Supplemental material: Better generalization with less data using robust gradient descent

Matthew J. Holland
Osaka University

Kazushi Ikeda
Nara Institute of Science and Technology

A Technical appendix

A.1 Preliminaries

Our generic data shall be denoted by $\boldsymbol{z} \in \mathcal{Z}$. Let μ denote a probability measure on $\mathcal{Z}$, equipped with an appropriate σ-field. Data samples shall be assumed independent and identically distributed (iid), written $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$. We shall work with loss function $l : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}_+$ throughout, with $l(\cdot; \boldsymbol{z})$ assumed differentiable for each $\boldsymbol{z} \in \mathcal{Z}$. Write $\mathbf{P}$ for a generic probability measure, most commonly the product measure induced by the sample. Let $f : \mathcal{Z} \to \mathbb{R}$ be an measurable function. Expectation is written $\mathbf{E}_\mu f(\boldsymbol{z}) := \int f \, d\mu$, with variance $\operatorname{var}_\mu f(\boldsymbol{z})$ defined analogously. For d-dimensional Euclidean space $\mathbb{R}^d$, the usual (ℓ_2) norm shall be denoted $\|\cdot\|$ unless otherwise specified. For function F on $\mathbb{R}^d$ with partial derivatives defined, write the gradient as $F'(\boldsymbol{u}) := (F'_1(\boldsymbol{u}), \ldots, F'_d(\boldsymbol{u}))$ where for short, we write $F'_j(\boldsymbol{u}) := \partial F(\boldsymbol{u})/\partial u_j$. For integer k, write $[k] := \{1, \ldots, k\}$ for all the positive integers from 1 to k. Risk shall be denoted $R(\boldsymbol{w}) := \mathbf{E}_\mu \, l(\boldsymbol{w}; \boldsymbol{z})$, and its gradient $\boldsymbol{g}(\boldsymbol{w}) := R'(\boldsymbol{w})$. We make a running assumption that we can differentiate under the integral sign in each coordinate [1, 6], namely that

$$\boldsymbol{g}(\boldsymbol{w}) = \left(\mathbf{E}_\mu \frac{\partial l(\boldsymbol{w}; \boldsymbol{z})}{\partial w_1}, \ldots, \mathbf{E}_\mu \frac{\partial l(\boldsymbol{w}; \boldsymbol{z})}{\partial w_d} \right). \tag{1}$$

Smoothness and convexity of functions shall also be utilized. For convex function F on convex set $\mathcal{W}$, say that F is λ-*Lipschitz* if, for all $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathcal{W}$ we have $|F(\boldsymbol{w}_1) - F(\boldsymbol{w}_2)| \leq \lambda \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|$. We say that F is λ-*smooth* if F' is λ-Lipschitz. Finally, F is *strongly convex* with parameter $\kappa > 0$ if for all $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathcal{W}$,

$$F(\boldsymbol{w}_1) - F(\boldsymbol{w}_2) \geq \langle F'(\boldsymbol{w}_2), \boldsymbol{w}_1 - \boldsymbol{w}_2 \rangle + \frac{\kappa}{2} \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|^2$$

for any norm $\|\cdot\|$ on $\mathcal{W}$, though we shall be assuming $\mathcal{W} \subseteq \mathbb{R}^d$. If there exists $\boldsymbol{w}^* \in \mathcal{W}$ such that $F'(\boldsymbol{w}^*) = 0$, then it follows that $\boldsymbol{w}^*$ is the unique minimum of F on $\mathcal{W}$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable, convex, λ-smooth function. The following basic facts will be useful: for any choice of $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, we have

$$f(\boldsymbol{u}) - f(\boldsymbol{v}) \leq \frac{\lambda}{2} \|\boldsymbol{u} - \boldsymbol{v}\|^2 + \langle f'(\boldsymbol{v}), \boldsymbol{u} - \boldsymbol{v} \rangle \tag{2}$$

$$\frac{1}{2\lambda} \|f'(\boldsymbol{u}) - f'(\boldsymbol{v})\|^2 \leq f(\boldsymbol{u}) - f(\boldsymbol{v}) - \langle f'(\boldsymbol{v}), \boldsymbol{u} - \boldsymbol{v} \rangle. \tag{3}$$

Proofs of these results can be found in any standard text on convex optimization, e.g. [5].

We shall leverage a special type of M-estimator here, built using the following convenient class of functions.

Definition 1 (Function class for location estimates)**.** Let $\rho : \mathbb{R} \to [0, \infty)$ be an even function ($\rho(u) = \rho(-u)$) with $\rho(0) = 0$ and the following properties. Denote $\psi(u) := \rho'(u)$.

1. $\rho(u) = O(u)$ as $u \to \pm\infty$.
2. $\rho(u)/(u^2/2) \to 1$ as $u \to 0$.
3. $\psi' > 0$, and for some $C > 0$, and all $u \in \mathbb{R}$,

$$-\log(1 - u + Cu^2) \leq \psi(u) \leq \log(1 + u + Cu^2).$$

Of particular importance in the proceeding analysis is the fact that $\psi = \rho'$ is bounded, monotonically increasing and Lipschitz on $\mathbb{R}$, plus the upper/lower bounds which let us generalize the technique of Catoni [3].

Example 2 (Valid ρ choices). In addition to the Gudermannian function (section 2 footnote), functions such as $2(\sqrt{1 + u^2/2} - 1)$ and $\log\cosh(u)$ are well-known examples that satisfy the desired criteria. Note that the wide/narrow functions of Catoni do not meet all these criteria, nor does the classic Huber function.

A.2 Proofs

Proof of Lemma 1 (main text). For cleaner notation, write $x_1, \ldots, x_n \in \mathbb{R}$ for our iid observations. Here ρ is assumed to satisfy the conditions of Definition 1. A high-probability concentration inequality follows by direct application of the specified properties of ρ and $\psi := \rho'$, following the general technique laid out by Catoni [2, 3]. For $u \in \mathbb{R}$ and $s > 0$, writing $\psi_s(u) := \psi(u/s)$, and taking expectation over the random draw of the sample,

$$\begin{aligned}
\mathbf{E} \exp\left(\sum_{i=1}^{n} \psi_s(x_i - u)\right) &\leq \left(1 + \frac{1}{s}(\mathbf{E}\, x - u) + \frac{C}{s^2}\,\mathbf{E}(x^2 + u^2 - 2xu)\right)^n \\
&\leq \exp\left(\frac{n}{s}(\mathbf{E}\, x - u) + \frac{Cn}{s^2}(\operatorname{var} x + (\mathbf{E}\, x - u)^2)\right).
\end{aligned}$$

The inequalities above are due to an application of the upper bound on ψ, and and the inequality $(1 + u) \leq \exp(u)$. Now, letting

$$\begin{aligned}
A &:= \frac{1}{n}\sum_{i=1}^{n} \psi_s(x_i - u) \\
B &:= \frac{1}{s}(\mathbf{E}\, x - u) + \frac{C}{s^2}(\operatorname{var} x + (\mathbf{E}\, x - u)^2)
\end{aligned}$$

we have a bound on $\mathbf{E}\exp(nA) \leq \exp(nB)$. By Chebyshev's inequality, we then have

$$\begin{aligned}
\mathbf{P}\{A > B + \varepsilon\} &= \mathbf{P}\{\exp(nA) > \exp(nB + n\varepsilon)\} \\
&\leq \frac{\mathbf{E}\exp(nA)}{\exp(nB + n\varepsilon)} \\
&\leq \exp(-n\varepsilon).
\end{aligned}$$

Setting $\varepsilon = \log(\delta^{-1})/n$ for confidence level $\delta \in (0, 1)$, and for convenience writing

$$b(u) := \mathbf{E}\, x - u + \frac{C}{s}(\operatorname{var} x + (\mathbf{E}\, x - u)^2),$$

we have with probability no less than $1-\delta$ that

$$\frac{s}{n}\sum_{i=1}^{n}\psi_s(x_i-u) \leq b(u) + \frac{s\log(\delta^{-1})}{n}. \tag{4}$$

The right hand side of this inequality, as a function of u, is a polynomial of order 2, and if

$$1 \geq D \coloneqq 4\left(\frac{C^2 \operatorname{var} x}{s^2} + \frac{C\log(\delta^{-1})}{n}\right),$$

then this polynomial has two real solutions. In the hypothesis, we stated that the result holds "for large enough n and s_j." By this we mean that we require n and s to satisfy the preceding inequality (for each $j \in [d]$ in the multi-dimensional case). The notation D is for notational simplicity. The solutions take the form

$$u = \frac{1}{2}\left(2\,\mathbf{E}\,x + \frac{s}{C} \pm \frac{s}{C}(1-D)^{1/2}\right).$$

Looking at the smallest of the solutions, noting $D \in [0,1]$ this can be simplified as

$$\begin{aligned} u_+ &\coloneqq \mathbf{E}\,x + \frac{s}{2C}\frac{(1-\sqrt{1-D})(1+\sqrt{1-D})}{1+\sqrt{1-D}} \\ &= \mathbf{E}\,x + \frac{s}{2C}\frac{D}{1+\sqrt{1-D}} \\ &\leq \mathbf{E}\,x + sD/2C, \end{aligned} \tag{5}$$

where the last inequality is via taking the $\sqrt{1-D}$ term in the previous denominator as small as possible. Now, writing $\widehat{x}$ as the M-estimate using s and ρ as in (3, main text), note that $\widehat{x}$ equivalently satisfies $\sum_{i=1}^{n}\psi_s(\widehat{x}-x_i)=0$. Using (4), we have

$$\frac{s}{n}\sum_{i=1}^{n}\psi_s(x_i-u_+) \leq b(u_+) + \frac{s\log(\delta^{-1})}{n} = 0,$$

and since the left-hand side of (4) is a monotonically decreasing function of u, we have immediately that $\widehat{x} \leq u_+$ on the event that (4) holds, which has probability at least $1-\delta$. Then leveraging (5), it follows that on the same event,

$$\widehat{x} - \mathbf{E}\,x \leq sD/2C.$$

An analogous argument provides a $1-\delta$ event on which $\widehat{x} - \mathbf{E}\,x \geq -sD/2C$, and thus using a union bound, one has that

$$|\widehat{x} - \mathbf{E}\,x| \leq 2\left(\frac{C \operatorname{var} x}{s} + \frac{s\log(\delta^{-1})}{n}\right) \tag{6}$$

holds with probability no less than $1-2\delta$. Setting the x_i to $l'_j(\boldsymbol{w};\boldsymbol{z}_i)$ for $j \in [d]$ and some $\boldsymbol{w} \in \mathbb{R}^d$, $i \in [n]$, and $\widehat{x}$ to $\widehat{\theta}_j$ corresponds to the special case considered in this Lemma. Dividing δ by two yields the $(1-\delta)$ result. □

Proof of Lemma 3 (main text). For any fixed $\boldsymbol{w}$ and $j \in [d]$, note that

$$
\begin{aligned}
|\widehat{\theta}_j - g_j(\boldsymbol{w})| \leq \varepsilon_j & \\
& := 2\left(\frac{C \operatorname{var}_\mu l_j'(\boldsymbol{w};\boldsymbol{z})}{s_j} + s_j \log(2\delta^{-1})\right) \qquad (7)\\
& = 2\sqrt{\frac{\log(2\delta^{-1})}{n}} \left(\frac{C \operatorname{var}_\mu l_j'(\boldsymbol{w};\boldsymbol{z})}{\widehat{\sigma}_j} + \widehat{\sigma}_j\right)\\
& \leq \varepsilon^* := 2\sqrt{\frac{V \log(2\delta^{-1})}{n}} c_0 \qquad (8)
\end{aligned}
$$

holds with probability no less than $1-\delta$. The first inequality holds via direct application of Lemma 1 (main text), which holds under (10, main text) and using ρ which satisfies (7, main text). The equality follows immediately from (5, main text). The final inequality follows from (A4) and (9, main text), along with the definition of c_0.

Making the dependence on $\boldsymbol{w}$ explicit with $\widehat{\theta}_j = \widehat{\theta}_j(\boldsymbol{w})$, an important question to ask is how sensitive this estimator is to a change in $\boldsymbol{w}$. Say we perturb $\boldsymbol{w}$ to $\widetilde{\boldsymbol{w}}$, so that $\|\boldsymbol{w} - \widetilde{\boldsymbol{w}}\| = a > 0$. By (A2), for any sample we have

$$\|l'(\boldsymbol{w};\boldsymbol{z}_i) - l'(\widetilde{\boldsymbol{w}};\boldsymbol{z}_i)\| \leq \lambda\|\boldsymbol{w} - \widetilde{\boldsymbol{w}}\| = \lambda a, \quad i \in [n]$$

which immediately implies $|l_j'(\boldsymbol{w};\boldsymbol{z}_i) - l_j'(\widetilde{\boldsymbol{w}};\boldsymbol{z}_i)| \leq \lambda a$ for all $j \in [d]$ as well. Given a sample of $n \geq 1$ points, the most extreme shift in $\widehat{\theta}_j(\cdot)$ that is feasible would be if, given the a-sized shift from $\boldsymbol{w}$ to $\widetilde{\boldsymbol{w}}$, *all* data points moved the maximum amount (namely λa) in the same direction. Since $\widehat{\theta}_j(\widetilde{\boldsymbol{w}})$ is defined by balancing the distance between points to its left and right, the most it could conceivably shift is thus equal to λa. That is, smoothness of the loss function immediately implies a Lipschitz property of the estimator,

$$|\widehat{\theta}_j(\boldsymbol{w}) - \widehat{\theta}_j(\widetilde{\boldsymbol{w}})| \leq \lambda\|\boldsymbol{w} - \widetilde{\boldsymbol{w}}\|.$$

Considering the vector of estimates $\widehat{\boldsymbol{\theta}}(\boldsymbol{w}) := (\widehat{\theta}_1(\boldsymbol{w}), \ldots, \widehat{\theta}_d(\boldsymbol{w}))$, we then have

$$\|\widehat{\boldsymbol{\theta}}(\boldsymbol{w}) - \widehat{\boldsymbol{\theta}}(\widetilde{\boldsymbol{w}})\| \leq \sqrt{d}\lambda\|\boldsymbol{w} - \widetilde{\boldsymbol{w}}\|. \qquad (9)$$

This will be useful for proving uniform bounds on the estimation error shortly.

First, let's use these one-dimensional results for statements about the vector estimator of interest. In d dimensions, using $\widehat{\boldsymbol{\theta}}(\boldsymbol{w})$ just defined for any pre-fixed $\boldsymbol{w}$, then for any $\varepsilon > 0$ we have

$$
\begin{aligned}
\mathbf{P}\left\{\|\widehat{\boldsymbol{\theta}}(\boldsymbol{w}) - \boldsymbol{g}(\boldsymbol{w})\| > \varepsilon\right\} &= \mathbf{P}\left\{\|\widehat{\boldsymbol{\theta}}(\boldsymbol{w}) - \boldsymbol{g}(\boldsymbol{w})\|^2 > \varepsilon^2\right\}\\
&\leq \sum_{j=1}^{d} \mathbf{P}\left\{|\widehat{\theta}_j(\boldsymbol{w}) - \boldsymbol{g}_j(\boldsymbol{w})| > \frac{\varepsilon}{\sqrt{d}}\right\}.
\end{aligned}
$$

Using the notation of ε_j and ε^* from (7), filling in $\varepsilon = \sqrt{d}\varepsilon^*$, we thus have

$$
\begin{aligned}
\mathbf{P}\left\{\|\widehat{\boldsymbol{\theta}}(\boldsymbol{w}) - \boldsymbol{g}(\boldsymbol{w})\| > \sqrt{d}\varepsilon^*\right\} &\leq \sum_{j=1}^{d} \mathbf{P}\left\{|\widehat{\theta}_j(\boldsymbol{w}) - g_j(\boldsymbol{w})| > \varepsilon^*\right\}\\
&\leq \sum_{j=1}^{d} \mathbf{P}\left\{|\widehat{\theta}_j(\boldsymbol{w}) - g_j(\boldsymbol{w})| > \varepsilon_j\right\}\\
&\leq d\delta.
\end{aligned}
$$

The second inequality is because $\varepsilon_j \leq \varepsilon^*$ for all $j \in [d]$. It follows that the event

$$\mathcal{E}(\boldsymbol{w}) := \left\{ \|\widehat{\boldsymbol{\theta}}(\boldsymbol{w}) - \boldsymbol{g}(\boldsymbol{w})\| > 2\sqrt{\frac{dV \log(2d\delta^{-1})}{n}} c_0 \right\}$$

has probability $\mathbf{P}\,\mathcal{E}(\boldsymbol{w}) \leq \delta$. In practice, however, $\widehat{\boldsymbol{w}}_{(t)}$ for all $t > 0$ will be random, and depend on the sample. We seek uniform bounds using a covering number argument. By (A1), $\mathcal{W}$ is closed and bounded, and thus compact, and it requires no more than $N_\epsilon \leq (3\Delta/2\epsilon)^d$ balls of ϵ radius to cover $\mathcal{W}$, where Δ is the diameter of $\mathcal{W}$.[1] Write the centers of these ϵ balls by $\{\widetilde{\boldsymbol{w}}_1, \ldots, \widetilde{\boldsymbol{w}}_{N_\epsilon}\}$. Given $\boldsymbol{w} \in \mathcal{W}$, denote by $\widetilde{\boldsymbol{w}} = \widetilde{\boldsymbol{w}}(\boldsymbol{w})$ the center closest to $\boldsymbol{w}$, which satisfies $\|\boldsymbol{w} - \widetilde{\boldsymbol{w}}\| \leq \epsilon$. Estimation error is controllable using the following new error terms:

$$\|\widehat{\boldsymbol{\theta}}(\boldsymbol{w}) - \boldsymbol{g}(\boldsymbol{w})\| \leq \|\widehat{\boldsymbol{\theta}}(\boldsymbol{w}) - \widehat{\boldsymbol{\theta}}(\widetilde{\boldsymbol{w}})\| + \|\boldsymbol{g}(\boldsymbol{w}) - \boldsymbol{g}(\widetilde{\boldsymbol{w}})\| + \|\widehat{\boldsymbol{\theta}}(\widetilde{\boldsymbol{w}}) - \boldsymbol{g}(\widetilde{\boldsymbol{w}})\|. \tag{10}$$

The goal is to be able to take the supremum over $\boldsymbol{w} \in \mathcal{W}$. We bound one term at a time. The first term can be bounded, for any $\boldsymbol{w} \in \mathcal{W}$, by (9) just proven. The second term can be bounded by

$$\|\boldsymbol{g}(\boldsymbol{w}) - \boldsymbol{g}(\widetilde{\boldsymbol{w}})\| \leq \lambda \|\boldsymbol{w} - \widetilde{\boldsymbol{w}}\| \tag{11}$$

which follows immediately from (A2). Finally, for the third term, fixing any $\boldsymbol{w} \in \mathcal{W}$, $\widetilde{\boldsymbol{w}} = \widetilde{\boldsymbol{w}}(\boldsymbol{w}) \in \{\widetilde{\boldsymbol{w}}_1, \ldots, \widetilde{\boldsymbol{w}}_{N_\epsilon}\}$ is also fixed, and can be bounded on the δ event $\mathcal{E}(\widetilde{\boldsymbol{w}})$ just defined. The important fact is that

$$\sup_{\boldsymbol{w} \in \mathcal{W}} \left\| \widehat{\boldsymbol{\theta}}(\widetilde{\boldsymbol{w}}(\boldsymbol{w})) - \boldsymbol{g}(\widetilde{\boldsymbol{w}}(\boldsymbol{w})) \right\| = \max_{k \in [N_\epsilon]} \left\| \widehat{\boldsymbol{\theta}}(\widetilde{\boldsymbol{w}}_k) - \boldsymbol{g}(\widetilde{\boldsymbol{w}}_k) \right\|.$$

We construct a "good event" naturally as the event in which the bad event $\mathcal{E}(\cdot)$ holds for no center on our ϵ-net, namely

$$\mathcal{E}_+ = \left(\bigcap_{k \in [N_\epsilon]} \mathcal{E}(\widetilde{\boldsymbol{w}}_k) \right)^c.$$

Taking a union bound, we can say that with probability no less than $1 - \delta$, for all $\boldsymbol{w} \in \mathcal{W}$, we have

$$\|\widehat{\boldsymbol{\theta}}(\widetilde{\boldsymbol{w}}(\boldsymbol{w})) - \boldsymbol{g}(\widetilde{\boldsymbol{w}}(\boldsymbol{w}))\| \leq 2\sqrt{\frac{dV \log(2dN_\epsilon \delta^{-1})}{n}} c_0. \tag{12}$$

Taking the three new bounds together, we have with probability no less than $1 - \delta$ that

$$\sup_{\boldsymbol{w} \in \mathcal{W}} \|\widehat{\boldsymbol{\theta}}(\boldsymbol{w}) - \boldsymbol{g}(\boldsymbol{w})\| \leq \lambda\epsilon(\sqrt{d} + 1) + 2\sqrt{\frac{dV \log(2dN_\epsilon \delta^{-1})}{n}} c_0.$$

Setting $\epsilon = 1/\sqrt{n}$ we have

$$\sup_{\boldsymbol{w} \in \mathcal{W}} \|\widehat{\boldsymbol{\theta}}(\boldsymbol{w}) - \boldsymbol{g}(\boldsymbol{w})\| \leq \frac{\lambda(\sqrt{d} + 1)}{\sqrt{n}} + 2c_0\sqrt{\frac{dV(\log(2d\delta^{-1}) + d\log(3\Delta\sqrt{n}/2))}{n}}.$$

Since every step of Algorithm 1 (main text), with orthogonal projection if required, has $\widehat{\boldsymbol{w}}_{(t)} \in \mathcal{W}$, the desired result follows from this uniform confidence interval. $\square$

[1] This is a basic property of covering numbers for compact subsets of Euclidean space [4].

Proof of Lemma 4 (main text). Given $\widehat{\boldsymbol{w}}_{(t)}$, running the approximate update (2, main text), we have

$$\begin{aligned}\|\widehat{\boldsymbol{w}}_{(t+1)} - \boldsymbol{w}^*\| &= \|\widehat{\boldsymbol{w}}_{(t)} - \alpha\widehat{\boldsymbol{g}}(\widehat{\boldsymbol{w}}_{(t)}) - \boldsymbol{w}^*\| \\ &\leq \|\widehat{\boldsymbol{w}}_{(t)} - \alpha\boldsymbol{g}(\widehat{\boldsymbol{w}}_{(t)}) - \boldsymbol{w}^*\| + \alpha\|\widehat{\boldsymbol{g}}(\widehat{\boldsymbol{w}}_{(t)}) - \boldsymbol{g}(\widehat{\boldsymbol{w}}_{(t)})\|.\end{aligned}$$

The first term looks at the distance from the target given an optimal update, using $\boldsymbol{g}$. Using the κ-strong convexity of R, via Nesterov [5, Thm. 2.1.15] it follows that

$$\|\widehat{\boldsymbol{w}}_{(t)} - \alpha\boldsymbol{g}(\widehat{\boldsymbol{w}}_{(t)}) - \boldsymbol{w}^*\|^2 \leq \left(1 - \frac{2\alpha\kappa\lambda}{\kappa+\lambda}\right)\|\widehat{\boldsymbol{w}}_{(t)} - \boldsymbol{w}^*\|^2.$$

Writing $\beta := 2\kappa\lambda/(\kappa+\lambda)$, the coefficient becomes $(1-\alpha\beta)$.

To control the second term simply requires unfolding the recursion. By hypothesis, we can leverage (6, main text) to bound the statistical estimation error by ε for every step, all on the same $1-\delta$ "good event." For notational ease, write $a := \sqrt{1-\alpha\beta}$. On the good event, we have

$$\begin{aligned}\|\widehat{\boldsymbol{w}}_{(t+1)} - \boldsymbol{w}^*\| &\leq a^{t+1}\|\widehat{\boldsymbol{w}}_{(0)} - \boldsymbol{w}^*\| + \alpha\varepsilon\left(1 + a + a^2 + \cdots + a^t\right) \\ &= a^{t+1}\|\widehat{\boldsymbol{w}}_{(0)} - \boldsymbol{w}^*\| + \alpha\varepsilon\frac{(1-a^{t+1})}{1-a}.\end{aligned}$$

To clean up the second summand,

$$\begin{aligned}\alpha\varepsilon\frac{(1-a^{t+1})}{1-a} &\leq \frac{\alpha\varepsilon(1+a)}{(1-a)(1+a)} \\ &= \frac{\alpha\varepsilon(1+\sqrt{1-\alpha\beta})}{\alpha\beta} \\ &\leq \frac{2\varepsilon}{\beta}.\end{aligned}$$

Taking this to the original inequality yields the desired result. □

Proof of Theorem 5 (main text). Using strong convexity and (2), we have that

$$\begin{aligned}R(\widehat{\boldsymbol{w}}_{(T)}) - R^* &\leq \frac{\lambda}{2}\|\widehat{\boldsymbol{w}}_{(T)} - \boldsymbol{w}^*\|^2 \\ &\leq \lambda(1-\alpha\beta)^T D_0^2 + \frac{4\lambda\varepsilon^2}{\beta^2}.\end{aligned}$$

The latter inequality holds by direct application of Lemma 4 (main text), followed by the elementary fact $(a+b)^2 \leq 2(a^2+b^2)$. The particular value of ε under which Lemma 4 (main text) is valid (i.e., under which (6, main text) holds) is given by Lemma 3 (main text). Filling in ε with this concrete setting yields the desired result. □

Proof of Lemma 8 (main text). As in the result statement, we write

$$\Sigma_{(t)} := \mathbf{E}_\mu\left(l'(\widehat{\boldsymbol{w}}_{(t)};\boldsymbol{z}) - \boldsymbol{g}(\widehat{\boldsymbol{w}}_{(t)})\right)\left(l'(\widehat{\boldsymbol{w}}_{(t)};\boldsymbol{z}) - \boldsymbol{g}(\widehat{\boldsymbol{w}}_{(t)})\right)^T, \quad \boldsymbol{w} \in \mathcal{W}.$$

Running this modified version of Algorithm 1 (main text), we are minimizing the bound in Lemma 1 (main text) as a function of scale s_j, $j \in [d]$, which immediately implies that the estimates $\widehat{\boldsymbol{\theta}}_{(t)} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_d)$ at each step t satisfy

$$|\widehat{\theta}_j - g_j(\widehat{\boldsymbol{w}})| > 4\left(\frac{C\,\mathrm{var}_\mu\, l'_j(\widehat{\boldsymbol{w}}_{(t)};\boldsymbol{z})\log(2\delta^{-1})}{n}\right)^{1/2} \tag{13}$$

with probability no greater than δ. For clean notation, let us also denote

$$A := 4\left(\frac{C\log(2\delta^{-1})}{n}\right)^{1/2}, \quad \varepsilon^{*} := A\sqrt{\operatorname{trace}(\Sigma_{(t)})}.$$

For the vector estimates then, we have

$$\begin{aligned}
&\mathbf{P}\left\{\|\widehat{\boldsymbol{\theta}}_{(t)} - \boldsymbol{g}(\widehat{\boldsymbol{w}}_{(t)})\| > \varepsilon^{*}\right\}\\
&\quad = \mathbf{P}\left\{\sum_{j=1}^{d}\frac{(\widehat{\theta}_j - g_j(\widehat{\boldsymbol{w}}_{(t)}))^2}{A^2} > \operatorname{trace}(\Sigma_{(t)})\right\}\\
&\quad = \mathbf{P}\left\{\sum_{j=1}^{d}\left(\frac{(\widehat{\theta}_j - g_j(\widehat{\boldsymbol{w}}_{(t)}))^2}{A^2} - \operatorname{var}_{\mu} l_j'(\widehat{\boldsymbol{w}}_{(t)};\boldsymbol{z})\right) > 0\right\}\\
&\quad \leq \mathbf{P}\bigcup_{j=1}^{d}\left\{\frac{(\widehat{\theta}_j - g_j(\widehat{\boldsymbol{w}}_{(t)}))^2}{A^2} > \operatorname{var}_{\mu} l_j'(\widehat{\boldsymbol{w}}_{(t)};\boldsymbol{z})\right\}\\
&\quad \leq d\delta.
\end{aligned}$$

The first inequality uses a union bound, and the second inequality follows from (13). Plugging in A and taking confidence δ/d implies the desired result. □

Proof of Theorem 9 (main text). From Lemma 8 (main text), the estimation error has exponential tails, as follows. Writing

$$A_1 := 2d, \quad A_2 := 4\left(\frac{C\operatorname{trace}(\Sigma_{(t)})}{n}\right)^{1/2},$$

for each iteration t we have

$$\mathbf{P}\{\|\widehat{\boldsymbol{\theta}}_{(t)} - \boldsymbol{g}(\widehat{\boldsymbol{w}}_{(t)})\| > \varepsilon\} \leq A_1\exp\left(-\left(\frac{\varepsilon}{A_2}\right)^2\right).$$

Controlling moments using exponential tails can be done using a fairly standard argument. For random variable $X \in \mathcal{L}_p$ for $p \geq 1$, we have the classic inequality

$$\mathbf{E}\,|X|^p = \int_0^{\infty}\mathbf{P}\{|X|^p > t\}\,dt$$

as a starting point. Setting $X = \|\widehat{\boldsymbol{\theta}}_{(t)} - \boldsymbol{g}(\widehat{\boldsymbol{w}}_{(t)})\| \geq 0$, and using substitution of variables twice, we have

$$\begin{aligned}
\mathbf{E}\,|X|^p &= \int_0^{\infty}\mathbf{P}\{X > t^{1/p}\}\,dt\\
&= \int_0^{\infty}\mathbf{P}\{X > t\}pt^{p-1}\,dt\\
&\leq A_1 p\int_0^{\infty}\exp\left(-(t/A_2)^2\right)t^{p-1}\,dt\\
&= \frac{A_1A_2^p p}{2}\int_0^{\infty}\exp(-t)t^{p/2-1}\,dt.
\end{aligned}$$

The last integral on the right-hand side, written $\Gamma(p/2)$, is the usual Gamma function of Euler evaluated at $p/2$. Setting $p = 2$, we have $\Gamma(1) = 0! = 1$, and plugging in the values of A_1 and A_2 yields the desired result. □

A.3 Computational methods

Here we discuss precisely how to compute the implicitly-defined M-estimates of (3, main text) and (5, main text). Assuming $s > 0$ and real-valued observations $x_1, \ldots, x_n$, we first look at the program

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \rho_s (x_i - \theta)$$

assuming ρ is as specified in Definition 1, with $\psi = \rho'$. Write $\widehat{\theta}$ for this unique minimum, and note that it satisfies

$$\frac{s}{n} \sum_{i=1}^{n} \psi_s \left(x_i - \widehat{\theta} \right) = 0.$$

Indeed, by monotonicity of ψ, this $\widehat{\theta}$ can be found via ρ minimization or root-finding. The latter yields standard fixed-point iterative updates, such as

$$\widehat{\theta}_{(k+1)} = \widehat{\theta}_{(k)} + \frac{s}{n} \sum_{i=1}^{n} \psi_s \left(x_i - \widehat{\theta}_{(k)} \right).$$

Note the right-hand side has a fixed point at the desired value. In our routines, we use the Gudermannian function

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

which can be readily confirmed to satisfy all requirements of Definition 1.

For the dispersion estimate to be used in re-scaling, we introduce function χ, which is even, non-decreasing on $\mathbb{R}_+$, and satisfies

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

In practice, we take dispersion estimate $\widehat{\sigma} > 0$ as any value satisfying

$$\frac{1}{n} \sum_{i=1}^{n} \chi \left(\frac{x_i - \gamma}{\widehat{\sigma}} \right) = 0$$

where $\gamma = n^{-1} \sum_{i=1}^{n} x_i$, computed by the iterative procedure

$$\widehat{\sigma}_{(k+1)} = \widehat{\sigma}_{(k)} \left(1 - \frac{1}{\chi(0)n} \sum_{i=1}^{n} \chi \left(\frac{x_i - \gamma}{\widehat{\sigma}_{(k)}} \right) \right)^{1/2}$$

which has the desired fixed point, as in the location case. Our routines use the quadratic Geman-type χ, defined

$$\chi(u) := \frac{u^2}{1 + u^2} - c$$

with parameter $c > 0$, noting $\chi(0) = -c$. Writing the first term as χ_0 so $\chi(u) = \chi_0(u) - c$, we set $c = \mathbf{E}\, \chi_0(x)$ under $x \sim N(0, 1)$. Computed via numerical integration, this is $c \approx 0.34$.

References

[1] Ash, R. B. and Doleans-Dade, C. (2000). *Probability and Measure Theory.* Academic Press.

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

[3] 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.

[4] Kolmogorov, A. N. (1993). ε-entropy and ε-capacity of sets in functional spaces. In Shiryayev, A. N., editor, *Selected Works of A. N. Kolmogorov, Volume III: Information Theory and the Theory of Algorithms*, pages 86–170. Springer.

[5] Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course.* Springer.

[6] Talvila, E. (2001). Necessary and sufficient conditions for differentiating under the integral sign. *American Mathematical Monthly*, 108(6):544–548.