Supplemental material: Robust descent using smoothed multiplicative noise

Matthew J. Holland
Osaka University

Assumptions and examples

No algorithm can achieve arbitrarily good performance across all possible distributions [11]. In order to obtain meaningful theoretical results, we must place conditions on the underlying distribution, as well as the model and objective functions used. We give concrete examples to illustrate that these assumptions are reasonable, and that they include scenarios that allow for both sub-Gaussian and heavy-tailed data.

A0. $\mathcal{W}$ is a closed, convex subset of $\mathbb{R}^d$, with diameter $\Delta \coloneqq \sup\{\|\boldsymbol{u}-\boldsymbol{v}\| : \boldsymbol{u},\boldsymbol{v} \in \mathcal{W}\} < \infty$.

A1. Loss function $l(\cdot\,;\boldsymbol{z})$ is λ-smooth on $\mathcal{W}$.

A2. $R(\cdot)$ is λ-smooth, and continuously differentiable on $\mathcal{W}$.

A3. There exists $\boldsymbol{w}^{*} \in \mathcal{W}$ at which $\boldsymbol{g}(\boldsymbol{w}^{*}) = 0$.

A4. $R(\cdot)$ is κ-strongly convex on $\mathcal{W}$.

A5. There exists $v < \infty$ such that $\mathbf{E}_{\mu}(l_j'(\boldsymbol{w};\boldsymbol{z}))^2 \leq v$, for all $\boldsymbol{w} \in \mathcal{W}$, $j \in [d]$.

Of these assumptions, assuredly A0 is simplest: any ball (here in the ℓ_2 norm) with finite radius will suffice, though far more exotic examples are assuredly possible. The remaining assumptions require some checking, but hold under very weak assumptions on the underlying distribution, as the following examples show.

Example 1 (Concrete example of assumption A1). Consider the linear regression model $y = \langle \boldsymbol{w}^{*}, \boldsymbol{x}\rangle + \eta$, where $\boldsymbol{x}$ is almost surely bounded (say $\mathbf{P}\{\|\boldsymbol{x}\| \leq c\} = 1$), but the noise η can have any distribution we desire. Consider the squared loss $l(\boldsymbol{w};\boldsymbol{z}) = (\langle \boldsymbol{w}, \boldsymbol{x}\rangle - y)^2$, and observe that for any $\boldsymbol{w}, \boldsymbol{w}' \in \mathcal{W}$, we have

$$\boldsymbol{l}'(\boldsymbol{w};\boldsymbol{z}) - \boldsymbol{l}'(\boldsymbol{w}';\boldsymbol{z}) = 2(\langle \boldsymbol{w}-\boldsymbol{w}', \boldsymbol{x}\rangle)\boldsymbol{x}$$

and thus

$$\|\boldsymbol{l}'(\boldsymbol{w};\boldsymbol{z}) - \boldsymbol{l}'(\boldsymbol{w}';\boldsymbol{z})\| \leq 2\|\boldsymbol{x}\|^2\|\boldsymbol{w}-\boldsymbol{w}'\| \leq 2c^2\|\boldsymbol{w}-\boldsymbol{w}'\|.$$

Thus we have smoothness with $\lambda = 2c^2$, satisfying A1.

Example 2 (Concrete example of assumptions A2 and A3). Consider a similar setup as in Example 1, but instead of requiring $\boldsymbol{x}$ to be bounded, weaken the assumption to $\mathbf{E}\,\|\boldsymbol{x}\|^2 < \infty$.

Since taking the derivative under the integral we have $\boldsymbol{g}(\boldsymbol{w}) = \mathbf{E}\,\boldsymbol{l}'(\boldsymbol{w};\boldsymbol{z}) = 2(\langle \boldsymbol{w}-\boldsymbol{w}^{*},\boldsymbol{x}\rangle - \eta)\boldsymbol{x}$. Clearly, $\boldsymbol{g}(\boldsymbol{w}^{*}) = 0$, satisfying A3. Furthermore, it follows that

$$\begin{aligned}\boldsymbol{g}(\boldsymbol{w}) - \boldsymbol{g}(\boldsymbol{w}') &= \mathbf{E}\left(\boldsymbol{l}'(\boldsymbol{w};\boldsymbol{z}) - \boldsymbol{l}'(\boldsymbol{w}';\boldsymbol{z})\right)\\ &= 2\,\mathbf{E}(\langle \boldsymbol{w}-\boldsymbol{w}',\boldsymbol{x}\rangle)\boldsymbol{x}.\end{aligned}$$

We thus have

$$\|\boldsymbol{g}(\boldsymbol{w}) - \boldsymbol{g}(\boldsymbol{w}')\| \leq 2\,\mathbf{E}\,\|\boldsymbol{x}\|^{2}\|\boldsymbol{w}-\boldsymbol{w}'\| \leq 2c^{2}\|\boldsymbol{w}-\boldsymbol{w}'\|,$$

meaning smoothness of the risk holds with $\lambda = 2\,\mathbf{E}\,\|\boldsymbol{x}\|^{2}$, satisfying A2.

Example 3 (Concrete example of assumption A5). Again consider a setting similar to Examples 1–2, but with added assumptions that $\mathbf{E}\,\boldsymbol{x} = 0$, that the noise η and input $\boldsymbol{x}$ are independent, and that the components of $\boldsymbol{x} = (x_1,\ldots,x_d)$ are independent of each other. Some straightforward algebra shows that

$$\begin{aligned}\mathbf{E}(l_j'(\boldsymbol{w};\boldsymbol{z}))^{2} &= 4\left(\mathbf{E}\,x_j^{2}\langle \boldsymbol{w}-\boldsymbol{w}^{*},\boldsymbol{x}\rangle^{2} + \mathbf{E}\,\eta^{2}\,\mathbf{E}\,x_j^{2}\right)\\ &\leq 4\left(\|\boldsymbol{w}-\boldsymbol{w}^{*}\|^{2}\,\mathbf{E}\,x_j^{2}\|\boldsymbol{x}\|^{2} + \mathbf{E}\,\eta^{2}\,\mathbf{E}\,x_j^{2}\right).\end{aligned}$$

It follows that as long as the noise η has finite variance ($\mathbf{E}\,\eta^{2} < \infty$), and all inputs have finite fourth moments $\mathbf{E}\,x_j^{4} < \infty$, then using assumption A0, we get

$$\mathbf{E}(l_j'(\boldsymbol{w};\boldsymbol{z}))^{2} \leq 4\left(\Delta^{2}\,\mathbf{E}\,x_j^{2}\|\boldsymbol{x}\|^{2} + \mathbf{E}\,\eta^{2}\,\mathbf{E}\,x_j^{2}\right) < \infty.$$

This holds for all $\boldsymbol{w} \in \mathcal{W}$, satisfying A5.

Example 4 (Concrete example of assumption A4). Consider the same setup as Example 3. Since $\mathbf{E}\,x_j\eta = (\mathbf{E}\,x_j)(\mathbf{E}\,\eta) = 0$ for each $j \in [d]$, it follows that the risk induced by the squared loss under this model takes a convenient quadratic form,

$$R(\boldsymbol{w}) = \mathbf{E}\,l(\boldsymbol{w};\boldsymbol{z}) = (\boldsymbol{w}-\boldsymbol{w}^{*})^{T}A(\boldsymbol{w}-\boldsymbol{w}^{*}) + b^{2},$$

with $A = \mathbf{E}\,\boldsymbol{x}\boldsymbol{x}^{T}$ and $b^{2} = \mathbf{E}\,\eta^{2}$. For concreteness, say all the components of $\boldsymbol{x}$ have variance $\mathbf{E}\,x_j^{2} = \sigma^{2}$, recalling that $\mathbf{E}\,x_j = 0$ by assumption. Then the Hessian matrix of $R(\cdot)$ is $R''(\boldsymbol{w}) = \mathbf{E}\,\boldsymbol{x}\boldsymbol{x}^{T} = \sigma^{2}I_d$, for all $\boldsymbol{w} \in \mathcal{W}$. For any $\boldsymbol{w},\boldsymbol{w}' \in \mathcal{W}$, taking an exact Taylor expansion, we have that

$$R(\boldsymbol{w}) = R(\boldsymbol{w}') + \langle \boldsymbol{g}(\boldsymbol{w}), \boldsymbol{w}-\boldsymbol{w}'\rangle + \frac{1}{2}\langle \boldsymbol{w}-\boldsymbol{w}', R''(\boldsymbol{u})(\boldsymbol{w}-\boldsymbol{w}')\rangle$$

for some appropriate $\boldsymbol{u}$ on the line segment between $\boldsymbol{w}$ and $\boldsymbol{w}'$. Since the Hessian is positive definite with factor σ^{2}, the last term on the right-hand side can be no smaller than $\|\boldsymbol{w}-\boldsymbol{w}'\|^{2}\sigma^{2}/2$. This implies a lower bound,

$$R(\boldsymbol{w}) \geq R(\boldsymbol{w}') + \langle \boldsymbol{g}(\boldsymbol{w}), \boldsymbol{w}-\boldsymbol{w}'\rangle + \frac{\sigma^{2}}{2}\|\boldsymbol{w}-\boldsymbol{w}'\|^{2}.$$

The exact same inequality holds for any choice of $\boldsymbol{w}$ and $\boldsymbol{w}'$. This is precisely the definition of strong convexity of $R(\cdot)$ given in (5), with convexity parameter $\kappa = \sigma^{2}$, satisfying A4.

Experimental setup

For controlled tests (see section 4.1, main text)

Noisy convex minimization We begin with a "noisy convex risk minimization" task, designed as follows. The risk function itself takes a quadratic form, as $R(\boldsymbol{w}) = \langle \Sigma \boldsymbol{w}, \boldsymbol{w} \rangle / 2 + \langle \boldsymbol{w}, \boldsymbol{u} \rangle + c$, where $\Sigma \in \mathbb{R}^{d \times d}$, $\boldsymbol{u} \in \mathbb{R}^{d}$, and $c \in \mathbb{R}$ are constants set in advance. The learning task is to find a minimizer of $R(\cdot)$, without direct access to R, rather only access to n random function data $r_1, \ldots, r_n$, with $r : \mathbb{R}^d \to \mathbb{R}$ mapping from parameter space to a numerical penalty. This data is generated independently from a common distribution, and are centered at the true risk, namely $\mathbf{E}\, r(\boldsymbol{w}) = R(\boldsymbol{w})$ for all $\boldsymbol{w} \in \mathbb{R}^d$. More concretely, we generate $r_i(\boldsymbol{w}) = (\langle \boldsymbol{w}^* - \boldsymbol{w}, \boldsymbol{x}_i \rangle + \epsilon_i)^2/2$, $i \in [n]$, with $\boldsymbol{x}$ and ϵ independent. The true minimum is denoted $\boldsymbol{w}^*$, and $\Sigma = \mathbf{E}\, \boldsymbol{x}\boldsymbol{x}^T$. The inputs $\boldsymbol{x}$ are set to have a d-dimensional Gaussian distribution with all components uncorrelated. This means that Σ is positive definite, and R is strongly convex.

We make use of three metrics for evaluating performance here: average excess empirical risk (averaging of $r_1, \ldots, r_n$), average excess risk (computed using true R), and variance of the risk. The latter two are computed by averaging over trials; each trial means a new independent random sample. In all tests, we conduct 250 trials.

Regarding methods tested, we run three representative procedures. First is the idealized gradient descent procedure (1, main text), denoted `oracle`, which is possible here since R is designed by us. Second, as a *de facto* standard for most machine learning algorithms, we use ERM-GD, written `erm`. Here the update direction is simply the sample mean of the loss gradient. Finally, we compare our Algorithm 1 (main text), written `rgdmult`, against these two procedures. Variance bounds $\boldsymbol{v}_{(t)}$ are computed using the simplest possible procedure, namely the empirical mean of the second moments of $\boldsymbol{l}'(\widehat{\boldsymbol{w}}_{(t)}; \boldsymbol{z})$, divided by two.

Our first inquiry is a basic proof of concept: are there natural problem settings under which using `rgdmult` over ERM-GD is advantageous? How does this procedure perform when ERM-GD is known to be effectively optimal? Under Gaussian noise, ERM-GD is effectively optimal [7, Appendix C]. As a baseline, we start with Gaussian noise (mean 0, standard deviation 20), and then consider centered log-Normal noise (log-location 0, log-scale 1.75) as a representative example of asymmetric, heavy-tailed data. Performance results are given in Figure 2.

Comparison with robust loss minimizer Our next inquiry in the main text is a comparison with the robust loss minimizer approach of Brownlees et al. [1], which chiefly considered theoretical analysis of a robust learning procedure that minimizes a robust *objective*, in contrast to our use of a robust update direction. Our proposed procedure enjoys essentially the same theoretical guarantees, and we have claimed that it is more practical. Here we attempt to verify this claim empirically. Denote the method of Brownlees et al. [1] by `bjl`. To implement their approach, which does not specify any particular algorithmic technique, we implement `bjl` using the non-linear conjugate gradient method of Polak and Ribière [10]. This can be found as part of the the `optimize` module of the SciPy scientific computation library, called `fmin_cg`, with default parameter settings. We believe that using this standard first-order solver makes for a fair comparison between `bjl` and our Algorithm 1 (main text), again denoted `rgdmult`, and again with variance bound $\boldsymbol{v}_{(t)}$ set to the empirical second moments of $\boldsymbol{l}'(\widehat{\boldsymbol{w}}_{(t)}; \boldsymbol{z})$, multiplied by $1/\sqrt{d}$. For our routine, we have fixed the number of iterations to be $T = 30$ for all settings. We compute the time required for computation using the Python `time` module. Multiple independent trials of each learning task (analogous to those previous) are carried out, with the median time taken over trials (for each d setting) used as the final time record. We consider settings of $d = 2, 4, 8, 16, 32, 64$. These times along with performance results are given in Figure 3.

Regression application Finally, we look at a more general regression task, under a variety of data distributions. We then compare Algorithm 1 with well-known procedures specialized to regression, both classical and recent. In each experimental condition, and for each trial, we generate n observations of the form $y_i = \boldsymbol{x}_i^T \boldsymbol{w}^* + \epsilon_i, i \in [n]$ for training. Each condition is defined by the setting of (n, d) and μ. Throughout, we have inputs $\boldsymbol{x}$ which are generated from a d-dimensional Gaussian distribution, with each coordinate independent of the others. As such, to set μ requires setting the distribution of the noise, ϵ. We consider several families of distributions, each with 15 distinct parameter settings, or "noise levels." These settings are carried out such that the standard deviation of ϵ increases over the range 0.3–20.0, in a roughly linear fashion as we increase from level 1 (lowest) to 15 (highest).

A range of signal/noise ratios can be captured by controlling the norm of the vector $\boldsymbol{w}^* \in \mathbb{R}^d$ determining the model. For each trial, we generate $\boldsymbol{w}^*$ randomly as follows. Considering the sequence $w_k := \pi/4 + (-1)^{k-1}(k-1)\pi/8, k = 1, 2, \ldots$, sample $i_1, \ldots, i_d \in [d_0]$ uniformly, with $d_0 = 500$. The underlying vector is then set as $\boldsymbol{w}^* = (w_{i_1}, \ldots, w_{i_d})$. The signal to noise ratio $\mathrm{SN}_\mu = \|\boldsymbol{w}^*\|_2^2 / \operatorname{var}_\mu(\epsilon)$ then varies over the range $0.2 \leq SN_\mu \leq 1460.6$. Here we consider four noise families: log-logistic (denoted `llog` in figures), log-Normal (`lnorm`), Normal (`norm`), and symmetric triangular (`tri_s`).

Here we do not compute the risk R exactly, but rather use off-sample prediction error as the key metric for evaluating performance. This is computed as excess root mean squared error (RMSE) computed on an independent testing set. Performance is averaged over independent trials. For each condition and trial, a test set of m independent observations is generated identically to the n-sized training set that precedes testing. All competing methods use common samples for training and testing, for each condition and trial. In the kth trial, each algorithm outputs an estimate $\widehat{\boldsymbol{w}}(h)$. Using RMSE to approximate the ℓ_2-risk, compute $e_k(\widehat{\boldsymbol{w}}) := (m^{-1} \sum_{i=1}^m (\widehat{\boldsymbol{w}}^T \boldsymbol{x}_{k,i} - y_{k,i})^2)^{1/2}$, outputting prediction error as the excess error $e_k(\widehat{\boldsymbol{w}}(k)) - e_k(\boldsymbol{w}^*(k))$, averaged over K trials. In all experiments, we have $K = 250$, $m = 1000$.

We consider several methods against which we compare the proposed Algorithm 1 (main text). As classical choices, we have ordinary least squares (ERM under the squared error, `ols`) and least absolute deviations (ERM under absolute error, `lad`). For more recent methods, as described in section 1, we consider robust regression routines as given by Minsker [8] (`geomed`) and Hsu and Sabato [4] (`hs`). In the former, we partition the data, obtaining the `ols` solution on each subset, and these candidates are aggregated using the geometric median in the ℓ_2 norm [12]. The number of partitions is set to $\max\{2, \lfloor n/(2d) \rfloor\}$. In the latter, we used source code published online by the authors. To compare our Algorithm 1 (main text) with these routines, we initialize `rgdmult` to the analytical `ols` solution, with step size $\alpha_{(t)} = 0.01$ for all iterations, and $\delta = 0.005$. Variance bounds $\boldsymbol{v}_{(t)}$ are set to the empirical second moments of $\boldsymbol{l}'(\widehat{\boldsymbol{w}}_{(t)}, \boldsymbol{z})$, divided by 2. In total, the number of iterations is constrained by a fixed budget: we allow for $40n$ gradient evaluations in total. Representative results are provided in Figure 4.

For application to real-world benchmarks (see section 4.2, main text)

All methods use a common model, here multi-class logistic regression. If the number of classes is C, and we have F input features, then the dimension of the model will be $d = (C-1)F$. A basic property of this model is that the loss function is convex in the parameters, with gradients that exist, thus placing the model firmly within our realm of interest. Furthermore, for all of these tests we shall add a squared ℓ_2-norm regularization term $a\|\boldsymbol{w}\|^2$ to the loss, where a varies depending on the dataset. Once again, each algorithm is given a fixed budget, this time of $20n$, where n is the size of the training set available, which again depends on the dataset (details below).

Here we give results for two well-known data sets used for benchmarking: the forest cover type dataset from the UCI repository,[1] and the protein homology dataset used in a previous KDD Cup.[2] For each dataset, we execute 10 independent trials, with training/testing subsets randomly sampled without replacement as is described shortly. For all datasets, we normalize input features to the unit interval $[0, 1]$ in a per-feature fashion. For the cover type dataset, we consider binary classification of the second type against all other types. With $C = 2$ and $F = 54$, we have $d = 54$ and $a = 0.001$, with a training subset of size $n = 4d$. The protein homology dataset has highly unbalanced labels, with only 1296 positive labels our of over 145,000 examples. We balance out training and testing data, randomly selecting 296 positive examples and the same number of negative examples, yielding a test set of 592 points. As for the training set size, we use all positive examples not used for testing (1000 points each time), plus a random selection of 1000 negatively labeled examples, so $n = 2000$. With $C = 2$ and $F = 74$, the dimension is $d = 74$, and $a = 0.001$. In all settings, initialization is done uniformly over the interval $[-0.05, 0.05]$.

We investigate the utility of a random mini-batch version of Algorithm 1 (main text) here. We try mini-batch sizes of 10 and 20. Variance bounds $\boldsymbol{v}_{(t)}$ are set to k times the empirical mean of the second moments of $\boldsymbol{l}'(\widehat{\boldsymbol{w}}_{(t)}, \boldsymbol{z})$, with k ranging over $\{1/10, 1/5, 1/2, 1, 5, 25, 125, 625\}$. Furthermore, for the high-dimensional datasets, we consider a mini-batch in terms of random selection of which parameters to robustly update. At each iteration, we randomly choose $\min\{100, d\}$ indices, running Algorithm 1 (main text) for the resulting sub-vector, and the sample mean for the remaining coordinates. We compare our proposed algorithm with stochastic gradient descent (SGD), and stochastic variance-reduced gradient descent (SVRG) proposed by Johnson and Zhang [5]. For each method, pre-fixed step sizes ranging over $\{0.0001, 0.001, 0.01, 0.05, 0.10, 0.15, 0.20\}$ are tested. SGD has mini-batches of size 1, just as the SVRG inner loop. The inner loop of SVRG has $n/2$ iterations, and all methods continue running until the fixed budget of gradient evaluations is spent. Representative results are given in Figure 5.

A Technical appendix

A.1 Preliminaries

Consider two probability measures P and Q on measurable space $(\mathcal{X}, \mathcal{A})$. We say that Q is absolutely continuous with respect to P, written $Q \ll P$, whenever $P(A) = 0$ implies $Q(A) = 0$ for all $A \in \mathcal{A}$. The Radon-Nikodym theorem guarantees that there exists a measurable function $g \geq 0$, such that

$$Q(A) = \int_A g\, dP, \qquad \text{for all } A \in \mathcal{A}.$$

Furthermore, this g is unique in the sense that if another f exists satisfying the above equality, we have $f = g$ almost everywhere $[P]$. It is common to call this function g the Radon-Nikodym derivative of Q with respect to P, written dQ/dP. The relative entropy, or Kullback-Leibler

[1] http://archive.ics.uci.edu/ml/datasets/Covertype

[2] http://www.kdd.org/kdd-cup/view/kdd-cup-2004/Tasks

divergence, between two probability measures P and Q on measurable space $(\mathcal{X}, \mathcal{A})$ is defined

$$\boldsymbol{K}(P;Q) \coloneqq \begin{cases} -\int \log\left(\frac{dQ}{dP}\right)\, dP, & \text{if } Q \ll P \\ +\infty, & \text{else.} \end{cases} \tag{1}$$

The key property of the ψ truncation function utilized by Catoni and Giulini [3], defined in (4, main text), is that for all $u \in \mathbb{R}$, we have

$$-\log\left(1 - u + \frac{u^2}{2}\right) \leq \psi(u) \leq \log\left(1 + u + \frac{u^2}{2}\right). \tag{2}$$

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable, convex, λ-smooth function.

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

$$\frac{1}{2\lambda}\|f'(\boldsymbol{u}) - f'(\boldsymbol{v})\|^2 \leq f(\boldsymbol{u}) - f(\boldsymbol{v}) - \langle f'(\boldsymbol{v}), \boldsymbol{u} - \boldsymbol{v}\rangle \tag{4}$$

for all $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$.

Terminology For a function $F : \mathcal{W} \to \mathbb{R}$, we say that F is λ-*Lipschitz* if, for all $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathcal{W}$ we have $|F(\boldsymbol{w}_1) - F(\boldsymbol{w}_2)| \leq \lambda\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|$. If F is differentiable, and the derivative $\boldsymbol{w} \mapsto F'(\boldsymbol{w})$ is λ-Lipschitz, then we say that F is λ-*smooth*.

If F is a convex function on convex set $\mathcal{W}$, then we say F is κ-*strongly convex* if for all $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathcal{W}$,

$$F(\boldsymbol{w}_1) - F(\boldsymbol{w}_2) \geq \langle F'(\boldsymbol{w}_2), \boldsymbol{w}_1 - \boldsymbol{w}_2\rangle + \frac{\kappa}{2}\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|^2. \tag{5}$$

This definition can be made for any valid norm space, but we shall be assuming $\mathcal{W} \subseteq \mathbb{R}^d$ throughout, and use the Euclidean norm. 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 $\mathcal{W}$.

A.2 Proofs of results in the main text

Proof of Lemma 2 (main text). Let $\mathcal{P}(\mathbb{R})$ denote all probability measures on $\mathbb{R}$, with an appropriate σ-field tacitly assumed. Consider any two measures $\nu, \nu_0 \in \mathcal{P}(\mathbb{R})$, and $h : \mathbb{R} \to \mathbb{R}$ a measurable function. By Catoni [2, p. 159–160], it is proved that a Legendre transform of the mapping $\nu \mapsto \boldsymbol{K}(\nu; \nu_0)$ takes the form of a cumulant generating function, namely

$$\sup_{\nu}\left(\int h(u)\, d\nu(u) - \boldsymbol{K}(\nu; \nu_0)\right) = \log \int \exp(h(u))\, d\nu_0(u), \tag{6}$$

where the supremum is taken over $\nu \in \mathcal{P}(\mathbb{R})$. This identity is a technical tool, and the choice of h and ν_0 are parameters that can be adjusted to fit the application.

In actually setting these parameters, we adapt the general argument of Catoni and Giulini [3] to our setting. Recalling the estimator (3, main text), we start with a quasi average of the points $x_1, \ldots, x_n$, modified by some data-sensitive additive noise, and passed through a truncation function. The expectation of this sum is then taken over the noise distribution. The ν in the definition of (3, main text) will correspond to ν here, and thus to reflect the whole estimator within (6), it makes sense to include the data-dependent sum in our choice of h. Note that the summands in the estimator definition

$$\psi\left(\frac{x_i + \epsilon_i x_i}{s}\right), \quad i \in [n]$$

depend on two random quantities, namely the data x_i, and the artificial noise ϵ_i (since $s > 0$ is assumed pre-fixed). Reflecting dependence on these quantities directly, we write

$$f(\epsilon, x) \coloneqq \psi\left(\frac{x + \epsilon x}{s}\right), \quad \epsilon, x \in \mathbb{R}.$$

Note that by definition of ψ in (4, main text), the function $f : \mathbb{R}^2 \to \mathbb{R}$ is measurable and bounded. With this cleaner notation, let us now set

$$h(\epsilon) = \sum_{i=1}^{n} f(\epsilon, x_i) - c(\epsilon)$$

where $c(\epsilon)$ is a term to be determined shortly. Plugging this in to (6) yields the following quantity:

$$\begin{aligned} B &\coloneqq \sup_{\nu} \left(\int h(\epsilon)\, d\nu(\epsilon) - \boldsymbol{K}(\nu; \nu_0) \right) \\ &= \log \int \exp\left(\sum_{i=1}^{n} f(\epsilon, x_i) - c(\epsilon) \right) d\nu(\epsilon). \end{aligned}$$

Taking the exponential of this B and then taking expectation with respect to the sample, we have

$$\begin{aligned} \mathbf{E}_{\mu} \exp(B) &= \mathbf{E}_{\mu} \int \left(\frac{\exp\left(\sum_{i=1}^{n} f(\epsilon, x_i)\right)}{\exp(c(\epsilon))} \right) \nu(\epsilon) \\ &= \int \left(\frac{\prod_{i=1}^{n} \mathbf{E}_{\mu} f(\epsilon, x_i)}{\exp(c(\epsilon))} \right) \nu(\epsilon). \end{aligned}$$

The first equality comes from simple log/exp manipulations, and the second equality from taking the integration over the sample inside the integration with respect to ν, valid via Fubini's theorem. It will be useful to have $\mathbf{E}_{\mu} \exp(B) \leq 1$. This can be achieved easily by setting

$$c(\epsilon) = n \log \mathbf{E}_{\mu} \exp(f(\epsilon, x)),$$

which yields

$$\mathbf{E}_{\mu} \exp(B) = \int \left(\frac{\prod_{i=1}^{n} \mathbf{E}_{\mu} \exp(f(\epsilon, x_i))}{\left(\mathbf{E}_{\mu} \exp(f(\epsilon, x_i))\right)^{n}} \right) \nu(\epsilon) = 1. \tag{7}$$

With this preparation done, we can start on the high-probability upper bound of interest:

$$\begin{aligned} \mathbf{P}\{B \geq \log(\delta^{-1})\} &= \mathbf{P}\{\exp(B) \geq 1/\delta\} \\ &= \mathbf{E}_{\mu} I\{\delta \exp(B) \geq 1\} \\ &\leq \mathbf{E}_{\mu} \delta \exp(B) \\ &= \delta. \end{aligned}$$

The inequality follows immediately since $\delta \exp(B) \geq 0$, and the final equality holds due to (7). Note that since our setting of $c(\epsilon)$ is such that $c(\cdot)$ is measurable (via measurability of f), the resulting $h(\cdot)$ is indeed measurable, as required. Of importance here is the fact that

$$\sup_{\nu} \left(\int h(\epsilon)\, d\nu(\epsilon) - \boldsymbol{K}(\nu;\nu_0) \right) \leq \log(\delta^{-1}) \tag{8}$$

with probability no less than $1-\delta$, noting that the event is uniform in ν. Using (6) once again, and dividing both sides by n, we have that with high probability, for any choice of ν, we can bound this generic empirical mean as follows:

$$\frac{1}{n}\sum_{i=1}^{n} \int f(\epsilon, x_i)\, d\nu(\epsilon) \leq \int \log \mathbf{E}_{\mu} \exp\left(f(\epsilon, x)\right)\, d\nu(\epsilon) + \frac{\boldsymbol{K}(\nu,\nu_0) + \log(\delta^{-1})}{n}. \tag{9}$$

Bridging the gap between these preparatory facts and the estimator of interest is now easy; since the noise terms $\epsilon_1, \ldots, \epsilon_n$ are assumed to be independent copies of $\epsilon \sim \nu$, it follows immediately that

$$\begin{aligned} \widehat{x} &= \frac{s}{n}\sum_{i=1}^{n} \int \left(\psi\left(\frac{x_i + \varepsilon_i x_i}{s} \right) \right) d\nu(\epsilon_i) \\ &= \frac{s}{n}\sum_{i=1}^{n} \int f(\epsilon, x_i)\, d\nu(\epsilon). \end{aligned}$$

That is to say, we have

$$\widehat{x} \leq s \int \log \mathbf{E}_{\mu} \exp\left(\psi\left(\frac{x(1+\epsilon)}{s} \right) \right) d\nu(\epsilon) + \frac{s}{n}\left(\boldsymbol{K}(\nu;\nu_0) + \log(\delta^{-1}) \right) \tag{10}$$

on the high-probability event, uniformly in choice of ν. Let us work step by step through each of the terms in the upper bound.

Starting with the first term, recall the definition of the truncation function ψ given in (4, main text), and in particular the logarithmic upper/lower bounds given in (2). These bounds will be convenient because it offers us polynomial bounds when passing ψ through $\exp(\cdot)$, which is precisely what occurs in (10) above. To get the first term in (10) in a more useful form, we can bound it as

$$\begin{aligned} \int \log \mathbf{E}_{\mu} \exp\left(\psi\left(\frac{x(1+\epsilon)}{s} \right) \right) d\nu(\epsilon) &\leq \int \log\left(1 + \frac{(1+\epsilon)\,\mathbf{E}_{\mu}\, x}{s} + \frac{(1+\epsilon)^2\,\mathbf{E}_{\mu}\, x^2}{2s^2} \right) d\nu(\epsilon) \\ &\leq \int \left(\frac{(1+\epsilon)\,\mathbf{E}_{\mu}\, x}{s} + \frac{(1+\epsilon)^2\,\mathbf{E}_{\mu}\, x^2}{2s^2} \right) d\nu(\epsilon) \\ &= \frac{\mathbf{E}_{\nu}(1+\epsilon)\,\mathbf{E}_{\mu}\, x}{s} + \frac{\mathbf{E}_{\nu}(1+\epsilon)^2\,\mathbf{E}_{\mu}\, x^2}{2s^2} \\ &= \frac{\mathbf{E}_{\mu}\, x}{s} + \frac{\mathbf{E}_{\mu}\, x^2}{2s^2}\left(\frac{1}{\beta} + 1 \right). \end{aligned}$$

The first inequality follows from (2), and the second from the fact that $\log(1+u) \leq u$ for all $u > -1$. As for the final equality, note that with $\epsilon \sim \nu = N(0, \beta^{-1})$, it follows immediately that

$$\mathbf{E}_{\nu}(1+\epsilon)^2 = \frac{1}{\beta} + (\mathbf{E}_{\nu}(1+\epsilon))^2 = \frac{1}{\beta} + 1.$$

Moving on to the second term, evaluating $\boldsymbol{K}(\nu;\nu_0)$ depends completely on how we define the pre-fixed ν_0. One approach is to set ν_0 such that the KL divergence is easily computed; for example, $\nu_0 = N(1,\beta^{-1})$. In this case, simple computations show that

$$\begin{aligned}\boldsymbol{K}(\nu;\nu_0) &= \int_{-\infty}^{\infty} \log\left(\exp\left(\frac{\beta(u-1)^2}{2} - \frac{\beta u^2}{2}\right)\right)\sqrt{\frac{\beta}{2\pi}}\exp\left(-\frac{\beta u^2}{2}\right)du \\ &= \int_{-\infty}^{\infty} \frac{(1-2u)\beta}{2}\sqrt{\frac{\beta}{2\pi}}\exp\left(-\frac{\beta u^2}{2}\right)du \\ &= \frac{\beta}{2}.\end{aligned}$$

With this computation done, an upper bound is complete, taking the form

$$\widehat{x} \leq \mathbf{E}_\mu x + \frac{\mathbf{E}_\mu x^2}{2s}\left(\frac{1}{\beta}+1\right) + \frac{s}{n}\left(\frac{\beta}{2} + \log(\delta^{-1})\right). \tag{11}$$

Optimizing this upper bound with respect to $s > 0$, we have

$$s^2 = \left(1+\frac{1}{\beta}\right)\frac{n\,\mathbf{E}_\mu x^2}{2}\left(\frac{\beta}{2}+\log(\delta^{-1})\right)^{-1}$$

and with respect to $\beta > 0$, we have

$$\beta^2 = \frac{n\,\mathbf{E}_\mu x^2}{s^2}. \tag{12}$$

Plugging this setting of β in to the setting of s yields

$$s^2 = \frac{n\,\mathbf{E}_\mu x^2}{2\log(\delta^{-1})}. \tag{13}$$

With this setting of s, the upper bound (11) can be cleaned up to the form

$$\widehat{x} \leq \mathbf{E}_\mu x + \sqrt{\frac{2\,\mathbf{E}_\mu x^2 \log(\delta^{-1})}{n}} + \sqrt{\frac{\mathbf{E}_\mu x^2}{n}}.$$

To get lower bounds on $\widehat{x} - \mathbf{E}_\mu x$, we can equivalently seek out upper bounds on $(-1)\widehat{x} + \mathbf{E}_\mu x$. This can be easily done via

$$-\widehat{x} \leq s\int \log \mathbf{E}_\mu \exp\left(-\psi\left(\frac{x(1+\epsilon)}{s}\right)\right)d\nu(\epsilon) + \frac{s}{n}\left(\boldsymbol{K}(\nu;\nu_0) + \log(\delta^{-1})\right). \tag{14}$$

Only the first term on the right-hand side is different from before. Note that by the lower bound of (2), we have

$$\log \mathbf{E}_\mu \exp\left(-\psi\left(\frac{x(1+\epsilon)}{s}\right)\right) \leq (-1)\frac{(1+\epsilon)\,\mathbf{E}_\mu x}{s} + \frac{(1+\epsilon)^2\,\mathbf{E}_\mu x^2}{2s^2}.$$

The rest plays out analogously to the upper bound, yielding

$$(-1)\widehat{x} \leq (-1)\,\mathbf{E}_\mu x + \sqrt{\frac{2\,\mathbf{E}_\mu x^2 \log(\delta^{-1})}{n}} + \sqrt{\frac{\mathbf{E}_\mu x^2}{n}}$$

which implies, as desired,

$$\widehat{x} - \mathbf{E}_{\mu} x \geq \sqrt{\frac{2\,\mathbf{E}_{\mu} x^2 \log(\delta^{-1})}{n}} + \sqrt{\frac{\mathbf{E}_{\mu} x^2}{n}}. \tag{15}$$

Since $-\psi(u) = \psi(-u)$, both of these settings can be interpreted as different settings of the distribution of the noise factor: $(1+\epsilon)$ in the upper bound case, and $-(1+\epsilon)$ in the lower bound case, both with $\epsilon \sim \nu$. Since the inequality (8) is uniform in the distribution of this noise, both bounds hold on the same event, which has probability no less than $1-\delta$. We may thus conclude that with probability at least $1-\delta$ over the random draw of the sample $x_1, \ldots, x_n$, the estimator $\widehat{x}$ satisfies

$$|\widehat{x} - \mathbf{E}_{\mu} x| \leq \sqrt{\frac{2\,\mathbf{E}_{\mu} x^2 \log(\delta^{-1})}{n}} + \sqrt{\frac{\mathbf{E}_{\mu} x^2}{n}}.$$

In practice, since $\mathbf{E}_{\mu} x^2$ will typically be unknown, this factor can be replaced by any valid upper bound $v \geq \mathbf{E}_{\mu} x^2$. The only impact to the final upper bound is that the unknown $\mathbf{E}_{\mu} x^2$ factors are replaced by the known v, concluding the proof. □

Proof of Lemma 4 (main text). Consider two data sets, the original $x_1, \ldots, x_n$ and a perturbed version $x_1', \ldots, x_n'$. For clean notation, organize these into vectors $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{x}' = (x_1', \ldots, x_n')$. Taking the difference between the estimator evaluated on these distinct data sets, we have

$$\begin{aligned}
\widehat{x}(\boldsymbol{x}) - \widehat{x}(\boldsymbol{x}') &= \int \frac{s}{n} \sum_{i=1}^{n} \left(\psi\left(\frac{(1+\epsilon)x_i}{s}\right) - \psi\left(\frac{(1+\epsilon)x_i'}{s}\right) \right) d\nu(\epsilon) \\
&\leq \int \frac{s}{n} \sum_{i=1}^{n} \left| \frac{1+\epsilon}{s} \right| |x_i - x_i'| \; d\nu(\epsilon) \\
&= \mathbf{E}_{\nu} |1+\epsilon| \frac{1}{n} \sum_{i=1}^{n} |x_i - x_i'| \\
&= \frac{\mathbf{E}_{\nu} |1+\epsilon|}{n} \|\boldsymbol{x} - \boldsymbol{x}'\|_1.
\end{aligned}$$

The first equality follows by linearity and the definition of the estimators. The subsequent inequality follows from the 1-Lipschitz property of ψ defined in (4, main text), which is that for all $u, v \in \mathbb{R}$, we have that $|\psi(u) - \psi(v)| \leq |u - v|$.

Evaluating $\mathbf{E}_{\nu} |1+\epsilon|$ is straightforward under the assumption that $\epsilon \sim \nu = N(0, 1/\beta)$, since the random variable $|1+\epsilon|$ follows a Folded Normal distribution. More generally, if $X \sim N(a, b^2)$, then $Y = |X|$ follows a folded normal distribution, with expected value

$$\mathbf{E}\, Y = a\left(1 - 2\Phi\left(\frac{-a}{b}\right)\right) + b\sqrt{\frac{2}{\pi}} \exp\left(\frac{-a^2}{2b^2}\right).$$

Since in our case, we have $a = 1$ and $b^2 = 1/\beta$, it follows that

$$\mathbf{E}_{\nu} |1+\epsilon| = 1 - 2\Phi\left(-\sqrt{\beta}\right) + \sqrt{\frac{2}{\beta\pi}} \exp\left(\frac{-\beta}{2}\right).$$

Reflecting this factor in the above inequalities concludes the proof. □

Proof of Lemma 5 (main text). In order to obtain bounds that hold uniformly over the choice of $\boldsymbol{w}$, we adopt a rather standard strategy utilizing covering numbers of $\mathcal{W}$. Using assumption A0, since $\mathcal{W}$ is closed and bounded, the Heine-Borel theorem implies that $\mathcal{W}$ is compact. This means the number of balls of radius ε required to cover $\mathcal{W}$ (denoted N_ε) is bounded above[3] as

$$N_\varepsilon \leq (3\Delta/2\varepsilon)^d. \tag{16}$$

Denote the centers of this ε-net by $\{\widetilde{\boldsymbol{w}}_1, \ldots, \widetilde{\boldsymbol{w}}_{N_\varepsilon}\}$. Given an abitrary $\boldsymbol{w} \in \mathcal{W}$ and center $\widetilde{\boldsymbol{w}} \in \{\widetilde{\boldsymbol{w}}_1, \ldots, \widetilde{\boldsymbol{w}}_{N_\varepsilon}\}$, we break the quantity to be controlled into three error terms, each to be tackled separately, as

$$\|\widehat{\boldsymbol{g}}(\boldsymbol{w}) - \boldsymbol{g}(\boldsymbol{w})\| \leq \|\widehat{\boldsymbol{g}}(\boldsymbol{w}) - \widehat{\boldsymbol{g}}(\widetilde{\boldsymbol{w}})\| + \|\boldsymbol{g}(\boldsymbol{w}) - \boldsymbol{g}(\widetilde{\boldsymbol{w}})\| + \|\widehat{\boldsymbol{g}}(\widetilde{\boldsymbol{w}}) - \boldsymbol{g}(\widetilde{\boldsymbol{w}})\|. \tag{17}$$

Let us start with the first term, $\|\widehat{\boldsymbol{g}}(\boldsymbol{w}) - \widehat{\boldsymbol{g}}(\widetilde{\boldsymbol{w}})\|$. Using Lemma 4 (main text), we have that

$$\begin{aligned}
\|\widehat{\boldsymbol{g}}(\boldsymbol{w}) - \widehat{\boldsymbol{g}}(\widetilde{\boldsymbol{w}})\|^2 &\leq \sum_{j=1}^{d} \left(\frac{c_\nu}{n} \sum_{i=1}^{n} |l_j'(\boldsymbol{w}; \boldsymbol{z}_i) - l_j'(\widetilde{\boldsymbol{w}}; \boldsymbol{z}_i)| \right)^2 \\
&\leq \sum_{j=1}^{d} (c_\nu \lambda \|\boldsymbol{w} - \widetilde{\boldsymbol{w}}\|)^2 \\
&= d c_\nu^2 \lambda^2 \|\boldsymbol{w} - \widetilde{\boldsymbol{w}}\|^2.
\end{aligned}$$

The first inequality is via Lemma 4 (main text), and the second via smoothness of the loss (via A1). We may thus control the first error term as

$$\|\widehat{\boldsymbol{g}}(\boldsymbol{w}) - \widehat{\boldsymbol{g}}(\widetilde{\boldsymbol{w}})\| \leq c_\nu \lambda \sqrt{d} \|\boldsymbol{w} - \widetilde{\boldsymbol{w}}\|. \tag{18}$$

Moving on to the second error term in the upper bound, this follows easily by smoothness of the risk (via A2), namely a Lipschitz property of the risk gradient. It immediately follows that

$$\|\boldsymbol{g}(\boldsymbol{w}) - \boldsymbol{g}(\widetilde{\boldsymbol{w}})\| \leq \lambda \|\boldsymbol{w} - \widetilde{\boldsymbol{w}}\| \tag{19}$$

for any choice of $\boldsymbol{w} \in \mathcal{W}$ and ε-ball center $\widetilde{\boldsymbol{w}}$.

Finally for the third error term, given any center $\widetilde{\boldsymbol{w}}$, as long as $\mathbf{E}_\mu \, l_j'(\widetilde{\boldsymbol{w}}; \boldsymbol{z})^2 < \infty$, then we can apply Lemma 2 (main text), implying

$$|\widehat{g}_j(\boldsymbol{w}) - g_j(\boldsymbol{w})| \leq \varepsilon_j := \sqrt{\frac{2 v_j \log(\delta^{-1})}{n}} + \sqrt{\frac{v_j}{n}}$$

where $v_j > 0$ is an upper bound on $\mathbf{E}_\mu \, |l_j'(\boldsymbol{w}; \boldsymbol{z})|^2$ used in the setting of s_j, in accordance with Lemma 2 (main text). For any pre-fixed $\boldsymbol{w}$, then for any $\varepsilon > 0$ we have

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

[3] This is a basic property of covering numbers for compact subsets of Euclidean space [6].

Using ε_j just defined, taking the maximum over $j \in [d]$, it follows that

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

Note that the second inequality follows immediately from $\varepsilon_j \leq \max_k \varepsilon_k$ and monotonicity of probability measures. Writing $V := \max_j v_j$, it follows immediately that fixing any $\boldsymbol{w} \in \mathcal{W}$, the nearest center $\widetilde{\boldsymbol{w}} := \widetilde{\boldsymbol{w}}(\boldsymbol{w})$ can be determined, and the event

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

has probability no greater than δ. The whole reason for utilizing a ε-cover of $\mathcal{W}$ in the first place is to avoid having to take a supremum over $\boldsymbol{w} \in \mathcal{W}$, which spoils union bounds, and instead to simply take a maximum over a finite number of ε-covers. The critical fact for our purposes is that

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

holds. The "good event" of interest is the one in which the bad event $\mathcal{E}(\cdot)$ holds for none of the centers on our ε-cover. In other words, the event

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

which taking a union bound, occurs with probability no less than $1 - \delta N_\varepsilon$. To get a $1 - \delta$ guarantee, simply pay the price of an extra logarithmic factor in the upper bound; that is to say, we equivalently have

$$\|\widehat{\boldsymbol{g}}(\widetilde{\boldsymbol{w}}(\boldsymbol{w})) - \boldsymbol{g}(\widetilde{\boldsymbol{w}}(\boldsymbol{w}))\| \leq \sqrt{\frac{2dV\log(dN_\varepsilon\delta^{-1})}{n}} + \sqrt{\frac{dV}{n}} \tag{20}$$

with probability no less than $1 - \delta$, uniformly in the choice of $\boldsymbol{w} \in \mathcal{W}$.

Taking these intermediate results together, we can form a useful uniform upper bound on (17), taking the form

$$\begin{aligned}
\sup_{\boldsymbol{w} \in \mathcal{W}} \|\widehat{\boldsymbol{g}}(\boldsymbol{w}) - \boldsymbol{g}(\boldsymbol{w})\| &\leq \sup_{\boldsymbol{w} \in \mathcal{W}} \left(c_\nu\lambda\sqrt{d}\|\boldsymbol{w} - \widetilde{\boldsymbol{w}}\| + \lambda\|\boldsymbol{w} - \widetilde{\boldsymbol{w}}\| + \|\widehat{\boldsymbol{g}}(\widetilde{\boldsymbol{w}}) - \boldsymbol{g}(\widetilde{\boldsymbol{w}})\|\right) \\
&\leq c_\nu\lambda\sqrt{d}\varepsilon + \lambda\varepsilon + \max_{k \in [N_\varepsilon]} \|\widehat{\boldsymbol{g}}(\widetilde{\boldsymbol{w}}_k) - \boldsymbol{g}(\widetilde{\boldsymbol{w}}_k)\| \\
&\leq \lambda\varepsilon(1 + c_\nu\sqrt{d}) + \sqrt{\frac{2dV\log(dN_\varepsilon\delta^{-1})}{n}} + \sqrt{\frac{dV}{n}}
\end{aligned}$$

with probability no less than $1 - \delta$ over the random draw of the sample. The bounds on the first, second, and third terms in the original upper bound come from (18), (19), and (20) respectively, with the ε factors following immediately from the definition of an ε-cover. To obtain the desired result, simply bound N_ε as in (16), and set $\varepsilon = 1/\sqrt{n}$, yielding updates to two of the terms, as

$$\begin{aligned} \lambda\varepsilon(1 + c_\nu\sqrt{d}) &= \frac{\lambda(1 + c_\nu\sqrt{d})}{\sqrt{n}} \\ \sqrt{\frac{2Vd\log(dN_\varepsilon\delta^{-1})}{n}} &\leq \sqrt{\frac{2Vd(\log(d\delta^{-1}) + d\log(3\Delta\sqrt{n}/2))}{n}} \end{aligned}$$

which, when plugged into the bound just obtained, concludes the proof. □

Proof of Lemma 6 (main text). Given $\widehat{\boldsymbol{w}}_{(t)}$, running the approximate update (2, main text), we have

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

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 [9, Thm. 2.1.15] it follows that

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

Writing $\gamma := 2\kappa\lambda/(\kappa+\lambda)$, the coefficient becomes $(1 - \alpha_{(t)}\gamma)$.

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 $a_{(t)} := \sqrt{1 - \alpha_{(t)}\gamma}$. Unfolding the recursion, on the good event, we have

$$\|\widehat{\boldsymbol{w}}_{(t+1)} - \boldsymbol{w}^*\| \leq \|\widehat{\boldsymbol{w}}_{(0)} - \boldsymbol{w}^*\| \prod_{k=0}^{t} a_{(k)} + \varepsilon\left(\alpha_{(t)} + \sum_{k=0}^{t-1}\alpha_{(k)}\prod_{l=k+1}^{t} a_{(l)}\right). \tag{21}$$

In the case of $\alpha_{(t)} = \alpha/\gamma$, things are very simple. We have $a_{(t)} = a := \sqrt{1-\alpha}$ for all t. The above inequality simplifies to

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

To clean up the second summand in (21),

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

This gives us the first statement as desired. For the case of $\alpha_{(t)} = 1/((2+t)\gamma)$, things are only slightly more complicated. First observe that

$$\prod_{m=2}^{M}\left(1 - \frac{1}{m}\right) = \prod_{m=2}^{M}\frac{m-1}{m} = \frac{1}{M}, \tag{22}$$

where the last equality follows by simply cancelling terms. We can now handle the first summand in (21) as

$$\left(\prod_{k=0}^{t} a_{(k)}\right)^2 = \prod_{k=0}^{t} \left(1 - \alpha_{(k)}\gamma\right) = \prod_{k=0}^{t} \left(1 - \frac{1}{2+k}\right) = \prod_{k=2}^{t+2} \left(1 - \frac{1}{k}\right) = \frac{1}{t+2},$$

where the final equality uses (22). As for the second summand in (21), first note that for any $k \geq 1$, we have

$$\frac{\alpha_{(k)}}{a_{(k)}\alpha_{(k-1)}} = \frac{(2+k-1)}{(2+k)\left(1 - \frac{1}{(2+k)}\right)} = 1.$$

Then recalling the second term on the right-hand side of (21), consider any two consecutive summands within the parentheses, say

$$\alpha_{(k)}a_{(k+1)} \cdots a_{(t)} \text{ and } \alpha_{(k-1)}a_{(k)} \cdots a_{(t)} \tag{23}$$

for any $1 \leq k < t$. Dividing the first term by the second term, note that almost all the factors cancel, yielding

$$\frac{\alpha_{(k)}a_{(k+1)} \cdots a_{(t)}}{\alpha_{(k-1)}a_{(k)} \cdots a_{(t)}} = \frac{\alpha_{(k)}}{a_{(k)}\alpha_{(k-1)}} = 1,$$

by what we just proved in (23). It follows that all terms inside the parentheses next to ε are identical, and indeed equal to $\alpha_{(t)}$, which is to say

$$\varepsilon\left(\alpha_{(t)} + \sum_{k=0}^{t-1} \alpha_{(k)} \prod_{l=k+1}^{t} a_{(l)}\right) = (t+1)\alpha_{(t)}\varepsilon = \frac{(t+1)\varepsilon}{(t+2)\gamma} \leq \frac{\varepsilon}{\gamma}.$$

Plugging these two new forms into the original inequality (21) yields our second desired result, and concludes the proof. □

Proof of Theorem 7 (main text). Using the strong convexity of R (via A4) and (3), it follows that

$$\begin{aligned} R(\widehat{\boldsymbol{w}}_{(T)}) - R^* &\leq \frac{\lambda}{2}\|\widehat{\boldsymbol{w}}_{(T)} - \boldsymbol{w}^*\|^2 \\ &\leq \lambda(1-\alpha)^T\|\widehat{\boldsymbol{w}}_{(0)} - \boldsymbol{w}^*\|^2 + \frac{4\lambda\varepsilon^2}{\gamma^2}. \end{aligned}$$

The latter inequality holds by direct application of Lemma 6 (main text) under fixed step size, followed by the elementary fact $(a+b)^2 \leq 2(a^2+b^2)$. The particular value of ε under which Lemma 6 (main text) is valid (i.e., under which (6, main text) holds) is given by Lemma 5 as $\widetilde{\varepsilon}$. Setting $\varepsilon = \widetilde{\varepsilon}$ yields the desired result. □

Proof of Theorem 9 (main text). We construct an upper bound using

$$\begin{aligned} \mathbf{E}\,\|\widehat{\boldsymbol{w}}_{(t+1)} - \widehat{\boldsymbol{w}}_{(t)}\|^2 &= \alpha_{(t)}^2\,\mathbf{E}\,\|\widehat{\boldsymbol{g}}(\widehat{\boldsymbol{w}}_{(t)})\|^2 \\ &\leq \alpha_{(t)}^2\,\mathbf{E}\left(\|\widehat{\boldsymbol{g}}(\widehat{\boldsymbol{w}}_{(t)}) - \boldsymbol{g}(\widehat{\boldsymbol{w}}_{(t)})\| + \|\boldsymbol{g}(\widehat{\boldsymbol{w}}_{(t)})\|\right)^2 \\ &\leq 2\alpha_{(t)}^2\left(\mathbf{E}\,\|\widehat{\boldsymbol{g}}(\widehat{\boldsymbol{w}}_{(t)}) - \boldsymbol{g}(\widehat{\boldsymbol{w}}_{(t)})\|^2 + \|\boldsymbol{g}(\widehat{\boldsymbol{w}}_{(t)})\|^2\right). \end{aligned}$$

Now, if we condition on $\widehat{\boldsymbol{w}}_{(t)}$, by assumption the loss gradients $\boldsymbol{l}'(\widehat{\boldsymbol{w}}_{(t)};\boldsymbol{z}_1),\ldots,\boldsymbol{l}'(\widehat{\boldsymbol{w}}_{(t)};\boldsymbol{z}_n)$ are iid. With independence, just as in the proof of Lemma 5 (main text), we have

$$\|\widehat{\boldsymbol{g}}(\widehat{\boldsymbol{w}}_{(t)}) - \boldsymbol{g}(\widehat{\boldsymbol{w}}_{(t)})\| \leq \sqrt{\frac{2dV\log(d\delta^{-1})}{n}} + \sqrt{\frac{dV}{n}},$$

with probability no less than $1-\delta$. Setting the right-hand side of this equation to ε and solving for δ, we have exponential tails of the form

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

with constant defined $a \coloneqq \sqrt{dV/n}$. Controlling moments using exponential tails can be done as follows. For random variable $X \in \mathcal{L}_p$ for $p \geq 1$, recall the classic inequality

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

Our setting of interest is $X = \|\widehat{\boldsymbol{g}}(\widehat{\boldsymbol{w}}_{(t)}) - \boldsymbol{g}(\widehat{\boldsymbol{w}}_{(t)})\|$, with $p = 2$. It follows that

$$\begin{aligned}
\mathbf{E}\,|X|^2 &= \int_0^\infty \mathbf{P}\{|X|^2 > u\}\,du \\
&= \int_0^\infty \mathbf{P}\{X > \sqrt{u}\}\,du \\
&= \int_0^\infty \mathbf{P}\{X > u\}\frac{u}{2}\,du \\
&\leq \frac{d}{2}\int_0^\infty \exp\left(-\frac{(u-a)^2}{2a^2}\right)u\,du.
\end{aligned}$$

The third equality uses substitution of variables, and the inequality at the end uses the exponential tail inequality given above. This integral is the expectation of the Normal distribution $N(a, a^2)$ taken over just the positive half-line. A simple upper bound can be constructed by

$$\begin{aligned}
\int_0^\infty \exp\left(-\frac{(u-a)^2}{2a^2}\right)u\,du &= \int_{-\infty}^\infty I\{u \geq 0\}\exp\left(-\frac{(u-a)^2}{2a^2}\right)u\,du \\
&\leq \int_{-\infty}^\infty \exp\left(-\frac{(u-a)^2}{2a^2}\right)|u|\,du,
\end{aligned}$$

easily recognized (after rescaling by $1/\sqrt{2\pi a^2}$) as the expectation of a Folded Normal random variable, induced by $N(a, a^2)$. Recalling the proof of Lemma 4 (main text), the expected value of this Folded Normal random variable is

$$\begin{aligned}
\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}a}\exp\left(-\frac{(u-a)^2}{2a^2}\right)|u|\,du &= a\left(1 - 2\Phi\left(\frac{-a}{a}\right)\right) + a\sqrt{\frac{2}{\pi}}\exp\left(\frac{-a^2}{2a^2}\right) \\
&= \sqrt{\frac{dV}{n}}\left(1 - 2\Phi\left(-1\right)\right) + \sqrt{\frac{2dV}{n\pi}}e^{-2} \\
&\leq \sqrt{\frac{dV}{n}}\left(1 + e^{-2}\sqrt{\frac{2}{\pi}}\right).
\end{aligned}$$

Taking into account the normalization factor, our upper bound takes the form

$$\begin{aligned}\mathbf{E}\,|X|^2 &\leq \frac{d\sqrt{2\pi}a}{2}\left(\sqrt{\frac{dV}{n}}\left(1+e^{-2}\sqrt{\frac{2}{\pi}}\right)\right)\\ &= \frac{d^2V}{n}\left(\sqrt{\frac{\pi}{2}}+\frac{1}{e^2}\right).\end{aligned}$$

Plugging this in for $X = \|\widehat{\boldsymbol{g}}(\widehat{\boldsymbol{w}}_{(t)}) - \boldsymbol{g}(\widehat{\boldsymbol{w}}_{(t)})\|$ in the upper bound constructed at the start of this proof yields the desired result. □

Computation

From Catoni and Giulini [3], Lemma 3.2, it follows that the correction term $C(a,b)$ used in (5, main text) can be computed as follows. First, some preparatory definitions to keep notation clean.

$$V_- := \frac{\sqrt{2}-a}{b}, \quad V_+ := \frac{\sqrt{2}+a}{b}$$

$$F_- := \Phi(-V_-), \quad F_+ := \Phi(-V_+)$$

$$E_- := \exp\left(-\frac{V_-^2}{2}\right), \quad E_+ := \exp\left(-\frac{V_+^2}{2}\right).$$

As seen in other parts of the text, Φ denotes the standard Normal CDF. With these atomic elements defined to keep things a bit cleaner, we break the final quantity into five terms to be summed:

$$\begin{aligned}T_1 &:= \frac{2\sqrt{2}}{3}\,(F_- - F_+)\\ T_2 &:= -\left(a - \frac{a^3}{6}\right)(F_- + F_+)\\ T_3 &:= \frac{b}{\sqrt{2\pi}}\left(1-\frac{a^2}{2}\right)(E_+ - E_-)\\ T_4 &:= \frac{ab^2}{2}\left(F_+ + F_- + \frac{1}{\sqrt{2\pi}}\,(V_+E_+ + V_-E_-)\right)\\ T_5 &:= \frac{b^3}{6\sqrt{2\pi}}\left((2+V_-^2)E_- - (2+V_+^2)E_+\right).\end{aligned}$$

With these terms in hand, the final computation is just summation, as

$$C(a,b) = T_1 + T_2 + T_3 + T_4 + T_5.$$

References

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

[2] Catoni, O. (2004). *Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour XXXI-2001*, volume 1851 of *Lecture Notes in Mathematics*. Springer.

[3] Catoni, O. and Giulini, I. (2017). Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv preprint arXiv:1712.02747*.

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

[5] Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26*, pages 315–323.

[6] 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.

[7] Lin, J. and Rosasco, L. (2016). Optimal learning for multi-pass stochastic gradient methods. In *Advances in Neural Information Processing Systems 29*, pages 4556–4564.

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

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

[10] Nocedal, J. and Wright, S. (1999). *Numerical Optimization.* Springer Series in Operations Research. Springer.

[11] Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press.

[12] 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.