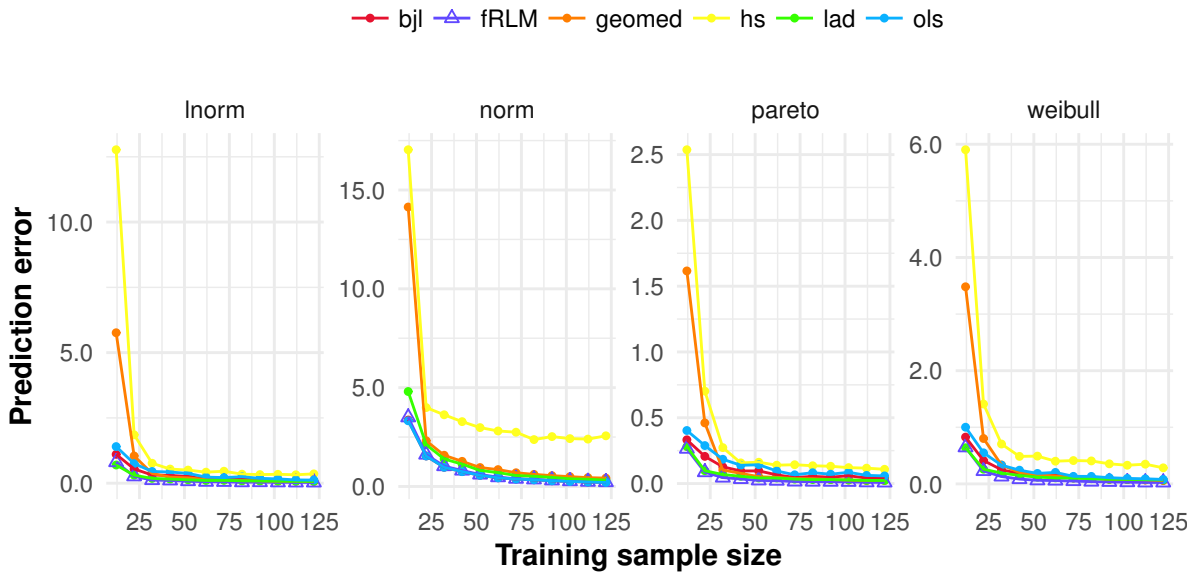

Figure 7: Prediction error as a function of sample size n, with $d = 5$, at noise level = 8.

Prediction error results are given in Fig. 8, at the middle noise level, for different model dimensions ranging over $5 \leq d \leq 140$. The sample size is determined such that $d/n = 1/6$ holds; this is a rather generous size, and thus where we observe deterioration in performance, we infer a lack of utility in more complex models, even when a sample of sufficient size is available. We see clearly that most procedures considered see a performance drop as model dimension grows, whereas our routine performs exactly the same, regardless of dimension size. This is a particularly appealing result illustrating the scalability of our `fRLM` in "bigger" tasks.

5.4 Test results: real-world data

We have seen extremely strong performance in the simulated situation; let us see how this extends to a number of real-world domains. The algorithms run are precisely the same as in the simulated cases, just the data is new. We have selected four data sets from a database of benchmark data sets for testing regression algorithms.[2] Our choices were such that the data come from a wide class of domains. For reference, the response variable in `bpres` is blood pressure, in `psych` is psychiatric assessment scores, in `rent` is cost to rent land, and in `oct` is petrol octane rating. All the data sets used here are included with a description in the online code repository referred at the start of this section. Depending on the data set, the dimensionality and sample size of the data sets naturally differ. Our protocol for evaluation is as follows. If the full data set is $\{\boldsymbol{z}_i\}_{i=1}^N$, then we take $n = \lceil 0.3N \rceil$ for training, and $m = N - n$ observations for testing. We carry out 100 trials, each time randomly choosing the train/test indices, and averaging over these trials to get prediction error.

Results are given in Fig. 9. While the data sets come from wildly varying domains (economics, manufacturing of petroleum products, human physiology and psychology), it is apparent that the results here very closely parallel those of our simulations, which again are the kind of performance that the theoretical exposition of sections 3–4 would have us expect. Strong performance is achieved with no *a priori* information, and with no fine-tuning whatsoever. Exactly the same routine is deployed in all problems. Of particular importance here is that we

[2]Compiled online by J. Burkardt at `http://people.sc.fsu.edu/~jburkardt/`.

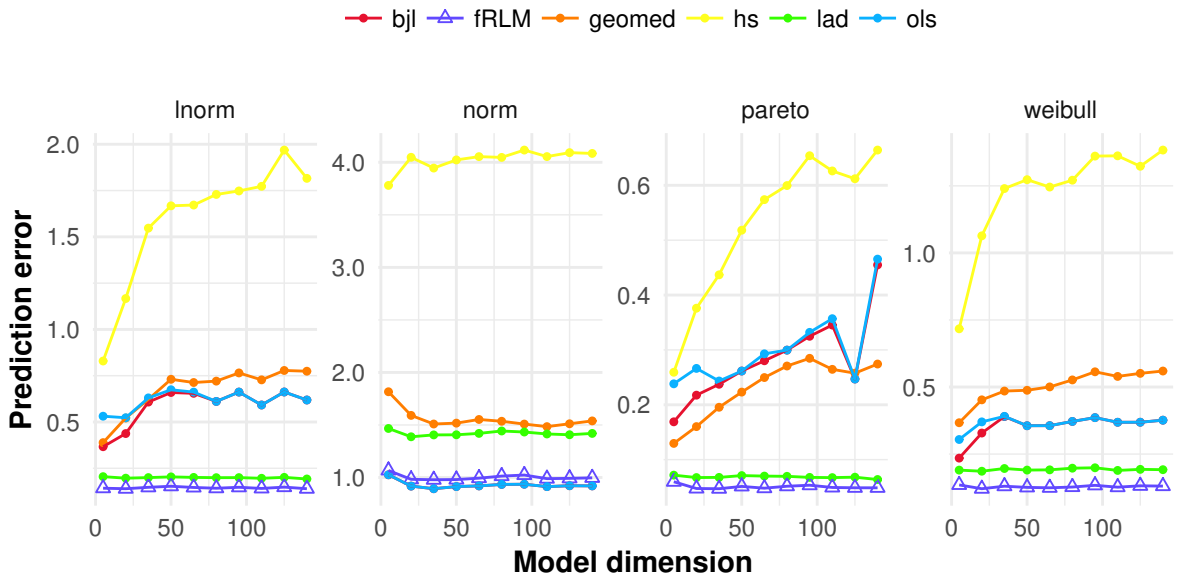

Figure 8: Prediction error as a function of model dimension d with fixed ratio $d/n = 1/6$, at noise level $= 8$.

are able to beat or match the `bjl` routine under all settings here as well; both of these routines attempt to minimize similar robust losses (defined implicitly), however our routine does it at a fraction of the cost, since we have no need to appeal to general-purpose non-linear optimizers, a very promising result.

6 Concluding remarks

In this work, we have introduced and explored a novel approach to the regression problem, using robust loss estimates and an efficient routine for minimizing these estimates without requiring prior knowledge of the underlying distribution. In addition to theoretical analysis of the fundamental properties of the algorithm being used, we showed through comprehensive empirical testing that the proposed technique indeed has extremely desirable robustness properties. In a wide variety of problem settings, our routine was shown to uniformly outperform well-known competitors both classical and modern, with cost requirements that are tolerable, suggesting a strong general approach for regression in the non-parametric setting.

Looking ahead, there are a number of interesting lines of work to be taken up. Extending this work to unsupervised learning problems is an immediate goal. Beyond this, a more careful look at the optimality of different algorithms from a cost/performance standpoint would assuredly be of interest. When is it more profitable (under some metric) to use "balanced" methods such as that of Minsker [33], Brownlees et al. [11], Hsu and Sabato [25] or ours, rather than committing to one of two extremes, say OLS or LAD? The former perform very well, but require extra computation. Characterizing such situations in terms of the underlying data distribution is both technically and conceptually interesting. Clear tradeoffs between formal assurances and extra computational cost could shed new light on precisely where traditional ERM algorithms and close variants fail to be economical.

Figure 9: Prediction error on four distinct real-world data sets.

References

[1] Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, volume 55 of *National Bureau of Standards Applied Mathematics Series*. National Bureau of Standards.

[2] Alon, N., Ben-David, S., Cesa-Bianchi, N., and Haussler, D. (1997). Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4):615–631.

[3] Ash, R. B. and Doléans-Dade, C. A. (2000). *Probability and Measure Theory*. Academic Press, 2nd edition.

[4] Audibert, J.-Y. and Catoni, O. (2011). Robust linear least squares regression. *Annals of Statistics*, 39(5):2766–2794.

[5] Bartlett, P. L., Long, P. M., and Williamson, R. C. (1996). Fat-shattering and the learnability of real-valued functions. *Journal of Computer and System Sciences*, 52(3):434–452.

[6] Bartlett, P. L. and Mendelson, S. (2006). Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334.

[7] Bartlett, P. L., Mendelson, S., and Neeman, J. (2012). ℓ_1-regularized linear regression: persistence and oracle inequalities. *Probability Theory and Related Fields*, 154(1-2):193–224.

[8] Breiman, L. (1968). *Probability*. Addison-Wesley.

[9] Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

[10] Brent, R. P. (1973). *Algorithms for minimization without derivatives*. Prentice-Hall.

[11] Brownlees, C., Joly, E., and Lugosi, G. (2015). Empirical risk minimization for heavy-tailed losses. *Annals of Statistics*, 43(6):2507–2536.

[12] Catoni, O. (2009). High confidence estimates of the mean of heavy-tailed real random variables. *arXiv preprint arXiv:0909.5366*.

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

[14] Cucker, F. and Smale, S. (2002). On the mathematical foundations of learning. *Bulletin (New Series) of the American Mathematical Society*, 39(1):1–49.

[15] Dellacherie, C. and Meyer, P.-A. (1978). *Probabilities and Potential*, volume 29 of *North-Holland Mathematics Studies*. North-Holland.

[16] Devroye, L., Lerasle, M., Lugosi, G., and Oliveira, R. I. (2015). Sub-Gaussian mean estimators. *arXiv preprint arXiv:1509.05845*.

[17] Dudley, R. M. (1978). Central limit theorems for empirical measures. *Annals of Probability*, 6(6):899–929.

[18] Dudley, R. M. (2014). *Uniform Central Limit Theorems*. Cambridge University Press, 2nd edition.

[19] Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

[20] Geman, D. and Reynolds, G. (1992). Constrained restoration and the recovery of discontinuities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(3):367–383.

[21] Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741.

[22] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons.

[23] Hsu, D., Kakade, S. M., and Zhang, T. (2014). Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14(3):569–600.

[24] Hsu, D. and Sabato, S. (2014). Heavy-tailed regression with a generalized median-of-means. In *Proceedings of the 31st International Conference on Machine Learning (ICML2014)*, pages 37–45.

[25] Hsu, D. and Sabato, S. (2016). Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, 17(18):1–40.

[26] Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1):73–101.

[27] Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons, 1st edition.

[28] Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*. John Wiley & Sons, 2nd edition.

[29] Kearns, M. J. and Schapire, R. E. (1994). Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48:464–497.

[30] Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

[31] Lerasle, M. and Oliveira, R. I. (2011). Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*.

[32] Lugosi, G. and Mendelson, S. (2016). Risk minimization by median-of-means tournaments. *arXiv preprint arXiv:1608.00757*.

[33] Minsker, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4):2308–2335.

[34] Pollard, D. (1981). Limit theorems for empirical processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57(2):181–195.

[35] Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer-Verlag.

[36] R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[37] Rousseeuw, P. and Yohai, V. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis*, volume 26 of *Lecture Notes in Statistics*, pages 256–272. Springer.

[38] Salibian-Barrera, M. and Yohai, V. J. (2006). A fast algorithm for S-regression estimates. *Journal of Computational and Graphical Statistics*, 15(2):1–14.

[39] Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. (2010). Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11:2635–2670.

[40] Srebro, N., Sridharan, K., and Tewari, A. (2010). Smoothness, low noise and fast rates. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 2199–2207.

[41] Steele, J. M. (1975). *Combinatorial entropy and uniform limit laws*. PhD thesis, Stanford University.

[42] Takeuchi, I., Le, Q. V., Sears, T. D., and Smola, A. J. (2006). Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7:1231–1264.

[43] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, pages 267–288.

[44] Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280.

[45] Vardi, Y. and Zhang, C.-H. (2000). The multivariate L_1-median and associated data depth. *Proceedings of the National Academy of Sciences*, 97(4):1423–1426.

[46] Yu, Y., Aslan, Ö., and Schuurmans, D. (2012). A polynomial-time form of robust regression. In *Advances in Neural Information Processing Systems 25*, pages 2483–2491.

A Helper functions for M-estimation

Here we define the ρ and χ functions referred to throughout the text. Starting with the ρ functions, we have referred to

$$\rho(u) = 2\left(\sqrt{1+u^2/2}-1\right) \quad (\texttt{al})$$

$$\psi(u) = 2\arctan\left(\exp\left(u\right)\right) - \pi/2 \quad (\texttt{gud})$$

$$\rho(u) = \begin{cases} c|u| + c^2(1-\pi/2) & |u| > c\pi/2 \\ c^2(1-\cos(u/c)) & |u| \le c\pi/2 \end{cases} \quad (\texttt{hmd})$$

$$\rho(u) = \begin{cases} c^2\left(\frac{|u|}{c} - \frac{1}{2}\right) & |u| > c \\ u^2/2 & |u| \le c \end{cases} \quad (\texttt{hub})$$

$$\rho(u) = \log(\cosh(u)) \quad (\texttt{lch})$$

where the text in parentheses (e.g., `al`) refers to the short-form used in Figure 5. As discussed in Example 3, the Gudermannian (`gud`) ρ function is defined implicitly by the ψ given here.

Next are χ functions for scaling:

$$\chi(u) = \frac{|u|^p}{1+|u|^p} - \beta \quad (\texttt{ga}, \texttt{gq})$$
$$\chi(u) = \log(1+\psi(u)^2) - \beta \quad (\texttt{lq})$$
$$\chi(u) = \psi(u)^2 - \beta \quad (\texttt{pq})$$
$$\chi(u) = \begin{cases} \frac{c^2}{6} - \beta & |u| \geq c \\ \frac{x^6}{6c^4} - \frac{x^4}{2c^2} + \frac{x^2}{2} - \beta & |u| < c \end{cases} \quad (\texttt{tuk})$$

Here `ga` and `gq` refer to settings $p = 1$ and $p = 2$ respectively, and the $\psi = \rho'$ here is for any choice of ρ according to Defn. 6. For these two χ in our experiments, we have used `lch` for the ρ that specifying them.

B Proofs of results in the main text

Proof of Lemma 8. For notational simplicity, given any $h \in \mathcal{H}$, write $x_i = l(h; \boldsymbol{z}_i)$, $i \in [n]$. Taking $u \in [\min_i\{x_i\}, \max_i\{x_i\}]$, clearly the right-hand side of (7) is non-empty, i.e., an M-estimate exists. Since ρ is differentiable and strongly convex on $\mathbb{R}$, the minimum is uniquely determined, characterized by the $\mathbf{E}_{\mu_n}\psi$ condition in the Lemma statement, noting ψ is monotone increasing on its domain, we have that $\widehat{\theta}(h)$ is well-defined.

Regarding $\theta^*(h)$, writing $x = l(h; \boldsymbol{z})$, since $|\rho(u)| \leq c|u|$ for some $c > 0$, integrability follows by monotonicity of the Lebesgue integral, that is for any $u \in \mathbb{R}$, we have by $x \in \mathcal{L}_2(\mu)$ that

$$\int \rho\left(\frac{x-u}{s}\right) d\mu \leq \int \frac{c|x-u|}{s}\, d\mu < \infty.$$

Since $\rho' = \psi$ is bounded, again for any u we have that

$$\frac{d}{du}\,\mathbf{E}_\mu\,\rho\left(\frac{x-u}{s}\right) = \frac{-1}{s}\,\mathbf{E}_\mu\,\psi\left(\frac{x-u}{s}\right)$$

holds [3, Ch. 1.6]. Existence of the minimum, given as a root of the right-hand side of this equation, is now immediate. Uniqueness follows from the strong convexity of ρ, noting for any functions u and v of $\boldsymbol{z}$,

$$\mathbf{E}_\mu\,\rho(\alpha u(\boldsymbol{z}) + (1-\alpha)v(\boldsymbol{z})) < \alpha\,\mathbf{E}_\mu\,\rho(u(\boldsymbol{z})) + (1-\alpha)\,\mathbf{E}_\mu\,\rho(v(\boldsymbol{z}))$$

for any $\alpha \in (0,1)$. □

Proof of Lemma 9. Fix arbitrary values $l_1, \ldots, l_n \in \mathbb{R}_+$ and $s_1, \ldots, s_n > 0$. To compactly denote these variables, write $\boldsymbol{a} = (l_1, \ldots, l_n, s_1, \ldots, s_n)$. Denote $\mathcal{B}_0 := \mathcal{B}(\mathbb{R}^{2n})$ here, and define

$$F(u, \boldsymbol{a}) := \sum_{i=1}^n \rho\left(\frac{l_i - u}{s_i}\right), \quad f(u, \boldsymbol{a}) := \frac{d}{dt}F(t, \boldsymbol{a})|_{t=u}, \quad u \in \mathbb{R}.$$

Let $\widehat{u} := \inf \arg\min_u F(u, \boldsymbol{a})$, a map from $\mathbb{R}^{2n}$ to $\mathbb{R}$. If ρ specifies a robust penalty, then from Lemma 8 the minimizer is unique and thus the infimum is superfluous. More generally, even when the minimizer is not unique, the infimum $\widehat{u}$ will be a valid minimizer. To see this, denoting $\rho_0 := \min_u F(u, \boldsymbol{a})$, say we have $F(\widehat{u}, \boldsymbol{a}) > \rho_0$. By continuity and monotonicity, there exists $u_1 > \widehat{u}$ such that $\rho_0 < F(u_1, \boldsymbol{a}) < F(\widehat{u}, \boldsymbol{a})$, and thus u_1 lower bounds the set

$\arg\min_u F(u, \boldsymbol{a})$, a contradiction of $\widehat{\theta}(h)$ being the greatest lower bound. Thus $F(\widehat{u}, \boldsymbol{a}) = \rho_0$. It follows that $\widehat{u}$ is also root of $f(\cdot, \boldsymbol{a})$.

For arbitrary $\alpha \in \mathbb{R}$, define events

$$\begin{aligned}\boldsymbol{A}_\alpha &\coloneqq \{\boldsymbol{a} \in \mathbb{R}^{2n} : \widehat{u} \leq \alpha\} \\ \boldsymbol{A}' &\coloneqq \bigcap_{k=1}^{\infty} \bigcup_{u \in U_\alpha} \left\{\boldsymbol{a} \in \mathbb{R}^{2n} : |f(u, \boldsymbol{a})| < \frac{1}{k}\right\}, \quad U_\alpha \coloneqq \{q \in \mathbb{Q} : q \leq \alpha\}.\end{aligned}$$

Indexing over the rationals is to make the union countable. First note that as $f(u, \cdot)$ is continuous, it is measurable for every u, and equivalently

$$\{|f(u, \boldsymbol{a})| < 1/k\} \in \mathcal{B}_0, \quad \forall\, u \in \mathbb{R}, k \in \mathbb{N}.$$

As such every set indexed in $\boldsymbol{A}'$ is measurable. As $\boldsymbol{A}'$ is a countable intersection of a countable union of measurable sets, $\boldsymbol{A}'$ itself is measurable. First, say $\boldsymbol{a} \in \boldsymbol{A}'$. On this occasion, for each integer $k > 0$, there exists a rational $u \leq \alpha$ such that the objective $f(\cdot, \boldsymbol{a})$ falls within $\pm\, k^{-1}$ of zero. Now assume $\widehat{u}(\boldsymbol{a}) > \alpha$ for this $\boldsymbol{a}$. By definition $f(\widehat{u}(\boldsymbol{a}), \boldsymbol{a}) = 0$. As f depends monotonically on u, and $\widehat{u}$ is infimal, we have for some $\epsilon > 0$ that

$$\exists\, u_1 \in (\alpha, \widehat{u}(\boldsymbol{a})) \cap \mathbb{Q}, \quad f(u_1, \boldsymbol{a}) \geq \epsilon.$$

Taking $k \in \mathbb{N}$ large enough (so that $1/k < \epsilon$), we can necessarily secure a rational $q \leq \alpha$ such that $|f(q, \boldsymbol{a})| < \epsilon$. However as $q < u_1$, this means that

$$f(q, \boldsymbol{a}) \geq f(u_1, \boldsymbol{a}) \geq \epsilon > 0,$$

which is a contradiction. Thus $\widehat{u}(\boldsymbol{a}) \leq \alpha$. The $\boldsymbol{a}$ choice was arbitrary, so $\boldsymbol{A}' \subseteq \boldsymbol{A}_\alpha$.

The converse is even simpler. Let $\boldsymbol{a} \in \boldsymbol{A}_\alpha$. We can always take a sequence (q_m) of $q_m \in \mathbb{Q}$ where $q_m \uparrow \widehat{u}(\boldsymbol{a})$. For any $k \in \mathbb{N}$, there exists $m_0 < \infty$ where

$$m \geq m_0 \implies f(q_m, w, \boldsymbol{z}) - f(\widehat{u}(\boldsymbol{a}), \boldsymbol{a}) < 1/k$$

which in turn implies $|f(q_m, \boldsymbol{a})| < 1/k$, that is $\boldsymbol{a} \in \boldsymbol{A}'$. We have $\boldsymbol{A}_\alpha \subseteq \boldsymbol{A}'$ and thus $\boldsymbol{A}_\alpha = \boldsymbol{A}'$, concluding that $\boldsymbol{A}_\alpha \in \mathcal{B}_0$ for any choice of α, and any $w \in \mathcal{W}$. Note $\boldsymbol{A}_\alpha$ is just $\widehat{u}^{-1}(-\infty, \alpha]$, the inverse image of this segment induced by $\widehat{u}$. Denoting these intervals as $\mathcal{D} = \{(-\infty, \alpha] : \alpha \in \mathbb{R}\}$, the σ-field generated by this class is $\sigma(\mathcal{D}) = \mathcal{B}(\mathbb{R})$, and the class $\mathcal{D}' = \{B \in \mathcal{B} : \widehat{u}^{-1}(B) \in \mathcal{B}_0\}$ is a σ-field [8, Ch. 2.7]. We proved above that $\mathcal{D} \subseteq \mathcal{D}'$, and by minimality of the generated field, $\mathcal{D}' = \mathcal{B}^1$. We conclude $\widehat{u}^{-1}(B) \in \mathcal{B}_0$ for all $B \in \mathcal{B}(\mathbb{R})$. With this, and the measurability of $l(\cdot; \cdot)$ and s_h, the Lemma follows; the specific requirement is $\mathcal{B}(\mathcal{H}) \times \mathcal{B}_{d+1}$ measurability of l and either $\mathcal{B}(\mathcal{H}) \times \mathcal{B}_{d+1}^n$ or $\mathcal{B}(\mathcal{H}) \times \mathcal{B}_{d+1}$ measurability of s_h, depending on whether it is determined by μ_n or individual instances. □

Proof of Theorem 10. Use $\widehat{\theta}(h)$ as in the statement of Lemma 9. Fix an arbitrary set of instances $\boldsymbol{Z} \coloneqq (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) \in \mathcal{Z}$, and

$$\begin{aligned}\widehat{\theta}(\mathcal{H}) &\coloneqq \inf\left\{\widehat{\theta}(h) : h \in \mathcal{H}\right\} \\ f(u, h; \boldsymbol{Z}) &\coloneqq \sum_{i=1}^{n} \psi\left(\frac{l(h; \boldsymbol{z}_i) - u}{s_h(\boldsymbol{z}_i)}\right), \quad h \in \mathcal{H}, u \in \mathbb{R}.\end{aligned}$$

Construct a sequence (θ_m) of $\theta_m \in \{\widehat{\theta}(h) : h \in \mathcal{H}\}$ such that $\theta_m \downarrow \widehat{\theta}(\mathcal{H})$. To each θ_m, there is an accompanying $h_m \in \mathcal{H}$ such that $f(\theta_m, h_m, \boldsymbol{Z}) = 0$. As $\sup_m \|h_m\| < \infty$, there exists

a convergent subsequence (h_k). Denote $\widehat{h} \coloneqq \lim_{k\to\infty} h_k$. Subsequence θ_k converges to $\widehat{\theta}(\mathcal{H})$. Continuity of L and s implies $f(\cdot,\cdot,\boldsymbol{Z})$ is continuous, and thus

$$f(\widehat{\theta}(\mathcal{H}),\widehat{h},\boldsymbol{Z}) = \lim_{k\to\infty} f(\theta_k,h_k,\boldsymbol{Z}) = 0,$$

which by uniqueness of the root of $f(\cdot,\widehat{h},\boldsymbol{Z})$ (Lemma 8) implies that

$$\forall\, \boldsymbol{Z}, \exists\, \widehat{h} \in \mathcal{H}, \widehat{\theta}(\widehat{h}) = \widehat{\theta}(\mathcal{H}). \tag{16}$$

That is, for any set of observations $\boldsymbol{Z}$, we can find such an $\widehat{h}$ minimizing the new objective function.

From this point, measurability is a purely technical endeavour. Useful references are Dudley [18, Ch. 5], Pollard [35, Appendix C], and Dellacherie and Meyer [15, Ch. 1–3]. We assume $\mathcal{H}$ ia separable; the special case of $\mathcal{H} \subset \mathbb{R}^d$ is an archetypal example. Index and assemble all possible (random) values of our objective in $\boldsymbol{\Theta} \coloneqq \{\widehat{\theta}(h) : h \in \mathcal{H}\}$, with $\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n$ left free to vary randomly. As $\widehat{\theta}(h)$ has been shown to be $\mathcal{H}\times\mathcal{Z}$-measurable (Lemma 9), under an innocuous regularity condition, [35, Appendix C, 1(ii)], the class $\boldsymbol{\Theta}$ is sufficiently regular, called "permissible." It is readily verified that $\widehat{\theta}(\mathcal{H})$ is $\mathcal{B}(\mathcal{Z})$-measurable. Next, define the set

$$\begin{aligned}\boldsymbol{A}_3 &\coloneqq \{(\boldsymbol{Z},h) : \widehat{\theta}(h) = \widehat{\theta}(\mathcal{H})\}\\ &= \widetilde{\theta}^{-1}(-\infty,0] \cap \widetilde{\theta}^{-1}(-\infty,0)^c\end{aligned}$$

where we have written $\widetilde{\theta}(\boldsymbol{Z},h) \coloneqq (\widehat{\theta}(h) - \widehat{\theta}(\mathcal{H}))$. We have already verified the measurability of the two terms being subtracted, thus $\widetilde{\theta}$ is $\mathcal{B}(\mathcal{H})\times\mathcal{B}(\mathcal{Z})$ measurable. Looking at the second equality, we have that $\boldsymbol{A}_3$ is an analytic subset of $\mathcal{Z}\times\mathcal{H}$. Taking the projection π of $\boldsymbol{A}_3$ onto the observation space, namely

$$\pi(\boldsymbol{A}_3) \coloneqq \{\boldsymbol{Z} : (\boldsymbol{Z},h) \in \boldsymbol{A}_3, h \in \mathcal{H}\},$$

and note that by our existence result (16), $\mathbf{P}\,\pi(\boldsymbol{A}_3) = 1$. From Pollard [35, Appendix C(d)], it follows that there exists a random variable $\widehat{h}(\boldsymbol{Z})$ such that $(\boldsymbol{Z},\widehat{h}(\boldsymbol{Z})) \in \boldsymbol{A}_3$ for almost all $\boldsymbol{Z} \in \pi(\boldsymbol{A}_3)$. Since the latter set has $\mathbf{P}$-measure 1, we conclude that this $\widehat{h}$ realizes the properties sought in the statement of Theorem 10, concluding the argument. □

Proof of Proposition 12. Consider any sample $\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n$. Write $\gamma(h) = \gamma_{\mu_n}(h)$ for simplicity. Fix any $\varepsilon > 0$. By continuity of L, exists $\delta > 0$ where $\|h - h'\| \le \delta$ implies

$$\max\left\{|l(h;\boldsymbol{z}_i) - \gamma(h) - l(h';\boldsymbol{z}_i) + \gamma(h')|\right\}_{i=1}^n \le \varepsilon.$$

Denote $s \coloneqq s_h$ and $s' \coloneqq s_{h'}$. Now assume $|s - s'| > \varepsilon$, say for concreteness that $s + \varepsilon < \widetilde{s} < s'$. This implies that for any $\widetilde{s}$ taken such that $s + \varepsilon < \widetilde{s} < s'$, we have

$$\frac{l(h';\boldsymbol{z}_i) - \gamma(h')}{s'} < \frac{l(h';\boldsymbol{z}_i) - \gamma(h')}{\widetilde{s}} < \frac{l(h;\boldsymbol{z}_i) - \gamma(h)}{s}, \quad i = 1,\ldots,n$$

and by the weak monotonicity of χ, and the definitions of the two roots s and s',

$$\begin{aligned}0 = \sum_{i=1}^n \chi\left(\frac{l(h';\boldsymbol{z}_i) - \gamma(h')}{s'}\right) &\le \sum_{i=1}^n \chi\left(\frac{l(h';\boldsymbol{z}_i) - \gamma(h')}{\widetilde{s}}\right)\\ &\le \sum_{i=1}^n \chi\left(\frac{l(h;\boldsymbol{z}_i) - \gamma(h)}{s}\right)\\ &= 0,\end{aligned}$$

and thus the middle sum is in fact zero. This implies

$$\widetilde{s} \in \left\{ s > 0 : \sum_{i=1}^{n} \chi((l(h'; \boldsymbol{z}_i) - \gamma(h'))/s) = 0 \right\},$$

but since $\widetilde{s} < s'$, this is a contradiction of s' as the infimum of this set. An identical argument holds for the other case of $s' + \varepsilon < s$, and so $|s - s'| \leq \varepsilon$. We conclude for any $\varepsilon > 0$, there exists $\delta > 0$ such that $\|h - h'\| \leq \delta$ implies $|s_h - s_{h'}| \leq \varepsilon$. □

Proof of Theorem 13. For $h \in \mathcal{H}$, write $x = l(h; \boldsymbol{z})$ and $\widehat{\theta} = \widehat{\theta}(h)$, $\theta^* = \theta^*(h)$ for simplicity. Let s be either a fixed positive constant, or be generated on a per-observation basis, i.e., $s_1, \ldots, s_n$ are independent positive random variables, where say $s = s(\boldsymbol{z})$ for $\boldsymbol{z} \sim \mu$. The existence of $\widehat{\theta}$ and θ^* is given by Lemma 8. For convenience denote $\psi_u \coloneqq \psi((x-u)/s)$ and note that

$$\{\widehat{\theta} < u\} = \{\mathbf{E}_{\mu_n} \psi_u < 0\}, \quad \{\widehat{\theta} > u\} = \{\mathbf{E}_{\mu_n} \psi_u > 0\} \tag{17}$$

for any choice of u. Use the typical set lim inf definition, which is to say for any given sequence of sets A_m, let $\liminf_m A_m \coloneqq \bigcup_{m=1}^{\infty} \bigcap_{k \geq m} A_k$. For arbitrary fixed $\varepsilon > 0$, we have

$$\begin{aligned}
\mathbf{P}\left\{\lim_n \widehat{\theta} < \theta^* + \varepsilon\right\} &= \mathbf{P} \liminf_n \{\widehat{\theta}_n < \theta^* + \varepsilon\} \\
&= \mathbf{P} \liminf_n \{\mathbf{E}_{\mu_n} \psi_{\theta^*+\varepsilon} < 0\} \\
&= \mathbf{P}\left\{\lim_n \mathbf{E}_{\mu_n} \psi_{\theta^*+\varepsilon} < 0\right\} \\
&\geq \mathbf{P}\left\{\lim_n \mathbf{E}_{\mu_n} \psi_{\theta^*+\varepsilon} = \mathbf{E}_{\mu} \psi_{\theta^*+\varepsilon}\right\} \\
&= 1.
\end{aligned}$$

The final equality holds via the strong law of large numbers, which is where we require $\mathbf{E}_{\mu} x^2 < \infty$ [8, Theorem 3.27]. The inequality prior to that holds since $\mathbf{E}_{\mu} \psi_{\theta^*+\varepsilon} < 0$, and the remaining equalities by lim inf definition and (17). An indentical argument can be used to show $\mathbf{P}\{\lim_n \widehat{\theta} > \theta^* - \varepsilon\} = 1$, which implies

$$\begin{aligned}
\mathbf{P}\left\{\lim_n |\widehat{\theta} - \theta^*| \geq \varepsilon\right\} &\leq \mathbf{P}\left\{\lim_n \widehat{\theta}_n \geq \theta^* + \varepsilon\right\} \cup \left\{\lim_n \widehat{\theta} \leq \theta^* - \varepsilon\right\} \\
&= 0.
\end{aligned}$$

This holds for any choice of $\varepsilon > 0$, and thus $|\widehat{\theta} - \theta^*| \to 0$ almost surely, yielding strong consistency. □

Proof of Lemma 14. If ρ specifies a robust objective, then ψ is a bounded measurable function, and can be uniformly approximated by a sequence of weighted indicators as follows. For concreteness, say $|\psi| \leq M < \infty$. Let sequence $\varepsilon_m \downarrow 0$, and for each $m \in \mathbb{N}$ partition the range $[-M, M]$ into $k_m \coloneqq 2M/\varepsilon_m$ segments $A_j \coloneqq \{t : a_{j-1} \leq \psi(t) < a_j\}$ defined by

$$a_0 = -M, \quad a_j = a_{j-1} + \varepsilon_m, \quad j = 1, \ldots, k_m.$$

The approximating function s_m is then defined as

$$s_m(u) \coloneqq \sum_{j=1}^{k_m} a_j I_{A_j}(u), \quad u \in \mathbb{R}.$$

By strong convexity, there is no $u \in \mathbb{R}$ where $|\psi(u)| = M$, and thus the uniform approximation is immediate. That is, $|s_m(u) - \psi(u)| \le \varepsilon_m$ holds uniformly in u. Note that each A_j can be given as an interval. Defining b_j to be the unique element in $\overline{\mathbb{R}}$ where $\psi(b_j) = a_j$, the marginal sets are $A_1 = (-\infty, b_1)$ and $A_{k_m} = [b_{k_m-1}, \infty)$ respectively, and the remainder are half-closed real intervals $A_j = [b_{j-1}, b_j)$.

Denote $\mathbf{P}_n = \mathbf{E}_{\mu_n}$ and $\mathbf{E} = \mathbf{E}_\mu$ for clean notation. Our interest is with the quantity

$$\|\mathbf{P}_n\,\psi - \mathbf{E}\,\psi\| := \sup_{u,h,s} \left| \mathbf{P}_n\,\psi\left(\frac{l(h;\boldsymbol{z}) - u}{s}\right) - \mathbf{E}\,\psi\left(\frac{l(h;\boldsymbol{z}) - u}{s}\right)\right|$$

where $s > 0$, $h \in \mathcal{H}$, and $u \in \mathbb{R}$ when taking the supremum. For any observation $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ an application of the triangle inequality yields

$$\|\mathbf{P}_n\,\psi - \mathbf{E}\,\psi\| \le \|\mathbf{P}_n\,\psi - \mathbf{P}_n\,s_m\| + \|\mathbf{P}_n\,s_m - \mathbf{E}\,s_m\| + \|\mathbf{E}\,s_m - \mathbf{E}\,\psi\| \tag{18}$$

where the $\|\cdot\|$ terms on the right-hand side denote taking the exact same suprema as on the left-hand side. The first and third terms are readily dealt with. Note for example that

$$\|\,\mathbf{E}(s_m - \psi)\| \le \|s_m - \psi\|_\infty \le \varepsilon_m \to 0$$

whenever we set index $m = m(n) \to \infty$ as $n \to \infty$. An identical argument holds for the first term. This convergence is deterministic, in the sense that it holds for arbitrary observations, and thus also holds almost surely.

The second term in (18) is slightly more involved but the approach is rather standard. To get started, denoting for convenience the events

$$E_j := \{l(h;\boldsymbol{z}) \in [sb_{j-1} + u, sb_{j-1} + u)\}, \quad j = 1, \ldots, k_m$$

with the understanding that for the index $j = 1$ the interval is $(-\infty, sb_1 + u)$ and $j = k_m$ it is $[sb_{k_m-1} + u, \infty)$. The obvious but important fact is that each event E_j, specified by s, h, u, and the b_j values, is naturally captured by a larger class of sets $\mathcal{C}$

$$\mathcal{C} := \left\{\{\boldsymbol{z} : l(h;\boldsymbol{z}) \in [a,b)\} : h \in \mathbb{R}^d, a, b \in \overline{\mathbb{R}}, a < b\right\}.$$

Note we are assuming $\mathcal{H}$ is specified by elements of d-dimensional Euclidean space. Since each $E_j \in \mathcal{C}$, we have that

$$\begin{aligned}\|\mathbf{P}_n\,s_m - \mathbf{E}\,s_m\| &= \sup\left|\sum_{j=1}^{k_m} a_j\left(\mathbf{P}_n\,I_{E_j}(\boldsymbol{z}) - \mathbf{P}\,E_j\right)\right| \\ &\le M k_m \|\,\mathbf{P}_n\,I_C - \mathbf{P}\,C\|_{\mathcal{C}}\end{aligned} \tag{19}$$

where $\|\cdot\|_{\mathcal{C}}$ denotes taking the supremum over $C \in \mathcal{C}$. We will frequently use I_C to denote $I_C(\cdot)$, with domain $\mathbb{R}^{d+1}$. It remains to show the strong convergence to zero of the supremal factor in (19), with convergence rates to deal with the increasing k_m sequence.

A typical symmetrization inequality is of use next [44, Lemma 2]. Take an artificial sample $\boldsymbol{z}'_1, \ldots, \boldsymbol{z}'_n$, independent from $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, but identically distributed. For any $\varepsilon > 0$, whenever $n > 2/\varepsilon^2$, we have

$$\mathbf{P}\left\{\|\,\mathbf{P}_n\,I_C - \mathbf{P}\,C\|_{\mathcal{C}} > \varepsilon\right\} \le 2\,\mathbf{P}\left\{\|\,\mathbf{P}_n\,I_C - \mathbf{P}'_n\,I_C\|_{\mathcal{C}} \ge \varepsilon/2\right\} \tag{20}$$

where $\mathbf{P}'_n$ analogously denotes μ'_n supported on the new sample. Next a randomization technique due to Pollard [34]. Let $\sigma_1, \ldots, \sigma_n$ be iid, and independent from both samples, with

distribution $\mathbf{P}\{\sigma = -1\} = \mathbf{P}\{\sigma = 1\} = 1/2$. Checking cases one immediately confirms that for any $C \in \mathcal{C}$, the random quantities $I_C(z) - I_C(z')$ and $\sigma(I_C(z) - I_C(z'))$ have the same distribution. As such for any $\varepsilon > 0$,

$$\begin{aligned}\mathbf{P}\left\{\|\,\mathbf{P}_n\, I_C - \mathbf{P}'_n\, I_C\|_{\mathcal{C}} \geq \varepsilon\right\} &= \mathbf{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\sigma_i(I_C(z_i) - I_C(z'_i))\right\|_{\mathcal{C}} \geq \varepsilon\right\} \\ &\leq \mathbf{P}\left\{\|\,\mathbf{P}_n\, \sigma I_C\|_{\mathcal{C}} + \|\,\mathbf{P}'_n\, \sigma I_C\|_{\mathcal{C}} \geq \varepsilon\right\} \\ &\leq 2\,\mathbf{P}\left\{\|\,\mathbf{P}_n\, \sigma I_C\|_{\mathcal{C}} \geq \varepsilon/2\right\}\end{aligned}$$

where for the first inequality one leverages the triangle inequality, and for the second a union bound. We can conclude up to this point for large enough n that

$$\mathbf{P}\left\{\|\,\mathbf{P}_n\, I_C - \mathbf{P}\, C\|_{\mathcal{C}} > \varepsilon\right\} \leq 4\,\mathbf{P}\left\{\|\,\mathbf{P}_n\, \sigma I_C\|_{\mathcal{C}} \geq \varepsilon/4\right\}.$$

Fixing arbitrary sample $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, a combinatorial indicator of the complexity $\mathcal{C}$ is given by

$$\begin{aligned}\Delta_n(\mathcal{C}) &\coloneqq |\{C \cap \{z_1, \ldots, z_n\} : C \in \mathcal{C}\}| \\ &= |\{(I_C(z_1), \ldots, I_C(z_n)) \in \{0,1\}^n : C \in \mathcal{C}\}|.\end{aligned}$$

Naturally the number of distinct subsets captured by members of $\mathcal{C}$ is identical to the number of distinct n-length binary-valued vectors than can be built on the sample when indexing over $\mathcal{C}$. Trivially $\Delta_n(\mathcal{C}) \leq 2^n$. Again conditioning on a fixed sample, we can always take $C_1, \ldots, C_k \in \mathcal{C}$ such that all possible realizations of $\mathbf{P}_n\, \sigma I_C$ are captured by indexing over these $k = \Delta_n(\mathcal{C})$ sets. That is, denoting $\boldsymbol{Z} \coloneqq (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$,

$$\begin{aligned}\mathbf{P}\left\{\|\,\mathbf{P}_n\, \sigma I_C\|_{\mathcal{C}} \geq \varepsilon; \boldsymbol{Z}\right\} &= \mathbf{P}\left\{\max_{1\leq j\leq k}\left|\mathbf{P}_n\, \sigma I_{C_j}\right| \geq \varepsilon; \boldsymbol{Z}\right\} \\ &\leq \mathbf{P}\bigcup_{j=1}^{k}\left\{\left|\mathbf{P}_n\, \sigma I_{C_j}\right| \geq \varepsilon; \boldsymbol{Z}\right\} \\ &\leq \Delta_n(\mathcal{C})\max_{1\leq j\leq k}\mathbf{P}\left\{\left|\mathbf{P}_n\, \sigma I_{C_j}\right| \geq \varepsilon; \boldsymbol{Z}\right\}.\end{aligned}$$

The two multiplicands need to be controlled. Let us start with the former. When we do not fix $\boldsymbol{Z}$, naturally $\Delta_n(\mathcal{C})$ is a random quantity. Note that the possible forms any $C \in \mathcal{C}$ can take are characterized into three types as

$$\{\boldsymbol{z} : l \in [a, b)\}, \quad \{\boldsymbol{z} : l \in [a, \infty)\}, \quad \{\boldsymbol{z} : l \in (-\infty, b)\},$$

also $\{l \in (-\infty, \infty)\} = \mathbb{R}^{d+1}$, and setting $b \leq 0$ returns the empty set since $l \geq 0$. We have denoted $l(h; \boldsymbol{z})$ by l for simplicity. For concreteness consider $l(h; \boldsymbol{z}) = (y - h(\boldsymbol{x}))^2$ case, though the exact same argument clearly holds for other related losses. Take any $a, b \in \mathbb{R}$ where $a < b$. Then setting

$$\begin{aligned}G_1 &\coloneqq \left\{y - \boldsymbol{w}^T\boldsymbol{x} \geq \sqrt{|a|}\right\}, \quad G_2 \coloneqq \left\{y - \boldsymbol{w}^T\boldsymbol{x} \leq -\sqrt{|a|}\right\} \\ G'_1 &\coloneqq \left\{y - \boldsymbol{w}^T\boldsymbol{x} < \sqrt{|b|}\right\}, \quad G'_2 \coloneqq \left\{y - \boldsymbol{w}^T\boldsymbol{x} > -\sqrt{|b|}\right\}\end{aligned}$$

and recalling under the linear model assumption on $\mathcal{H}$, for any $h \in \mathcal{H}$ we have $h(\boldsymbol{x}) = w^T\boldsymbol{x}$ for some $\boldsymbol{w} \in \mathbb{R}^d$, thus clearly we have

$$\{l \in [a, b)\} = (G_1 \cup G_2) \cap G'_1 \cap G'_2.$$

If one defines $g(z) \coloneqq (y - \boldsymbol{w}^T\boldsymbol{x} - \sqrt{|a|})(-1)$, then $G_1 = \{g(\boldsymbol{z}) \leq 0\}$. Setting $g'(\boldsymbol{z}) \coloneqq (y - \boldsymbol{w}^T\boldsymbol{x} - \sqrt{|b|})(-1)$, have $G_1' = \{g'(\boldsymbol{z}) \leq 0\}^c$ where the superscript denotes the complement. If our observations are $d+1$ dimension vectors of the form $\boldsymbol{z} = (x_1, \ldots, x_d, y)$, define functions $f_0(\boldsymbol{z}) \coloneqq 1$ and $f_j(\boldsymbol{z}) \coloneqq \pi_j(\boldsymbol{z})$ for $j = 1, \ldots, d+1$, where π_j denotes the jth coordinate projection. That is, e.g., $f_1(\boldsymbol{z}) = x_1$ and so forth. Construct a linear space of functions on $\mathbb{R}^{d+1}$ as

$$\mathcal{F} \coloneqq \operatorname{span}\{f_0, \ldots, f_{d+1}\}.$$

One may check the linear independence of these functions, and thus the dimension of is precisely $\dim \mathcal{F} = d+2$. Note clearly that $g, g' \in \mathcal{F}$. From this one naturally induces two classes of sets, namely

$$\mathcal{G} \coloneqq \{\{f(\boldsymbol{z}) \leq 0\} : f \in \mathcal{F}\}, \quad \mathcal{G}^c \coloneqq \{G^c : G \in \mathcal{G}\}.$$

A classic result [41, 17] says that, using more modern parlance the class $\mathcal{G}$ has a VC dimension bounded by $\dim \mathcal{F}$. The fundamental property of classes with finite VC dimension is that the supremum of Δ_n taken over all samples is bounded by a polynomial in n. More precisely, for some constant c_0, for all n we have

$$\mathbf{E}\, \Delta_n(\mathcal{G}) \leq s_n(\mathcal{G}) \coloneqq \sup_{\boldsymbol{Z}} \Delta_n(\mathcal{G}) \leq c_0 n^{d+2},$$

where the expectation is being taken over the sample $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$. It is then clear that

$$\{\boldsymbol{z} : l \in [a, b)\} \in (\mathcal{G} \cup \mathcal{G}) \cap \mathcal{G}^c \cap \mathcal{G}^c.$$

For all the other forms the sets $C \in \mathcal{C}$ take, it is clear that each is composed of sets from $\mathcal{G}$, $\mathcal{G}^c$, or $\{\mathbb{R}^{d+1}, \emptyset\}$. The zero function $g_0(\boldsymbol{z}) = 0$, is $g_0 \in \mathcal{F}$, and as such $\mathbb{R}^{d+1} = \{g_0(\boldsymbol{z}) \leq 0\} \in \mathcal{G}$. Also the basis function f_0 used in defining $\mathcal{F}$ is such that $\emptyset = \{f_0(\boldsymbol{z}) \leq 0\} \in \mathcal{G}$. It thus follows that $\emptyset, \mathbb{R}^{d+1} \in \mathcal{G}^c$ as well. We thus conclude

$$\mathcal{C} \subseteq \mathcal{G}^* \coloneqq (\mathcal{G} \cup \mathcal{G}) \cap \mathcal{G}^c \cap \mathcal{G}^c.$$

Basic combinatorial arguments [35, Lemma 15] show that for a constant $c_1 > 0$ we have

$$s_n(\mathcal{G}^*) \leq s_n(\mathcal{G})^2 s_n(\mathcal{G}^c)^2 \leq c_1 n^{4d+8}$$

which implies $\mathbf{E}\, \Delta_n(\mathcal{C}) \leq c_1 n^{4d+8}$. This is the desired bound for the combinatorial parameter. As for the conditional probability term, note that with fixed $\boldsymbol{Z}$ and the σ_i left random, taking expectation with respect to σ we have for any $C \in \mathcal{C}$ that

$$\mathbf{E}\, \mathbf{P}_n\, \sigma I_C(\boldsymbol{z}) = (\mathbf{E}\, \sigma)\, \mathbf{P}_n\, I_C(\boldsymbol{z}) = 0,$$

and so $\mathbf{P}_n\, \sigma I_C(\boldsymbol{z})$ is a zero-mean sum of random variables taking values on $[-1/n, 1/n]$. Direct application of Hoeffding's inequality yields, with an application of the union bound to get two-sided inequalities,

$$\mathbf{P}\left\{|\mathbf{P}_n\, \sigma I_C| \geq \varepsilon | \boldsymbol{z}_{(n)}\right\} \leq 2 \exp\left(\frac{-n\varepsilon^2}{2}\right)$$

for all n. Since the exact same bound holds regardless of $\boldsymbol{z}_{(n)}$ and choice of C, we connect things by integrating, noting for large enough n and constant $c_2 > 0$ we have

$$\begin{aligned}\mathbf{P}\left\{\|\,\mathbf{P}_n\, I_C - \mathbf{P}\, C\|_{\mathcal{C}} > \varepsilon\right\} &\leq 4\,\mathbf{P}\left\{\|\,\mathbf{P}_n\, \sigma I_C\|_{\mathcal{C}} \geq \varepsilon/4\right\} \\ &= 4\,\mathbf{E}\left(\Delta_n(\mathcal{C}) \max_{1\leq j\leq k} \mathbf{P}\left\{\left|\mathbf{P}_n\, \sigma I_{C_j}\right| \geq \varepsilon; \boldsymbol{Z}\right\}\right) \\ &\leq c_2 n^{4d+8} \exp\left(\frac{-n\varepsilon^2}{32}\right).\end{aligned}$$

Application of the root test immediately shows that summing the right-hand side of the final inequality over n, the series converges and thus

$$\sum_{n=1}^{\infty} \mathbf{P}\left\{\|\,\mathbf{P}_n\, I_C - \mathbf{P}\, C\|_{\mathcal{C}} > \varepsilon\right\} < \infty.$$

The Borel-Cantelli lemma then says that for any $\varepsilon > 0$,

$$\mathbf{P} \limsup_n \left\{\|\,\mathbf{P}_n\, I_C - \mathbf{P}\, C\|_{\mathcal{C}} > \varepsilon\right\} = 0$$

and since

$$\left\{\lim_{n\to\infty} \|\,\mathbf{P}_n\, I_C - \mathbf{P}\, C\|_{\mathcal{C}} = 0\right\}^c = \bigcup_{k=1}^{\infty} \limsup_n \left\{\|\,\mathbf{P}_n\, I_C - \mathbf{P}\, C\|_{\mathcal{C}} > 1/k\right\},$$

using a union bound we have

$$\mathbf{P}\left\{\lim_{n\to\infty} \|\,\mathbf{P}_n\, I_C - \mathbf{P}\, C\|_{\mathcal{C}} = 0\right\} \geq 1 - \sum_{k=1}^{k} \mathbf{P} \limsup_n \left\{\|\,\mathbf{P}_n\, I_C - \mathbf{P}\, C\|_{\mathcal{C}} > 1/k\right\} = 1$$

which means $\|\,\mathbf{P}_n\, I_C - \mathbf{P}\, C\|_{\mathcal{C}} \to 0$ almost surely.

Returning to sequence k_m from (19), while we are free to make this grow as slow as we like, a convergence rate for the term converging to zero makes the argument more transparent. This is done applying Theorems 37 and the Approximation Lemma of Pollard [35, Ch. 2], using the fact that $\mathcal{C}$ has polynomial discrimination, which is precisely what was proved above. In particular, setting $k_{m(n)} = O(n^{1/3})$ is sufficient to imply $\|\,\mathbf{P}_n\, I_C - \mathbf{P}\, C\|_{\mathcal{C}} = o(k_{m(n)}^{-1})$ almost surely. Thus via (19) we have that $\|\mathbf{P}_n\, s_m - \mathbf{E}\, s_m\| \to 0$ almost surely, implying the desired result via (18). □

Proof of Theorem 16. We start by controlling the random sequence $\widehat{\theta}(h)$ from above. Fix any $h \in \mathcal{H}$. By the usual strong law of large numbers, for any fixed $\delta > 0$, the event

$$\boldsymbol{A} := \left\{\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \psi\left(\frac{l(h;\boldsymbol{z}_i) - (\theta^*(h) + \delta)}{s_h(\boldsymbol{z}_i)}\right) = \mathbf{E}_\mu\, \psi\left(\frac{l(h;\boldsymbol{z}) - (\theta^*(h) + \delta)}{s_h(\boldsymbol{z})}\right)\right\}$$

has $\mathbf{P}(\boldsymbol{A}) = 1$. Given arbitrary $\omega \in \boldsymbol{A}$, and any $\varepsilon > 0$, we can choose $N(\omega) < \infty$ where $n \geq N(\omega)$ implies $|\,\mathbf{E}_{\mu_n}\, \psi - \mathbf{E}_\mu\, \psi| \leq \varepsilon$, where this notation represents the absolute difference between the sample mean (pre-limit) and expectation taken in the definition of $\boldsymbol{A}$. By definition of $\theta^*(\cdot)$ and monotonicity of ψ, one has that

$$0 = \mathbf{E}_\mu\, \psi\left(\frac{l(h;\boldsymbol{z}) - \theta^*(h)}{s_h(\boldsymbol{z})}\right) > \mathbf{E}_\mu\, \psi\left(\frac{l(h;\boldsymbol{z}) - (\theta^*(h) + \delta)}{s_h(\boldsymbol{z})}\right).$$

It follows that there exists $\varepsilon' > 0$ such that

$$\frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{l(h;\boldsymbol{z}_i)-(\theta^*(h)+\delta)}{s_h(\boldsymbol{z}_i)}\right) \leq (-1)\varepsilon' < 0$$

eventually (in $n \in \mathbb{N}$), on this $\omega \in \boldsymbol{A}$, and similarly by the definition of $\widehat{\theta}(\cdot)$, we have

$$\widehat{\theta}(\mathcal{H}) \leq \widehat{\theta}(h) < \theta^* + \delta.$$

This shows us that $\boldsymbol{A} \subseteq \{\limsup_n \widehat{\theta}(\mathcal{H}) < \theta^*(h) + \delta\}$. Letting $\delta = 1/k$, for each $k = 1, 2, \ldots$ denote the corresponding convergence event $\boldsymbol{A}$ particularly as $\boldsymbol{A}_k$. Noting $\boldsymbol{A}_{k+1} \subseteq \boldsymbol{A}_k$, we have

$$\boldsymbol{A}_m \subseteq \bigcap_{k=1}^{m} \boldsymbol{A}_k, \quad m = 1, 2, \ldots$$

Basic continuity of measures gives us that

$$\mathbf{P}\left\{\limsup_n \widehat{\theta}(\mathcal{H}) \leq \theta^*(h)\right\} = \lim_{m\to\infty} \mathbf{P}\bigcap_{k=1}^{m} \boldsymbol{A}_k = 1,$$

and the same result holds for arbitrary choice of $h \in \mathcal{H}$. Similarly, construct a sequence (h_m) of $h_m \in \mathcal{H}$ such that $\theta^*(h_m) \downarrow \theta^*(\mathcal{H})$. Clearly

$$\left\{\limsup_n \widehat{\theta}(\mathcal{H}) \leq \theta^*(h_{m+1})\right\} \subseteq \left\{\limsup_n \widehat{\theta}(\mathcal{H}) \leq \theta^*(h_m)\right\}$$

with each event occurring with probability 1. Again via measure continuity it follows that

$$\mathbf{P}\left\{\limsup_n \widehat{\theta}(\mathcal{H}) \leq \theta^*(\mathcal{H})\right\} = \lim_{m\to\infty} \mathbf{P}\bigcap_{k=1}^{m}\left\{\limsup_n \widehat{\theta}(\mathcal{H}) \leq \theta^*(h_m)\right\} = 1.$$

Thus we have that

$$\limsup_n \widehat{\theta}(\mathcal{H}) \leq \theta^*(\mathcal{H}) \coloneqq \inf_{h\in\mathcal{H}} \theta^*(h), \quad \text{a.s.}$$

Now we look at the lim inf side of the argument. At this point, we have

$$0 \leq \liminf_n \widehat{\theta}(\mathcal{H}) \leq \limsup_n \widehat{\theta}(\mathcal{H}) \leq \theta^*(\mathcal{H}),$$

which follows from the above argument and the fact that $L \geq 0$, so

$$\widehat{\theta}(h) \geq 0 \text{ and } |\liminf_n \widehat{\theta}(h)| < \infty$$

almost surely. Label the event

$$\boldsymbol{A}' \coloneqq \left\{\liminf_n \widehat{\theta}(\mathcal{H}) < \theta^*(\mathcal{H})\right\},$$

and start by assuming $\mathbf{P}\,\boldsymbol{A}' > 0$. On this event, we can fix a distance $\delta > 0$ such that taking n over $\mathbb{N}$, the sequence $\widehat{\theta}(\mathcal{H})$ drops more than δ below $\theta^*(\mathcal{H})$ infinitely often. To make this more concrete, fix $\theta_L \coloneqq \liminf_n \widehat{\theta}(\mathcal{H})$, and take any $\delta \in (0, \theta^*(\mathcal{H}) - \theta_L)$. Then for all $N < \infty$, can

find index $n \geq N$ where $\widehat{\theta}(\mathcal{H}) < \theta^*(\mathcal{H}) - \delta < \theta^*(\mathcal{H})$. This gap, between $\widehat{\theta}(\mathcal{H})$ and $\theta^*(\mathcal{H})$, of at least δ, occurs infinitely often. For any such n, we have

$$\begin{aligned}\mathbf{E}_\mu \, \psi\left(\frac{l(\widehat{h}_n; \boldsymbol{z}) - \widehat{\theta}(\mathcal{H})}{s_{h_n}(\boldsymbol{z})}\right) &> \mathbf{E}_\mu \, \psi\left(\frac{l(\widehat{h}_n; \boldsymbol{z}) - \theta^*(\mathcal{H})}{s_{\widehat{h}_n}(\boldsymbol{z})}\right) \\ &\geq \mathbf{E}_\mu \, \psi\left(\frac{l(\widehat{h}_n; \boldsymbol{z}) - \theta^*(\widehat{h}_n)}{s_{\widehat{h}_n}(\boldsymbol{z})}\right) \\ &= 0.\end{aligned}$$

The second inequality and the final inequality hold for all n, by the optimality of $\theta^*(\mathcal{H})$ and the definition of $\theta^*(\cdot)$. Depending on the $\omega \in \boldsymbol{A}'$, the actual value of this $\delta > 0$ will differ, but what matters is that such a δ-gap is fixed as we take n over $\mathbb{N}$. By the lim sup bound shown above, taking any $\theta_U \in (\theta^*(\mathcal{H}), \infty)$, we have that $\widehat{\theta}(\mathcal{H}) \in [0, \theta_U]$ eventually. That is, there exists $N < \infty$ where $n \geq N$ implies $\widehat{\theta}(\mathcal{H}) \in [0, \theta_U]$. Using concavity, note for any $u^* > 0$, $u \in [0, u^* - \delta]$, $s > 0$ and $l \geq 0$, we have

$$\psi\left(\frac{l-u}{s}\right) - \psi\left(\frac{l-u^*}{s}\right) \geq \frac{\delta}{s}\psi'\left(\frac{l-u^*+\delta}{s}\right).$$

Set $l = l(h; \boldsymbol{z})$, $u^* = \theta^*(\mathcal{H})$, $s = s_h(\boldsymbol{z})$, and integrate with $\mathbf{E}_\mu$. Note that by assumption, there exists $\epsilon > 0$ such that

$$\delta \, \mathbf{E}_\mu \, \frac{1}{s_{\widehat{h}_n}(\boldsymbol{z})}\psi'\left(\frac{l(\widehat{h}_n; \boldsymbol{z}) - \theta^*(\mathcal{H}) + \delta}{s_{\widehat{h}_n}(\boldsymbol{z})}\right) \geq \frac{\delta}{s_2} \inf_{h \in \mathcal{H}} \mathbf{E}_\mu \, \psi'\left(\frac{l(h; \boldsymbol{z}) - \theta^*(\mathcal{H}) + \delta}{s_1}\right) \geq \epsilon$$

noting that ψ' is non-increasing on $\mathbb{R}_+$, by concavity of ψ. We thus have that on the event $\boldsymbol{A}'$, and any "bad index" n where $\widehat{\theta}(\mathcal{H}) < \theta^*(\mathcal{H}) - \delta$, we have

$$\mathbf{E}_\mu \, \psi\left(\frac{l(\widehat{h}_n; \boldsymbol{z}) - \widehat{\theta}(\mathcal{H})}{s_{\widehat{h}_n}(\boldsymbol{z})}\right) - \mathbf{E}_\mu \, \psi\left(\frac{l(\widehat{h}_n; \boldsymbol{z}) - \theta^*(\mathcal{H})}{s_{\widehat{h}_n}(\boldsymbol{z})}\right) \geq \epsilon > 0.$$

Since this occurs infinitely often as n ranges over $\mathbb{N}$ and ϵ is free of n, it implies

$$\boldsymbol{A}' \subseteq \left\{\lim_{n\to\infty} \mathbf{E}_\mu \, \psi\left(\frac{l(\widehat{h}_n; \boldsymbol{z}) - \widehat{\theta}(\mathcal{H})}{s_{\widehat{h}_n}(\boldsymbol{z})}\right) = 0\right\}^c,$$

which contradicts the strong convergence guaranteed by Corollary 15, noting $\widehat{\theta}(\widehat{h}_n) = \widehat{\theta}(\mathcal{H})$ by definition and Theorem 10. We conclude $\mathbf{P}\,\boldsymbol{A}' = 0$, which is to say that almost surely

$$\liminf_n \, \widehat{\theta}(\mathcal{H}) \geq \theta^*(\mathcal{H}) \geq \limsup_n \, \widehat{\theta}(\mathcal{H}).$$

We thus conclude that $\widehat{\theta}(\widehat{h}_n) = \widehat{\theta}(\mathcal{H}) \to \theta^*(\mathcal{H})$ as $n \to \infty$. □

Proof of Proposition 17. We verify the statements by adapting a standard comparison function technique [28, Lemma 7.7]. Fix arbitrary sample $x_1, \ldots, x_n$ where $x_i = l(h; \boldsymbol{z}_i)$ in the setting of this paper. We consider the case of arbitrary s, where it may be completely determined by μ_n. Here defining two functions

$$\begin{aligned}g(\theta) &\coloneqq \frac{1}{n}\sum_{i=1}^n \rho\left(\frac{x_i - \theta}{s}\right) s \\ \widetilde{g}(u; \theta) &\coloneqq g(\theta) + \frac{1}{2ns}\sum_{i=1}^n \left(\left(\psi\left(\frac{x_i-\theta}{s}\right)s - u\right)^2 - \psi\left(\frac{x_i-\theta}{s}\right)^2 s^2\right),\end{aligned}$$

for any choice of $\theta, u \in \mathbb{R}$, we have a bound from above in $\widetilde{g}(u;\theta) \geq g(\theta+u)$. To see this, note that the difference function $d_\theta(u) \coloneqq \widetilde{g}(u;\theta) - g(\theta+u)$ satisfies

$$d_\theta(0) = 0, \quad d'_\theta(0) = 0, \quad d''_\theta \geq 0$$

for any choice of θ. The first two follow immediately from definitions, and the final inequality follows from $\rho'' \leq 1$ assuming we've standardized ρ such that ρ' is 1-Lipschitz, noting

$$d''_\theta(u) = \frac{1}{s}\left(1 - \frac{1}{n}\sum_{i=1}^n \psi'\left(\frac{x_i - (\theta+u)}{s}\right)\right),$$

which implies $d_\theta(u) \geq 0$ for all $u \in \mathbb{R}$, and also

$$g(\theta) - g(\theta+u) \geq g(\theta) - \widetilde{g}(u;\theta). \tag{21}$$

To make the best possible update to θ, we should set u to maximize the right-hand side, equivalently minimize $\widetilde{g}(u;\theta)$. Noting $\widetilde{g}'' = 1/s > 0$, and defining

$$u_0(\theta) \coloneqq \frac{s}{n}\sum_{i=1}^n \psi\left(\frac{x_i - \theta}{s}\right)$$

we have $\widetilde{g}'(u_0(\theta)) = 0$ which is thus the unique minimum. Plugging $u_0(\theta)$ into (21), some algebra reveals

$$g(\theta) - g(\theta + u_0(\theta)) \geq \frac{1}{2s}u_0(\theta)^2 \geq 0. \tag{22}$$

Note that the right-hand side is zero iff $\theta = \widehat{\theta}(h)$, otherwise it is strictly positive. Defining $\widehat{\theta}_{(k)} \coloneqq \widehat{\theta}_{(k-1)} + u_0(\widehat{\theta}_{(k-1)})$ is equivalent to the update (3). Looking at sequences taking $k \in \mathbb{N}$, $g(\widehat{\theta}_{(k)})$ is bounded and monotonic, and thus convergent. Since it is also Cauchy, this naturally implies $u_0(\widehat{\theta}_{[t]}) \to 0$ as well, from which it follows that $\widehat{\theta}_{(k)} \to \widehat{\theta}(h)$. To see this, assume $\widehat{\theta}_{(k)}$ is not Cauchy. Then there exists scale $\varepsilon_0 > 0$ at which for any $K < \infty$, there exist $k, k' \geq K$ such that $|\widehat{\theta}_{(k)} - \widehat{\theta}_{(k')}| > \varepsilon_0$. By the update definition and (22), for fixed sample $\boldsymbol{Z}$ the sequence $\widehat{\theta}_{(k)}$ is bounded. For concreteness, denote these bounds as $0 \leq \widehat{\theta}_{(k)} \leq \theta_U$. By strong monotonicity of ψ it then follows that defining

$$\varepsilon_1 \coloneqq \inf_{\theta \in [0,\theta_U]} \left|\sum_{i=1}^n \left(\psi\left(\frac{x_i - \theta}{s}\right) - \psi\left(\frac{x_i - (\theta \pm \varepsilon_0)}{s}\right)\right)\right|$$

we have $\varepsilon_1 > 0$, and this constant is determined the moment that sample $\boldsymbol{Z}$ is observed and the update routine is intitialized. On the bad indices k, k' where $|\widehat{\theta}_{(k)} - \widehat{\theta}_{(k')}| > \varepsilon_0$, we always have

$$\left|\sum_{i=1}^n \left(\psi\left(\frac{x_i - \widehat{\theta}_{(k)}}{s}\right) - \psi\left(\frac{x_i - \widehat{\theta}_{(k')}}{s}\right)\right)\right| \geq \varepsilon_1$$

which would imply that $u_0(\widehat{\theta}_{(k)})$ is not Cauchy, contradicting $u_0(\widehat{\theta}_{(k)}) \to 0$. Thus $\widehat{\theta}_{(k)}$ is convergent. Using continuity of ψ, we have

$$u_0\left(\lim_{k\to\infty} \widehat{\theta}_{(k)}\right) = \lim_{k\to\infty} u_0(\widehat{\theta}_{(k)}) = 0,$$

implying $\widehat{\theta}_{(k)} \to \widehat{\theta}(h)$.

Shifting our focus to the scale result, consider χ as in Defn. 11, but with some additional restrictions. Similar to ψ in the location estimation setting, treat χ as a gradient of some convex objective to be minimized. The general form of the objective function is to be

$$g(s) \coloneqq \mathbf{E}_{\mu_n}\left(r\left(\frac{x-\gamma}{s}\right)+\beta\right)s, \quad s>0$$

where the function $r(\cdot)$ is assumed to be $r \geq 0$, convex and even, with a unique minimum at $r(0)=0$. In addition, $r(u)/u$ should be concave on $\mathbb{R}_+$. The idea then is to construct χ using the gradient of this auxiliary objective, namely we seek that

$$g'(s) = (-1)\,\mathbf{E}_{\mu_n}\left(\chi\left(\frac{x-\gamma}{s}\right)\right). \tag{23}$$

To achieve this given a valid r, one need only set $\chi(u) \coloneqq r'(u)u - r(u) - \beta$.

A brief remark on constructing valid robust control functions of this form. Perhaps the simplest setting of r with the desired properties is $r(u) = u^{1+k}$, for $k \in (0,1]$. Clearly $r'' > 0$ on $\mathbb{R}_+$, and since $(r(u)/u)'' = k(k-1)u^{k-2} \leq 0$, we have the concavity desired. Furthermore, $\chi(u) = ku^{1+k} - \beta$, and

$$s = \left(\frac{k}{\beta}\,\mathbf{E}_{\mu_n}\left(x-\gamma\right)^{1+k}\right)^{\frac{1}{1+k}}$$

is the unique root of $\mathbf{E}_{\mu_n}\chi((x-\gamma)/s)$ in $s>0$. There are no issues with zero-valued solutions given this formulation.

Returning to the main proof, a critical property of the update (4) is that for any $k = 1, 2, \ldots$ we have

$$g(s_{(k)}) - g(s_{(k+1)}) \geq \frac{\beta}{s_{(k)}}\left(s_{(k+1)} - s_{(k)}\right)^2. \tag{24}$$

To simplify notation even further, denote $l_i \coloneqq x_i - \gamma$ for $i = 1, \ldots, n$. We set $\chi(u) \coloneqq r'(u)u - r(u) - \beta$ as above, and denote $\widetilde{\chi}(u) \coloneqq \chi(u) + \beta$. Just as for the location case, a comparison function is introduced of the form

$$\widetilde{g}(u;s) \coloneqq g(s) + (u-s)\beta + \frac{1}{n}\sum_{i=1}^{n}\widetilde{\chi}\left(\frac{l_i}{s}\right)\left(\frac{s^2}{u} - s\right).$$

A few remarks regarding this form. First of all, we want $\widetilde{g}(s;s) = g(s)$, thus the need for the first constant. The second term ensures β appears in the first derivative of $\widetilde{g}$. The third term takes the form that it does such that in addition to $u = s$ implying $g = \widetilde{g}$, we also get that $g'(\cdot)$ and $\widetilde{g}'(\cdot;s)$ coincide when evaluated at s. With this form, it is immediate as

$$\widetilde{g}'(u;s) = \frac{1}{n}\sum_{i=1}^{n}\widetilde{\chi}\left(\frac{l_i}{s}\right)\frac{s^2}{u^2}(-1) + \beta$$

since we can note

$$\begin{aligned} g'(s) &= \frac{1}{n}\sum_{i=1}^{n} r'\left(\frac{l_i}{s}\right)\left(\frac{l_i}{s}\right)(-1) + \frac{1}{n}\sum_{i=1}^{n} r\left(\frac{l_i}{s}\right) + \beta \\ &= (-1)\frac{1}{n}\sum_{i=1}^{n}\chi\left(\frac{l_i}{s}\right) \\ &= \widetilde{g}'(s;s). \end{aligned}$$

Defining the difference function for pre-fixed arbitrary $s > 0$ by $d_s(u) \coloneqq \widetilde{g}(u; s) - g(u)$, we have that $d_s(s) = 0$, $d_s'(s) = 0$. Since we want to show $d_s(u) \geq 0$ for all $u > 0$, it remains to show that $d_s(\cdot)$ is convex. This is straightforward, if one notices that there are positive constants α_0 and α_1 which depend on s but are free of u such that

$$\begin{aligned} d_s(u) &= \alpha_0 + \frac{\alpha_1}{u} + \frac{-1}{n} \sum_{i=1}^{n} r\left(\frac{l_i}{u}\right) u \\ &= \alpha_0 + \alpha_1 \sigma + \frac{-1}{n} \sum_{i=1}^{n} r\left(l_i \sigma\right) \frac{1}{\sigma} \end{aligned}$$

when defining $\sigma \coloneqq 1/u$. The first two terms together form an affine function of σ, and by assumption $r(u)/u$ is a concave function on $\mathbb{R}_+$. Note that having l_i scaling this has no impact on convexity, since letting $f(u) \coloneqq r(u)/u$ and for any $\alpha \neq 0$ setting $\widetilde{f}(u) \coloneqq r(\alpha u)/(\alpha u)$, using first-order characterization of concavity, we have for any $u, v \geq 0$ that

$$\begin{aligned} \widetilde{f}(u) - \widetilde{f}(v) &= f(\alpha u) - f(\alpha v) \\ &\leq (u - v) \alpha f'(\alpha v) \\ &= (u - v) \widetilde{f}'(v), \end{aligned}$$

showing $\widetilde{f}$ is concave on $\mathbb{R}_+$ when f is. Thus the third summand in d_s is a convex function of $\sigma > 0$, and $d_s(1/\sigma) \geq 0$ for all $\sigma > 0$, implying $d_s(u) \geq 0$ for all $u > 0$ as desired, and $\widetilde{g}(u; s) \geq g(u)$ for all $u > 0$. Since we seek an update routine where g gets smaller, fixing $s > 0$ as the scale value from a previous iteration, we naturally seek that $g(s) - g(u)$ is maximized in u. Note that $\widetilde{g}(\cdot; s)$ has its unique critical point at

$$u_A = \left(\frac{1}{n\beta} \sum_{i=1}^{n} \widetilde{\chi}\left(\frac{l_i}{s}\right) s^2 \right)^{1/2} = s \left(1 + \frac{1}{n\beta} \sum_{i=1}^{n} \chi\left(\frac{l_i}{s}\right) \right)^{1/2}$$

noting that the term inside the square root is non-negative as $\chi \geq -\beta$ by definition. Plugging u_A into $\widetilde{g}(\cdot; s)$ and some algebra then readily yields

$$\begin{aligned} g(s) - g(u_A) &\geq g(s) - \widetilde{g}(u_A; s) \\ &= \frac{\beta}{s} (u_A - s)^2 \end{aligned}$$

and thus implying 24 by the update definition (4).

We now move on to the final step of this proof. Initialize using $s_{(k)} > 0$. Beginning with some basic facts, note that by (24), we have

$$g(s_{(0)}) \geq g(s_{(k)}) \geq g(s_{(k+1)}) \geq 0,$$

so $g(s_{(k)})$ is a bounded, monotone sequence, and thus the limit $\lim_{k \to \infty} g(s_{(k)})$ certainly exists and is finite. As for the sequence $s_{(k)}$, note first that

$$\widetilde{\chi}'(u) = \chi'(u) = u r''(u) > 0, \quad \forall\, u > 0.$$

It follows that χ and $\widetilde{\chi}'$ are uniquely minimized at 0, meaning in particular that unless $l_1 = \cdots = l_n = 0$, we have $\mathbf{E}_{\mu_n} \widetilde{\chi}(l_i / s_{[0]}) > 0$. Assuming a continuous distribution function, this occurs with probability zero. Thus by definition of the update rule, $s_{(k)} > 0$ almost surely for all $k \in \mathbb{N}$. Henceforth we assume at least once $l_i \neq 0$. An upper bound is also simple

to check. Taking $s \to \infty$, necessarily $g(s) \to \infty$, meaning $g(s) > g(s_{(0)})$ for s large enough. Since $g(s_{(k)}) > g(s_{(0)})$ is a contradiction, necessarily $s_{(k)}$ is bounded above as well. Regarding convergence, note that

$$g''(u) = \frac{1}{n}\sum_{i=1}^{n} r''\left(\frac{l_i}{u}\right)\left(\frac{l_i^2}{u^3}\right),$$

so by strong convexity of r, $g'' > 0$ on $\mathbb{R}_+$, and has a unique minimum. Denote this by $u_0 := \arg\min g(u)$. Certainly either $u_0 = 0$ or $u_0 > 0$ are possible, but convergence is readily confirmed as follows. Since

$$\begin{aligned} g(s_{(k)}) = g(s_{(k+1)}) &\iff \frac{1}{n}\chi\left(\frac{l_i}{s_{(k)}}\right) = 0 \\ &\iff g(s_{(k)}) = \min_u g(u), \end{aligned}$$

we have that $g(s_{(k)}) \to \min_u g(u) = g(u_0)$ as $t \to \infty$. Now say $s_{(k)}$ under update (A) is not Cauchy. Then, there exists some $\varepsilon_0 > 0$ such that for any $K \in \mathbb{N}$, we can find bad indices $k_1, k_2 \geq K$ such that $|s_{(k_1)} - s_{(k_2)}| > \varepsilon_0$. Note that by continuity and strong convexity of g, for any $\varepsilon > 0$, we can find a $\delta > 0$ such that $|s - u_0| > \delta \implies |g(s) - g(u_0)| > \varepsilon$. Taking ε arbitrarily small lets us take δ arbitrarily small. Choose $\varepsilon > 0$ such that $\delta \leq \varepsilon_0/2$. Since $g(s_{(k)}) \to g(u_0)$, exists K_0 such that $k \geq K_0$ implies $|g(s_{(k)}) - g(u_0)| \leq \varepsilon$. Taking $K \geq K_0$ and bad indices $k_1, k_2 \geq K$, we have $|s_{(k_1)} - s_{(k_2)}| > \varepsilon_0 \geq 2\delta$, but also $|g(s_{(k)}) - g(u_0)| \leq \varepsilon$ for both $t = k_1, k_2$. Taking k_1 for instance, note that $s_{(k_1)} \in [u_0 - \delta, u_0 + \delta]$. Looking at k_2 then, one sees

$$\begin{aligned} |s_{(k_1)} - s_{(k_2)}| > \varepsilon_0 &\implies s_{(k_2)} \notin [u_0 - \delta, u_0 + \delta] \\ &\implies |g(s_{(k_2)}) - g(u_0)| > \varepsilon, \end{aligned}$$

a contradiction since $k_2 \geq K \geq K_0$. We conclude that $s_{(k)}$ must be Cauchy and thus convergent to the unique minimizer u_0, implying the desired result. □

Remark 18. It should be noted that the convergence of (4) given by Proposition 17 is convergence to a *solution*, but it is possible that the solution may in fact be zero. This depends on the loss observations (and thus choice of $h \in \mathcal{H}$), the form of r, and the value of $\chi(0) < 0$ in a rather complex manner. For any given sample $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ and candidate h, the solution will be positive if and only if $\mathbf{E}_{\mu_n} \chi((l(h; \boldsymbol{z}) - \gamma)/s)$ can be made positive for small enough $s > 0$, and the natural control for this is to ensure $\chi(0)$ is far enough below zero. Thus if χ is built following (23) with a strictly convex r and small enough β, one can rest assured that the $s_{(k)}$ updates of 4 used as a sub-routine in Algorithm 1 will converge to a positive solution.

C One-dimensional example

Data generation To create Figure 1, a simple experiment was carried out, as follows. On the left half, we have the well-behaved symmetric noise setting, while the noise on the right-hand side is asymmetric and heavy-tailed. More precisely, for the left side, we generated $(x_1, y_1), \ldots, (x_n, y_n)$ by $x \sim \text{Unif}[0, 0.5]$ and $y = f(x) + \epsilon$, where the noise ϵ is independent of x, and zero-mean Normal with variance $\mathbf{E}\,\epsilon^2 = 0.25$. The functional relation is $f(x) = \sin(2\pi x)$. To generate the right half, the same procedure was done, this time with $x \sim U[0.5, 1]$, and noise ϵ being a Fréchet random variable with shift 0, scale 1, and shape 2.1, shifted such that $\mathbf{E}\,\epsilon = 0$. The noise magnitude on the right side is larger as well, with $\sqrt{\mathbf{E}\,\epsilon^2} \approx 4$.

Solid curves We illustrate the deterministic parameters of the conditional distribution of response y as a function of input x using solid curves. The solid red curve is the graph of $\mathbf{E}(y;x) = \mathbf{E}\,\epsilon + f(x) = f(x)$, taking $x \in [0,1]$. Solid green denotes the graph of $\operatorname{med}(y;x) = \operatorname{med}\epsilon + f(x)$. On the left side, $\operatorname{med}\epsilon = 0$, but on the right half this is not the case and the two graphs diverge. Finally, solid blue denotes the ρ-induced M-estimate of the location of y, conditioned on x. More precisely, the graph of $\widetilde{y}(x) \coloneqq \arg\min_\theta \mathbf{E}\,\rho((f(x)+\varepsilon-\theta)/s)$. Here we used the Gudermannian for ρ, and set $s = \operatorname{med}|\epsilon|$ independent of x.

Dashed curves Next we look at statistical estimates of the deterministic parameters just mentioned, based on the sample. A simple fifth-degree polynomial model is assumed, taking the form $\widehat{h}(x) = \sum_{k=1}^{5} w_k x^k$. Running OLS to specify the weights results in the red dashed curve (the graph of $\widehat{h}$). Similarly, the green dashed line is the estimate due to running LAD. Finally, the blue dashed line is the product of running `fRLM` (Algorithm 1) just as specified in section 5.2. Each algorithm was run twice: once for the well-behaved data on the left domain, and once for the uncongenial data on the right domain. Finally, we should remark that while OLS and LAD do indeed correspond to trying to learn $\mathbf{E}(y;x)$ and $\operatorname{med}(y;x)$ respectively, the correspondence is not quite so clear for `fRLM`. Here the routine explicitly tries to minimize $\theta^*(h)$ from Defn. 6, though the relation between h^* satisfying $\theta^*(h^*) = \theta^*(\mathcal{H})$ and the closeness of $h^* \approx \widetilde{y}$ remains a matter of both technical and conceptual interest.