*Supplemental material:*
# Classification using margin pursuit

Matthew J. Holland
Osaka University

## A Technical appendix

### A.1 Preliminaries

Here we put together few standard technical results that are utilized in the main proofs.

**Lemma 1.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable, convex, and $l$-smooth. Then, we have*

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

$$\|f'(\boldsymbol{u}) - f'(\boldsymbol{v})\|^2 \leq 2l\left(f(\boldsymbol{u}) - f(\boldsymbol{v}) - \langle f'(\boldsymbol{v}), \boldsymbol{u}-\boldsymbol{v}\rangle\right). \tag{2}$$

*Proof.* Given in chapter 2 of Nesterov [8]. □

**Lemma 2.** *The surrogate risk $R_\varphi(h)$ defined in (7, main text), for $h(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x}\rangle$, $\boldsymbol{w} \in \mathbb{R}^d$, is $l$-smooth with coefficient $l = \mathbf{E}\,\|\boldsymbol{x}\|^2 = v_X$.*

*Proof.* Assuming the order of integration and differentiation can be reversed, one can write $R'_\varphi$ as

$$R'_\varphi(\boldsymbol{w}) = -s\ \mathbf{E}\,\rho'\left(\frac{\gamma - y\langle \boldsymbol{w}, \boldsymbol{x}\rangle}{s}\right) y\boldsymbol{x}.$$

It follows that for arbitrary $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d$ we have

$$\begin{aligned}
\|R'_\varphi(\boldsymbol{w}_1) - R'_\varphi(\boldsymbol{w}_2)\| &\leq s\ \mathbf{E}\,\|\boldsymbol{x}\| \left|\rho'\left(\frac{\gamma - y\langle \boldsymbol{w}_1, \boldsymbol{x}\rangle}{s}\right) - \rho'\left(\frac{\gamma - y\langle \boldsymbol{w}_2, \boldsymbol{x}\rangle}{s}\right)\right| \\
&\leq s\ \mathbf{E}\,\|\boldsymbol{x}\| \left|\frac{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{x}\rangle}{s}\right| \\
&\leq \|\boldsymbol{w}_2 - \boldsymbol{w}_1\|\,\mathbf{E}\,\|\boldsymbol{x}\|^2
\end{aligned}$$

where we utilized the property that $\rho'$ is 1-Lipschitz. This implies that

$$\|R'_\varphi(\boldsymbol{w}_1) - R'_\varphi(\boldsymbol{w}_2)\| \leq l\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|, \quad \boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d \tag{3}$$

with coefficient $l = \mathbf{E}\,\|\boldsymbol{x}\|^2$, namely $R_\varphi$ is $\mathbf{E}\,\|\boldsymbol{x}\|^2$-smooth. □

**Lemma 3** (Confidence interval for sample mean of sub-Gaussian random vector)**.** *Let $\boldsymbol{x}$ be a random vector taking values in $\mathbb{R}^d$, with the sub-Gaussian property*

$$\mathbf{E}\exp(a\langle \boldsymbol{u}, \boldsymbol{x} - \mathbf{E}\,\boldsymbol{x}\rangle) \leq \exp(ca^2\langle \boldsymbol{u}, \Sigma_X \boldsymbol{u}\rangle), \quad a \geq 0$$

*for some constant $c > 0$ and $\Sigma_X := \mathbf{E}(\boldsymbol{x} - \mathbf{E}\,\boldsymbol{x})(\boldsymbol{x} - \mathbf{E}\,\boldsymbol{x})^T$. Given $n$ independent copies of $\boldsymbol{x}$, denoted $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, write $\bar{\boldsymbol{x}} := n^{-1} \sum_{i=1}^n \boldsymbol{x}_i$. Then with probability no less than $1 - \delta$, we have*

$$\|\bar{\boldsymbol{x}} - \mathbf{E}\,\boldsymbol{x}\| \leq 2\sqrt{\frac{c\|\Sigma_X\| \log(\delta^{-1})}{n}}.$$

*Proof.* We use the Chernoff extension of Markov's inequality to establish exponential tails for the deviation of the sample mean from its expectation, a standard technique [2]. For real-valued random variable $z \geq 0$, taking any $b > 0$ we have $b\, I\{z \geq b\} \leq z\, I\{z \geq b\}$ almost surely. Integrating both sides implies $b\, \mathbf{P}\{z \geq b\} \leq \mathbf{E}\, zI\{z \geq b\} \leq \mathbf{E}\, z$, using the non-negativity of $z$ for the latter inequality. Thus $\mathbf{P}\{z \geq b\} \leq \mathbf{E}\, z/b$, the classic Markov inequality. For non-decreasing function $f(z) \geq 0$, this naturally extends via $\mathbf{P}\{z \geq b\} \leq \mathbf{P}\{f(z) \geq f(b)\}$ to $\mathbf{P}\{z \geq b\} \leq \mathbf{E}\, f(z)/f(b)$, now for any real-valued random variable $z$. When $\mathbf{E}\, z = 0$, setting $f(z) = z^2$ yields the special case of Chebyshev's inequality. Chernoff's inequality follows from the special case of $f(z) = \exp(az)$, for $a > 0$, with the form

$$\mathbf{P}\{z \geq b\} \leq e^{-ab}\, \mathbf{E} \exp(az).$$

If the moment generating function of $z$ is not finite, then of course these bounds are vacuous, but in the sub-Gaussian case we have easily manipulated upper bounds. In our setup we have $z = \langle \boldsymbol{u}, \boldsymbol{x} - \mathbf{E}\,\boldsymbol{x} \rangle$, and by our hypothesis we have for any $\|\boldsymbol{u}\| = 1$ that

$$\begin{aligned}
\mathbf{P}\{\langle \boldsymbol{u}, \boldsymbol{x} - \mathbf{E}\,\boldsymbol{x} \rangle \geq b\} &\leq e^{-ab} \exp(ca^2 \langle \boldsymbol{u}, \Sigma_X \boldsymbol{u} \rangle) \\
&\leq \exp\left(ca^2 \|\Sigma_X\| - ab\right)
\end{aligned}$$

where $\|\Sigma_X\|$ denotes the $\ell_2$-induced matrix norm, equivalent to the spectral norm, i.e., the largest singular value of $\Sigma_X$ [6]. Since this holds for any $a > 0$, this upper bound can be made as tight as possible when we set $a = b/(2c\|\Sigma_X\|)$, resulting in

$$\mathbf{P}\{\langle \boldsymbol{u}, \boldsymbol{x} - \mathbf{E}\,\boldsymbol{x} \rangle \geq b\} \leq \exp\left(-\frac{b^2}{4c\|\Sigma_X\|}\right).$$

For the special case of $\boldsymbol{u} = (\boldsymbol{x} - \mathbf{E}\,\boldsymbol{x})/\|\boldsymbol{x} - \mathbf{E}\,\boldsymbol{x}\|$, we have $\langle \boldsymbol{u}, \boldsymbol{x} - \mathbf{E}\,\boldsymbol{x} \rangle = \|\boldsymbol{x} - \mathbf{E}\,\boldsymbol{x}\|$, yielding the same bound for $\mathbf{P}\{\|\boldsymbol{x} - \mathbf{E}\,\boldsymbol{x}\| \geq b\}$ as a special case.

Finally, for the sample mean, we note that

$$\langle \boldsymbol{u}, \bar{\boldsymbol{x}} - \mathbf{E}\,\boldsymbol{x} \rangle = \frac{1}{n} \sum_{i=1}^n \langle \boldsymbol{u}, (\boldsymbol{x}_i - \mathbf{E}\,\boldsymbol{x}) \rangle\,.$$

Plugging this in to our Chernoff equality,

$$\begin{aligned}
\mathbf{P}\{\langle \boldsymbol{u}, \bar{\boldsymbol{x}} - \mathbf{E}\,\boldsymbol{x} \rangle \geq b\} &\leq e^{-anb} \prod_{i=1}^n \exp(ca^2 \langle \boldsymbol{u}, \Sigma_X \boldsymbol{u} \rangle) \\
&\leq \exp\left(nca^2 \|\Sigma_X\| - anb\right).
\end{aligned}$$

Once again optimizing with respect to $a$, and setting $\boldsymbol{u} = (\bar{\boldsymbol{x}} - \mathbf{E}\,\boldsymbol{x})/\|\bar{\boldsymbol{x}} - \mathbf{E}\,\boldsymbol{x}\|$ as noted above, we have

$$\mathbf{P}\{\|\bar{\boldsymbol{x}} - \mathbf{E}\,\boldsymbol{x}\| \geq b\} \leq \exp\left(-\frac{nb^2}{4c\|\Sigma_X\|}\right)$$

which implies the desired result. □

## A.2 Proofs of results in the main text

*Derivation of (3, main text).* Considering the definition of $\psi$, for the case of $|u| \leq \sqrt{2}$, we have an indefinite integral

$$\int \psi(u)\, du = \frac{u^2}{2} - \frac{u^4}{24} + C_1,$$

and since we would like a loss function taking a value of zero at $u = 0$, we set $C_1 = 0$. For the case of $u > \sqrt{2}$, we have

$$\int \psi(u)\, du = \frac{2\sqrt{2}}{3} u + C_2.$$

Since we would like this integral to be continuous on $\mathbb{R}$, we must have

$$\frac{u^2}{2} - \frac{u^4}{24} = \frac{2\sqrt{2}}{3} u + C_2$$

when $u = \sqrt{2}$. Setting $C_2 = -1/2$ achieves this. □

*Proof of Remark 1 (main text).* For simplicity, consider instance space $\mathcal{X} = \mathbb{R}$. As an intuitive model, consider $\mathcal{H}_{\text{ray}}$, the set of all classifiers defined by rays in the "left" direction. That is, each $h \in \mathcal{H}_{\text{ray}}$ takes the form

$$h(x; \alpha) = I\{x \leq \alpha\} - I\{x > \alpha\}$$

for some $\alpha \in \mathbb{R}$. Upon the underlying distribution, break up the input space into three segments, $(-\infty, \alpha_l^*]$. $(\alpha_l^*, \alpha_u^*)$, $[\alpha_l^*, \infty)$, with probabilities

$$\mathbf{P}\{x <= \alpha_l^*\} = \mathbf{P}\{x >= \alpha_u^*\} = 1/3.$$

It follows that $\mathbf{P}\{x \in (\alpha_l^*, \alpha_u^*)\} = 1/3$ as well. Furthermore, assume that the labeling of pair $(x, y)$ is done as

$$x \mapsto y = \begin{cases} 1, & x \notin (\alpha_l^*, \alpha_u^*) \\ -1, & x \in (\alpha_l^*, \alpha_u^*) \end{cases}.$$

In this situation, given the probabilities, it is evident that in terms of minimizing the classification error $\mathbf{E}\, I\{y \neq h(x)\}$, the optimal choice is to select $h(\cdot; \alpha_l^*)$, in which case

$$\mathbf{E}\, I\{y \neq h(x; \alpha_l^*)\} = \mathbf{P}\{x >= \alpha_l^*\} = 1/3.$$

This gives us a lower bound on performance, namely

$$\mathbf{E}\, I\{y \neq h(x)\} \geq 1/3, \quad \forall\, h \in \mathcal{H}_{\text{ray}}.$$

Now, in the limiting case of $\rho(u) = u^2$, say we have

$$h^* \in \operatorname*{arg\,min}_{h \in \mathcal{H}_{\text{ray}}} \mathbf{E}\, (\gamma - y\, h(\boldsymbol{x}))^2$$

for a pre-fixed value of $\gamma > 1/3$, something we are free to do. Note that as $\mathbf{P}\{y\, h(x) <= 0\} = \mathbf{P}\{y \neq h(x)\} \geq 1/3$, we have

$$\begin{aligned}\mathbf{E}\, y\, h^*(x) &= \mathbf{P}\{y\, h^*(x) > 0\} - \mathbf{P}\{y\, h^*(x) <= 0\} \\ &\leq \frac{2}{3} - \frac{1}{3} \\ &= \frac{1}{3} \\ &< \gamma.\end{aligned}$$

Since we know that

$$\mathbf{E}\, y\, h^*(x) = \arg\min_{\gamma}\, \mathbf{E}\left(\gamma - y\, h^*(\boldsymbol{x})\right)^2,$$

it follows that the margin level $\gamma$ need not provide a reliable measure of the location of the distribution of $y\, h^*(x)$ over a random draw from the underlying data distribution. □

*Proof of Proposition 2 (main text).* We begin with a sufficient condition for $\gamma$ to equal $\widehat{\gamma}(h)$,

$$\sum_{i=1}^{n} \rho'\left(\gamma - y_i\, h(\boldsymbol{x}_i)\right) = 0.$$

This function is bounded on $\mathbb{R}$ by $\pm B$, where $B := \rho'(\sqrt{2})$. For clarity, write $a_i := y_i\, h(\boldsymbol{x}_i)$ for $i \in [n]$. Assume without loss of generality that $n > 1$ is odd and $a_i \leq a_{i+1}$ for $i \in [n-1]$. Writing $m := (n+1)/2$, the median value is $a_m$. Obviously, taking any scale such that

$$0 < s < \frac{|a_m - a_i|}{\sqrt{2}}, \quad i \neq m$$

it follows immediately that

$$\sum_{i=1}^{n} \rho'\left(a_m - a_i\right) = 0$$

since the $m$th summand is zero, the first $(n-1)/2$ summands equal $\sqrt{2}$, and the last $(n-1)/2$ summands equal $-\sqrt{2}$, canceling each other out. Thus for any $s$ small enough, the median is a valid solution. Extending this to the case of $n$ even is straightforward. Writing $m := n/2$ now, since $\rho'(u) = -\rho'(-u)$, it follows that $\rho'_s(\gamma - a_m) + \rho'_s(\gamma - a_{m+1}) = 0$ when we set $\gamma = (a_m + a_{m+1})/2$. Looking at the sum over $\{\rho'(\gamma - a_i)\}_{i\in[n]}$ There are $(n-2)/2$ terms no less than these two middle terms, and $(n-2)/2$ terms no greater than them. Just as before, taking $s > 0$ small enough, the former will equal $\sqrt{2}$ and the latter $-\sqrt{2}$, once again canceling each other out and leaving the median as a valid solution. This proves part 1 of the hypothesis.

As for the empirical mean case (part 2), we use a more general result, taken from Holland and Ikeda [5]:

**Lemma 4.** *Let $x$ be an arbitrary random variable with distribution $\nu$. Assuming $\mathbf{E}_\nu\, |x|^3 < \infty$, it follows that defining*

$$\theta^* := \arg\min_{\theta\in\mathbb{R}} \mathbf{E}_\nu\, \rho_s(\theta - x)$$

*the deviation can be controlled as*

$$|\theta^* - \mathbf{E}_\nu\, x| \leq c s^{-2}, \quad s > 0$$

*for constant $c > 0$.*

Substituting underlying data distribution for $\nu$, and $y\,h(\boldsymbol{x})$ for $x$, and considering the analogous

$$\gamma^*(h) \coloneqq \operatorname*{arg\,min}_{\gamma \in \mathbb{R}} \mathbf{E}\, \rho_s(\gamma - y\,h(\boldsymbol{x}))$$

it immediately follows that

$$|\gamma^*(h) - \mathbf{E}\, y\,h(\boldsymbol{x})| = O\left(\frac{1}{s^2}\right)$$

for any valid distribution (i.e., where the third moment condition holds). This holds for the case of the empirical distribution $P_n(A) \coloneqq n^{-1} \sum_{i=1}^n I\{y_i\,h(\boldsymbol{x}_i) \in A\}$, and plugging in $P_n$ for $\nu$ we have that $\gamma^*(h) = \widehat{\gamma}(h)$, and obtain part 2. □

*Proof of Proposition 4 (main text).* This stability property follows from basic properties of the function $\rho$, as follows. Since $h \in \mathcal{H}$ is pre-fixed, we suppress it in the notation. Write $\{q_i\}_{i=1}^n$ and $\{q_i'\}_{i=1}^n$ for the margins $y\,h(\boldsymbol{x})$ evaluated on the two data sets of interest (original and modified). By definition, on the pre-modification data set, we have

$$\sum_{i=1}^n \rho'\left(\frac{\widehat{\gamma} - q_i}{s}\right) = 0.$$

Assume that the $j$th index is the one where $\boldsymbol{z}_j \neq \boldsymbol{z}_j'$, and thus where $q_j \neq q_j'$. Without loss of generality, assume $q_j > q_j'$. In the optimistic case, where $\widehat{\gamma} - q_j \geq \sqrt{2}$, this does not impact the estimator at all, and $\widehat{\gamma} = \widehat{\gamma}'$. On the pessimistic side, the largest impact possible would be

$$\begin{aligned}
\sum_{i=1}^n \rho'\left(\frac{\widehat{\gamma} - q_i'}{s}\right) &= \rho'\left(\frac{\widehat{\gamma} - q_j'}{s}\right) - \rho'\left(\frac{\widehat{\gamma} - q_j}{s}\right) \\
&\geq \rho'(\sqrt{2}) - \rho'(-\sqrt{2}) \\
&= 2\rho'(\sqrt{2}) \\
&= 4\sqrt{2}/3.
\end{aligned}$$

Thus, in order to satisfy the first order condition

$$\sum_{i=1}^n \rho'\left(\frac{\widehat{\gamma}' - q_i'}{s}\right) = 0,$$

a shift from $\widehat{\gamma}$ to $\widehat{\gamma}'$ must in the worst case make up $4\sqrt{2}/3$, by sufficiently decreasing $\widehat{\gamma}'$. Define an index

$$\mathcal{I} \coloneqq \left\{ i \in [n] : \frac{|\widehat{\gamma} - q_i|}{s} \leq \frac{\sqrt{2}}{2} \right\}.$$

The value $\sqrt{2}/2$ in the definition of $\mathcal{I}$ is arbitrary; any value less than $\sqrt{2}$ would work fine, but this allows for a straightforward argument. The points with $i \in \mathcal{I}$ give us a worst-case value for how far we must shift from $\widehat{\gamma}$ to $\widehat{\gamma}'$, as follows. Write $m \coloneqq |\mathcal{I}|$, and note that if $m \geq 3$, a jump of width $\sqrt{(2/m)}$ from the edge of our "good range" of $[-\sqrt{2}/2, \sqrt{2}/2]$ in the slowest-changing direction (say positive side, without loss of generality) yields a slope of

$$\rho''\left(\frac{\sqrt{2}}{2} + \sqrt{\frac{2}{m}}\right) = \frac{1}{2} - \frac{2(\sqrt{m}+1)}{m}.$$

By symmetry, the slope on the negative side is the same. It follows that

$$\rho'\left(-\frac{\sqrt{2}}{2}\right) - \rho'\left(-\frac{\sqrt{2}}{2} - \sqrt{\frac{2}{m}}\right) \geq D \coloneqq \sqrt{\frac{2}{m}}\left(\frac{1}{2} - \frac{2(\sqrt{m}+1)}{m}\right).$$

Shifting $\widehat{\gamma}$ such that

$$\frac{\widehat{\gamma}' - q_i'}{s} - \frac{\widehat{\gamma} - q_i'}{s} = -\sqrt{\frac{2}{m}},$$

which is to say setting $\widehat{\gamma}' = \widehat{\gamma} - s\sqrt{(2/m)}$, we know that at the very least, each $i \in \mathcal{I}$ contributes $-D$ to the sum in the first-order optimality condition. That is to say, we have

$$\sum_{i=1}^{n} \rho'\left(\frac{\widehat{\gamma} - q_i'}{s}\right) - \sum_{i=1}^{n} \rho'\left(\frac{\widehat{\gamma}' - q_i'}{s}\right) \geq mD.$$

All that is left is to ensure $mD \geq 4\sqrt{2}/3$. Some basic arithmetic shows that $m \geq 24$ implies $mD \geq 4\sqrt{2}/3$. Thus, we conclude that with $s$ such that $m = |\mathcal{I}| \geq 24$, the true $\widehat{\gamma}'$ for the modified set can in the worst case be no farther from $\widehat{\gamma}$ than $s\sqrt{(2/m)}$, concluding the proof. Assuming $|\mathcal{I}| \geq n/2 \geq 24$ yields the desired result as a special case. □

*Proof of Lemma 6 (main text).* Extending the results of Catoni [3], as long as $\rho$ satisfies

$$-\log(1 - u + Cu^2) \leq \rho'(u) \leq \log(1 + u + Cu^2), \quad u \in \mathbb{R} \tag{4}$$

then exponential tails on the empirical estimator's deviation can be obtained; there is nothing particularly special about $\rho$ defined in (3, main text), except the computational convenience and ease of analysis. Given this inequality, Lemma 1 of Holland and Ikeda [5] implies that

$$\mathbf{P}\left\{\frac{|\widehat{\gamma}(h) - \mathbf{E}\, y\, h(\boldsymbol{x})|}{2} \leq \frac{C \operatorname{var} y\, h(\boldsymbol{x})}{s} + \frac{s \log(2\delta^{-1})}{n}\right\} \geq 1 - \delta.$$

For our setting, $\rho$ defined in (3, main text) indeed satisfies (4), with $C = 1/2$, which follows from Catoni and Giulini [4, Lemma 1], where this function is studied in the context of robust vector mean estimates. Optimizing the upper bound with respect to $s > 0$ and plugging in $C = 1/2$ yields the desired result. □

*Proof of Lemma 8 (main text).* We follow along with the now-standard framework set out by Bartlett et al. [1]. For simplicity, we start with the special case of $s = 1$, where the loss function becomes $\varphi(u) = \rho(\gamma - u)$.

Next we put together the analytical machinery that will be used. First, the conditional expected $\varphi$-risk takes the form

$$\mathbf{E}\left(\varphi(y\, h(\boldsymbol{x}))|\boldsymbol{x}\right) = \eta\varphi(y\, h(\boldsymbol{x})) + (1 - \eta)\varphi(y\, h(\boldsymbol{x}))$$

where $\eta$ denotes $\eta = \mathbf{P}\{y = 1\}$. A generalization of this quantity for arbitrary $\eta \in [0, 1]$ is constructed as

$$C_\eta(u) \coloneqq \eta\varphi(u) + (1 - \eta)\varphi(-u), \quad u \in \mathbb{R}.$$

The optimal value that this takes is denoted by

$$H(\eta) \coloneqq \inf_{u \in \mathbb{R}} C_\eta(u).$$

Denote any optimal value using $u^*$, namely any

$$u^* \in \arg\min_{u \in \mathbb{R}} C_\eta(u).$$

If this value is indeed unique, then it makes sense to map $\eta \mapsto u^*(\eta)$. In relating $R$ and $R_\varphi$, our intuitive concern is the degree to which, on average, $\varphi(y\, h(\boldsymbol{x}))$ can be small while $I\{h(\boldsymbol{x}) \neq y\}$ remains non-zero. This notion is captured formally by the following nice quantity:

$$H^-(\eta) := \inf \{C_\eta(u) : u(2\eta - 1) \leq 0\}.$$

Using the "best (generalized) conditional $\varphi$-risk that can be achieved despite having different signs from $(2\eta - 1)$." Of course, the word "despite" becomes appropriate when we replace $\eta$ with the conditional probability $\eta(\boldsymbol{x}) = \mathbf{P}\{y = 1|\boldsymbol{x}\}$, in which $\operatorname{sign}(2\eta(\boldsymbol{x}) - 1)$ is the Bayes decision rule for this classification task. For $\varphi$ to be a good surrogate, we would expect that $H^-$ should tend to be larger than $H$. If this was not the case, a small $\varphi$-risk could be achieved despite having mis-labeled some instances, which would immediately imply a small $R_\varphi$ but larger $R$. To ensure that this cannot happen, the condition put forward by Bartlett et al. [1] is very lucid: call $\varphi$ *classification-calibrated* if

$$H^-(\eta) > H(\eta), \quad \forall\, \eta \neq 1/2.$$

The size of this gap is defined as

$$\widetilde{\Psi}(a) := H^- \left(\frac{1+a}{2}\right) - H\left(\frac{1+a}{2}\right),$$

and the Fenchel-Legendre bi-conjugate of $\widetilde{\Psi}$ is denoted by $\Psi$. It is this function that has the desirable properties that interest us. For one, via their Theorem 1, for any non-negative $\varphi$, any distribution on $\mathcal{X} \times \{-1, 1\}$ and measurable function $h$, we have

$$\Psi\left(R(h) - R^*\right) \leq R_\varphi(h) - R_\varphi^*.$$

It is easy to characterize this classification-calibration in the convex case. If $\varphi$ is convex, then via their Theorem 2(1),

$$\varphi \text{ is classification calibrated } \iff \varphi \text{ is differentiable at zero with } \varphi'(0) < 0.$$

Since our function is $(d/du)\varphi(u) = \rho(\gamma - u)(-1)$, for $\gamma > 0$ we have $\rho(\gamma) > 0$ and thus $(d/du)\varphi(0) < 0$ as desired. Thus our $\varphi$, being a convex function on $\mathbb{R}$, is classification calibrated. Furthermore, via Theorem 2(2), the $\Psi$ function takes a particularly simple form:

$$\Psi_{1,\gamma}(a) = \varphi(0) - H_{1,\gamma}\left(\frac{1+a}{2}\right), \quad -1 \leq a \leq 1$$

where $\varphi(0) = \rho(\gamma)$ gives us the expression from the hypothesis in the case of $s = 1$. All that remains in order to obtain $\Psi_{1,\gamma}$ then is to compute $H_{1,\gamma}$ explicitly, which we carry out below.

Now, since both $\varphi(u)$ and $\varphi(-u)$ are convex functions of $u$, and $C_\eta(u)$ is a convex combination of these two, it follows that $C_\eta(u)$ is also convex. Furthermore, noting that both $u \to \infty$ and $u \to -\infty$ imply $C_\eta(u) \to \infty$. Thus a minimum clearly exists, and can be characterized by a first-order condition as follows. Taking the first derivative of $C_\eta(\cdot)$, we have

$$\frac{d}{du} C_\eta(u) = \eta\rho'(\gamma - u)(-1) + (1 - \eta)\rho'(\gamma + u) = 0$$

which using $\rho'(-u) = (-1)\rho'(u)$, can be equivalently stated as

$$\frac{\rho'(u-\gamma)}{\rho'(u+\gamma)} = \frac{\eta - 1}{\eta}. \tag{5}$$

That is to say, for $\eta \in (0,1)$, any $u^*$ satisfying (5) will be a minimizer in that $C_\eta(u^*) \leq C_\eta(u)$ for all $u$.

It should be clear that the value of $\gamma$ plays an important role in finding the solution. Note that $\sqrt{2}$ is an important threshold here, since

$$u \geq \sqrt{2} \implies \rho'(u) = \rho'(\sqrt{2}) = \frac{2\sqrt{2}}{3}.$$

On the "left" side as well, $u \leq -\sqrt{2}$ implies $\rho'(u) = -\rho'(\sqrt{2})$.

For the case of $\eta = 0$, we have $u^* = -\gamma$, and when $\eta = 1$ we have $u^* = \gamma$. This implies that $H_{1,\gamma}(0) = H_{1,\gamma}(1) = 0$. More generally, an obvious but important fact is that for any $\eta \in (0,1)$, any solution $u^*$ must fall on the open interval $(-\gamma, \gamma)$. This is because the right-hand side of (5) is always negative, but the left-hand side is negative if and only if $u - \gamma < 0 < u + \gamma$, equivalently $u \in (-\gamma, \gamma)$. Also, for the case of $\eta = 1/2$, we have that (5) is always satisfied by setting $u = 0$, for which case we have

$$H_{1,\gamma}(1/2) = \frac{1}{2}\rho(\gamma - 0) + \frac{1}{2}\rho(\gamma + 0) = \rho(\gamma).$$

This value, and thus the height of the peak of $H_{1,\gamma}(\cdot)$, changes as a function of $\gamma$. Let's proceed and look at evaluating $H_{1,\gamma}(\eta)$ for $\eta \in (0,1)$ when $\eta \neq 1/2$. We shall consider the following distinct settings:

1. $0 < \gamma \leq \sqrt{2}/2$
2. $\sqrt{2}/2 \leq \gamma < \sqrt{2}$
3. $\sqrt{2} \leq \gamma$

Doing these one at a time, first consider $0 < \gamma \leq \sqrt{2}/2$. This case is simple, since for any solution $u^* \in (-\gamma, \gamma)$ we have that $u^* \pm \gamma \in [-\sqrt{2}, \sqrt{2}]$ and thus we can set $\rho'(u) = u - u^3/6$, re-arrange equality (5), and solve for roots of the resulting cubic polynomial. The computations are quick, and writing

$$P(u; a, b, c, d) := au^3 + bu^2 + cu + d$$

and $\alpha := (\eta - 1)/\eta$, the new condition is

$$\begin{aligned} a &= 1 - \alpha \\ b &= -3\gamma(1 + \alpha) \\ c &= 3(1 - \alpha)(\gamma^2 - 2) \\ d &= (1 + \alpha)(6\gamma - \gamma^3) \\ P(u; a, b, c, d) &= 0. \end{aligned} \tag{6}$$

Call this (6) the *double-cube* condition (see Figure 1). The discriminant of an arbitrary cubic polynomial is defined

$$\Delta := 18abcd - 4b^3d + b^2c^2 - 4ac^3 - 27a^2d^2, \tag{7}$$

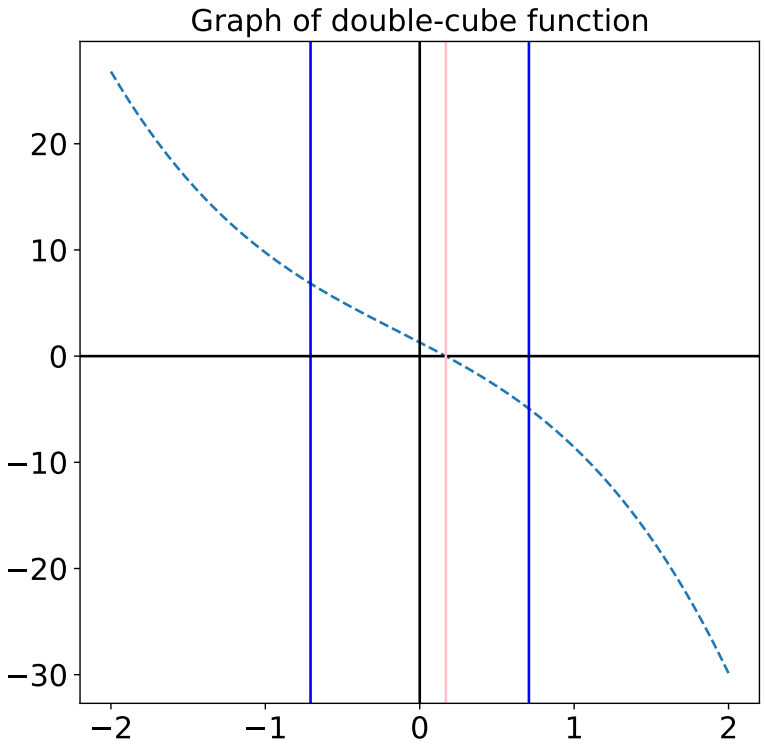


**Figure 1:** Graph of the third-degree polynomial used in the double-cube condition (6). Vertical blue lines denote $\pm\gamma$ (here $\gamma = \sqrt{2}/2$), and the vertical pink line denotes the root computed analytically.

and as long as $\Delta < 0$, the function $P(u)$ has only one real root, which can be computed analytically (see Appendix). Writing $u^*(\eta)$ for the real value satisfying $P(u^*(\eta)) = 0$ here, by plugging this into the original objective we get $H_{1,\gamma}(\eta) = C_\eta(u^*(\eta))$.

Next consider the case of $\sqrt{2}/2 < \gamma < \sqrt{2}$. This is the most complicated case. Writing $\delta := |\sqrt{2} - \gamma|$, if a solution exists on the interval $[-\delta, \delta]$, then it will naturally satisfy the double-cube condition given above. If there is no solution on this interval, then depending on whether $\eta > 1/2$ or $\eta < 1/2$, the appropriate condition will respectively be

$$\rho'(u - \gamma) - \rho'(\sqrt{2})\frac{\eta - 1}{\eta} = 0$$

or

$$\rho'(u + \gamma) + \frac{\eta}{\eta - 1}\rho'(\sqrt{2}) = 0.$$

Call these the *minus* and *plus single-cube* conditions. Re-arranged into more explicit terms, we have respectively

$$\begin{aligned} a &= 1 \\ b &= -3\gamma \\ c &= 3\gamma^2 - 6 \\ d &= -6\left(\gamma + \frac{\rho'(\sqrt{2})}{\alpha} - \frac{\gamma^3}{6}\right) \\ P(u; a, b, c, d) &= 0 \end{aligned} \tag{8}$$

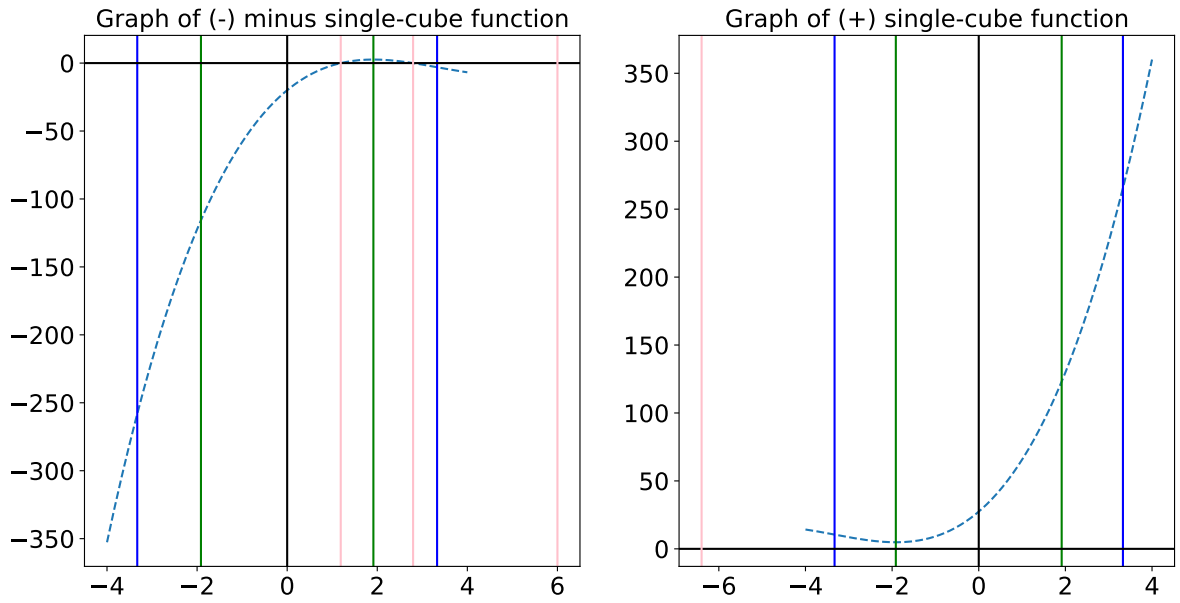


**Figure 2:** Graph of the third-degree polynomials used in the single-cube conditions. The left figure corresponds to condition (8), and the right figure corresponds to condition (9). The vertical blue lines are again $\pm\gamma$ (with $\gamma = 2\sqrt{2} + 1/2$ here), and the vertical green lines are $\pm\delta$.

and

$$\begin{aligned} a &= 1 \\ b &= 3\gamma \\ c &= 3\gamma^2 - 6 \\ d &= 6\left(\gamma + \alpha\rho'(\sqrt{2}) - \frac{\gamma^3}{6}\right) \\ P(u; a, b, c, d) &= 0 \end{aligned} \tag{9}$$

Graphs of these polynomials are plotted in Figure 2. Computationally determining which to use is straightforward. By the monotonicity of $\rho'$, we can simply check the edge case $u = \text{sign}(\eta - 1/2)\delta$. In the case of $\eta > 1/2$, noting that both the LHS and RHS are negative, if

$$\frac{\rho'(\delta - \gamma)}{\rho'(\delta + \gamma)} < \frac{\eta - 1}{\eta}, \tag{10}$$

then the solution must be larger than $\delta$, and thus the minus single-cube condition will be sufficient. Else, the double-cube condition will provide a solution. On the other hand, when $\eta < 1/2$, if

$$\frac{\rho'(-\delta - \gamma)}{\rho'(-\delta + \gamma)} > \frac{\eta - 1}{\eta}, \text{ or more cleanly, } \frac{\rho'(\delta - \gamma)}{\rho'(\delta + \gamma)} < \frac{\eta}{\eta - 1}, \tag{11}$$

then the solution must be below $-\delta$, and thus the plus single-cube condition will be sufficient. Else, the double-cube condition will provide a solution. This gives us a simple procedure for the current range of $\gamma$ values being considered[1], as follows:

[1]While there may be more than one real root of the cubic polynomials used in these conditions, there will not be more than one root in the range of $(\delta, \gamma)$ ($\eta > 1/2$ case) or $(-\gamma, -\delta)$ ($\eta < 1/2$ case).

- When $\eta > 1/2$:
  - If (10), then solve (8), take root falling in $(\delta, \gamma)$.
  - Else, solve (6).
- When $\eta < 1/2$:
  - If (11), then solve (9), take root falling in $(-\gamma, -\delta)$.
  - Else, solve (6).

Finally, consider the case of $\sqrt{2} \leq \gamma$. This situation is simple: if $\eta > 1/2$, find solutions to the minus single-cube condition, and if $\eta < 1/2$, find solutions to the plus single-cube condition.

With all these conditions in place, it follows that for any $\gamma > 0$ and any $\eta \in [0, 1]$, we can find a solution $u^*$ such that $C_\eta(u^*) = H_{1,\gamma}(\eta)$. It follows then that following the procedures outlined above, we can also compute $\Psi_{1,\gamma}(a) = \varphi(0) - H_{1,\gamma}((1+a)/2)$ for arbitrary $a \in [-1, 1]$.

The consistency part of the lemma statement follows immediately from the calibration of $\varphi$, using Bartlett et al. [1, Theorem 1(3)]. Invertibility of $\Psi_{1,\gamma}$ follows from convexity of $\varphi$ and Lemma 2 of Bartlett et al. [1].

It remains only to extend these results to the case of arbitrary $s > 0$, namely the general loss function $\varphi(u) = s^2\,\rho((\gamma - u)/s)$. That $\varphi(u)$ is convex and that $\varphi'(0) < 0$ under arbitrary $s > 0$ is immediate. Furthermore, the first-order optimality condition becomes

$$\frac{\eta - 1}{\eta} = \frac{\rho'((u-\gamma)/s)}{\rho'((u+\gamma)/s)} = \frac{\rho'((u/s) - (\gamma/s))}{\rho'((u/s) + (\gamma/s))}. \tag{12}$$

Writing $\widetilde{\gamma} = \gamma/s$, note that using the exact same procedures outlined above, we can always find a $u'$ such that

$$\frac{\eta - 1}{\eta} = \frac{\rho'(u' - \widetilde{\gamma})}{\rho'(u' + \widetilde{\gamma})},$$

which means that writing $u^* = su'$, we have that $u^*$ is a solution of (12). This means that for any $\gamma > 0$, $s > 0$, and $\eta \in [0, 1]$, we can find a solution $u^*$ such that $C_\eta(u^*) = H_{s,\gamma}(\eta)$, which yields the general $\Psi_{s,\gamma}$ as

$$\Psi_{s,\gamma}(u) = s^2\,\rho\left(\frac{\gamma}{s}\right) - H_{s,\gamma}\left(\frac{1+u}{2}\right),$$

concluding the proof. □

*Proof of Theorem 11, main text.* To keep notation clean, throughout this proof we denote the risk gradient by $\boldsymbol{g}^*(\boldsymbol{w}) \coloneqq R'(\boldsymbol{w})$, the surrogate risk gradient by $\boldsymbol{g}(\boldsymbol{w}) \coloneqq R'_\varphi(\boldsymbol{w})$, and the new loss gradient by $\widehat{\boldsymbol{g}}(\boldsymbol{w}) \coloneqq L'(\boldsymbol{w}; \gamma)$. By Lemma 8 (main text), we have that for any choice of $\boldsymbol{w} \in \mathcal{W}$,

$$\Psi_{s,\gamma}\left(R(\boldsymbol{w}) - R^*\right) \leq R_\varphi(\boldsymbol{w}) - R^*_\varphi.$$

To control the right-hand side, note that by strong convexity $\boldsymbol{w}^*$ is the unique minimum of $R_\varphi$, and so $R_\varphi(\boldsymbol{w}) - R^*_\varphi = R_\varphi(\boldsymbol{w}) - R_\varphi(\boldsymbol{w}^*)$. Since $R_\varphi$ is smooth via Lemma 2 with coefficient $v_X$, using the basic property (1) of smooth functions, we have

$$R_\varphi(\boldsymbol{w}) - R_\varphi(\boldsymbol{w}^*) \leq v_X \|\boldsymbol{w} - \boldsymbol{w}^*\|^2.$$

It remains to control $\|\widehat{\boldsymbol{w}}_{(t)} - \boldsymbol{w}^*\|$, where $\widehat{\boldsymbol{w}}_{(t)}$ is the output of a single iteration of the **for** loop in Algorithm 1 (main text). This can be broken up into computational and statistical elements, as follows. We can readily bound this distance from above as

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

The initial equality follows immediately by design of Algorithm 1 (main text) and the assumption that $\alpha_{(t)} = \alpha$. Using the triangle yields the upper bound, which is composed of two terms: the first term is the difference after doing one update of the ideal gradient descent routine (to minimize $R_\varphi$), and the second term is a statistical error term for the empirical mean estimate of the surrogate risk gradient vector.

For small enough step size $0 < \alpha < 2/(\kappa + v_X)$, the update improves on the previous error as

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

Writing $\beta \coloneqq 2\kappa v_X/(\kappa + v_X)$, we have that

$$\|\widehat{\boldsymbol{w}}_{(t+1)} - \boldsymbol{w}^*\| \leq \sqrt{1 - \alpha\beta}\|\widehat{\boldsymbol{w}}_{(t)} - \boldsymbol{w}^*\| + \alpha\|\widehat{\boldsymbol{g}}(\widehat{\boldsymbol{w}}_{(t)}) - \boldsymbol{g}(\widehat{\boldsymbol{w}}_{(t)})\|. \tag{13}$$

This deals with the computational error part. Now for the statistical error part, namely the accuracy of the $\widehat{\boldsymbol{g}} \approx \boldsymbol{g}$ approximation. Writing $\boldsymbol{b}(\boldsymbol{w}) \coloneqq -\rho((\gamma - y\langle \boldsymbol{w}, \boldsymbol{x}\rangle)/s)\, y\, \boldsymbol{x}$ and multiplying this by $s$, the sub-Gaussianity assumption of A3 (main text) gives us that for any $\boldsymbol{w} \in \mathbb{R}^d$ and $a \geq 0$,

$$\begin{aligned}\mathbf{E}\exp(a\langle \boldsymbol{u}, s\,\boldsymbol{b}(\boldsymbol{w}) - \mathbf{E}\,s\,\boldsymbol{b}(\boldsymbol{w})\rangle) &= \mathbf{E}\exp((as)\langle \boldsymbol{u}, \boldsymbol{b}(\boldsymbol{w}) - \mathbf{E}\,\boldsymbol{b}(\boldsymbol{w})\rangle) \\ &\leq \exp(c(as)^2\langle \boldsymbol{u}, \Sigma(\boldsymbol{w})\boldsymbol{u}\rangle) \\ &= \exp((cs^2)a^2\langle \boldsymbol{u}, \Sigma(\boldsymbol{w})\boldsymbol{u}\rangle)\end{aligned}$$

for all $\|\boldsymbol{u}\| = 1$, where $\Sigma(\boldsymbol{w})$ is the covariance matrix of $\boldsymbol{b}(\boldsymbol{w})$, and $c > 0$ is any constant such that the sub-Gaussian property holds. Now, suppressing $\boldsymbol{w}$ from the notation for readability, noting that

$$\begin{aligned}\Sigma &= \mathbf{E}(\boldsymbol{b} - \mathbf{E}\,\boldsymbol{b})(\boldsymbol{b} - \mathbf{E}\,\boldsymbol{b})^T \\ &= \mathbf{E}\,\boldsymbol{b}\boldsymbol{b}^T - (\mathbf{E}\,\boldsymbol{b})(\mathbf{E}\,\boldsymbol{b})^T,\end{aligned}$$

and using the positive semi-definiteness of $(\mathbf{E}\,\boldsymbol{b})(\mathbf{E}\,\boldsymbol{b})^T$, since for all $\boldsymbol{u}$ we have

$$\langle \boldsymbol{u}, (\mathbf{E}\,\boldsymbol{b}\boldsymbol{b}^T - \Sigma)\boldsymbol{u}\rangle = \langle \boldsymbol{u}, (\mathbf{E}\,\boldsymbol{b})(\mathbf{E}\,\boldsymbol{b})^T\boldsymbol{u}\rangle \geq 0,$$

for each $a \geq 0$ we can then bound

$$\begin{aligned}\mathbf{E}\exp(a\langle \boldsymbol{u}, \boldsymbol{b} - \mathbf{E}\,\boldsymbol{b}\rangle) &\leq \exp(cs^2a^2\langle \boldsymbol{u}, \Sigma\boldsymbol{u}\rangle) \\ &\leq \exp(cs^2a^2\langle \boldsymbol{u}, (\mathbf{E}\,\boldsymbol{b}\boldsymbol{b}^T)\boldsymbol{u}\rangle) \\ &\leq \exp(cs^2a^2\|\,\mathbf{E}\,\boldsymbol{b}\boldsymbol{b}^T\|) \\ &\leq \exp(cs^2a^2\,\mathbf{E}\,\|\boldsymbol{b}\boldsymbol{b}^T\|) \\ &\leq \exp(cs^2a^2\rho'(\sqrt{2})^2\,\mathbf{E}\,\|\boldsymbol{x}\boldsymbol{x}^T\|)\end{aligned}$$

With these inequalities, writing

$$\boldsymbol{b}_i(\boldsymbol{w}) \coloneqq -s\,\rho'\left(\frac{\gamma - y_i\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle}{s}\right) y_i\,\boldsymbol{x}_i, \quad i \in [n]$$

for any fixed $\boldsymbol{w}$ and noting that $\widehat{\boldsymbol{g}}(\boldsymbol{w}) = n^{-1}\sum_{i=1}^n \boldsymbol{b}_i(\boldsymbol{w})$, we can leverage Lemma 3 to prove that the "bad event"

$$\mathcal{E}(\boldsymbol{w}) \coloneqq \left\{\|\widehat{\boldsymbol{g}}(\boldsymbol{w}) - \boldsymbol{g}(\boldsymbol{w})\| > 2s\sqrt{\frac{c\rho'(\sqrt{2})^2\,\mathbf{E}\,\|\boldsymbol{x}\boldsymbol{x}^T\|\log(\delta^{-1})}{n}}\right\} \tag{14}$$

has probability $\mathbf{P}\,\mathcal{E}(\boldsymbol{w}) \leq \delta$.

Next, we must deal with the fact that in running Algorithm 1 (main text), the $\widehat{\boldsymbol{w}}_{(t)}$ for all $t > 0$ will be random and dependent on the sample. In general, there is not much choice but to pursue uniform bounds, namely high-probability events that hold over all $\boldsymbol{w} \in \mathcal{W}$. To do this is straightforward with an $\epsilon$-cover of $\mathcal{W}$. Since $\mathcal{W}$ is a compact subset of $\mathbb{R}^d$ by assumption A0 (main text), it follows that the size of an $\epsilon$-cover in the usual norm is bounded as $N_\epsilon \leq (3\Delta/2\epsilon)^d$ [7]. Denote the centers of the $\epsilon$ balls covering $\mathcal{W}$ by $\{\widetilde{\boldsymbol{w}}_1, \ldots, \widetilde{\boldsymbol{w}}_{N_\epsilon}\}$. Given any arbitrary $\boldsymbol{w} \in \mathcal{W}$, write $\widetilde{\boldsymbol{w}} = \widetilde{\boldsymbol{w}}(\boldsymbol{w})$ for the center closest to $\boldsymbol{w}$, which by definition satisfies $\|\boldsymbol{w} - \widetilde{\boldsymbol{w}}\| \leq \epsilon$. The statistical error can be bounded above by

$$\|\widehat{\boldsymbol{g}}(\boldsymbol{w}) - \boldsymbol{g}(\boldsymbol{w})\| \leq \|\widehat{\boldsymbol{g}}(\boldsymbol{w}) - \widehat{\boldsymbol{g}}(\widetilde{\boldsymbol{w}})\| + \|\boldsymbol{g}(\boldsymbol{w}) - \boldsymbol{g}(\widetilde{\boldsymbol{w}})\| + \|\widehat{\boldsymbol{g}}(\widetilde{\boldsymbol{w}}) - \boldsymbol{g}(\widetilde{\boldsymbol{w}})\|. \tag{15}$$

We want to take the supremum of both sides with respect to $\boldsymbol{w} \in \mathcal{W}$. Let's take it term by term.

Starting with the first term, by the 1-Lipschitz property of $\rho'$, it follows immediately that we can bound

$$\|\boldsymbol{b}_i(\boldsymbol{w}) - \boldsymbol{b}_i(\widetilde{\boldsymbol{w}})\| \leq \frac{\|\boldsymbol{x}_i\|\|\boldsymbol{w} - \widetilde{\boldsymbol{w}}\|}{s} \leq \frac{\|\boldsymbol{x}_i\|\epsilon}{s}.$$

This implies that

$$\|\widehat{\boldsymbol{g}}(\boldsymbol{w}) - \widehat{\boldsymbol{g}}(\widetilde{\boldsymbol{w}})\| \leq \left(\frac{\epsilon}{s}\right)\frac{s}{n}\sum_{i=1}^n \|\boldsymbol{x}_i\|^2 \leq \frac{\epsilon\,\mathbf{E}\,\|\boldsymbol{x}\|^2}{\delta} = \frac{\epsilon v_X}{\delta} \tag{16}$$

on an event of probability no less than $1-\delta$, where we have simply used Chebyshev's inequality to obtain tail bounds. Since regardless of what $\boldsymbol{w}$ we choose, the corresponding $\widetilde{\boldsymbol{w}}$ will be no farther than $\epsilon$, this (16) gives us a uniform bound.

For the second term, we just use the $v_X$-smoothness of $R_\varphi$, shown in Lemma 2. This implies

$$\|\boldsymbol{g}(\boldsymbol{w}) - \boldsymbol{g}(\widetilde{\boldsymbol{w}})\| \leq v_X\|\boldsymbol{w} - \widetilde{\boldsymbol{w}}\| \leq v_X\epsilon \tag{17}$$

again for arbitrary choice of $\boldsymbol{w} \in \mathcal{W}$.

Finally, for any fixed $\widetilde{\boldsymbol{w}} \in \{\widetilde{\boldsymbol{w}}_1, \ldots, \widetilde{\boldsymbol{w}}_{N_\epsilon}\}$, we can bound the third term using (14). Making the dependence of $\widetilde{\boldsymbol{w}}$ on $\boldsymbol{w}$ explicit for clarity, the critical fact is that

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

The "good event" of interest here is the event in which the bad event does not occur at any of the $\epsilon$-cover centers, that is

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

where $(\cdot)^c$ denotes the complement event. It thus follows that taking a union bound, we have that with probability no less than $1-\delta$, we can uniformly bound as

$$\|\widehat{\boldsymbol{g}}(\widetilde{\boldsymbol{w}}(\boldsymbol{w})) - \boldsymbol{g}(\widetilde{\boldsymbol{w}}(\boldsymbol{w}))\| \leq 2s\sqrt{\frac{c\rho'(\sqrt{2})^2\,\mathbf{E}\,\|\boldsymbol{x}\boldsymbol{x}^T\|\log(N_\epsilon\delta^{-1})}{n}}, \quad \forall\,\boldsymbol{w}\in\mathcal{W}. \tag{18}$$

Putting these three bounds together, and taking unions over the good events required for the first and third terms, we have with probability no less than $1-2\delta$ that

$$\sup_{\boldsymbol{w}\in\mathcal{W}}\|\widehat{\boldsymbol{g}}(\boldsymbol{w}) - \boldsymbol{g}(\boldsymbol{w})\| \leq \frac{\epsilon v_X}{\delta} + v_X\epsilon + 2s\sqrt{\frac{c\rho'(\sqrt{2})^2\,\mathbf{E}\,\|\boldsymbol{x}\boldsymbol{x}^T\|\log(N_\epsilon\delta^{-1})}{n}}.$$

Setting $\epsilon = \delta/\sqrt{n}$, this simplifies to

$$\sup_{\boldsymbol{w}\in\mathcal{W}}\|\widehat{\boldsymbol{g}}(\boldsymbol{w}) - \boldsymbol{g}(\boldsymbol{w})\| \leq \frac{(1+\delta)v_X}{\sqrt{n}} + \frac{\varepsilon^*}{\sqrt{n}}.$$

where we have defined

$$\varepsilon^* := 2s\sqrt{c\rho'(\sqrt{2})^2\,\mathbf{E}\,\|\boldsymbol{x}\boldsymbol{x}^T\|(d\log(3\sqrt{n}(2\delta)^{-1}) + \log(\delta^{-1}))}. \tag{19}$$

On the good event $\mathcal{E}_+$, then denoting $\varepsilon = (\varepsilon^* + (1+\delta)v_X)/\sqrt{n}$, for *all* steps $t$ we can re-write (13) as

$$\|\widehat{\boldsymbol{w}}_{(t+1)} - \boldsymbol{w}^*\| \leq \sqrt{1-\alpha\beta}\|\widehat{\boldsymbol{w}}_{(t)} - \boldsymbol{w}^*\| + \alpha\varepsilon.$$

Assuming the algorithm is run for $T$ updates, then with some straightforward algebra we can unfold and clean up the recursion such that

$$\|\widehat{\boldsymbol{w}}_{(T)} - \boldsymbol{w}^*\| \leq (\sqrt{1-\alpha\beta})^T\|\widehat{\boldsymbol{w}}_{(0)} - \boldsymbol{w}^*\| + \frac{2\varepsilon}{\beta}.$$

Let us connect all the inequalities now. We can bound the excess surrogate risk as

$$R_\varphi(\widehat{\boldsymbol{w}}_{(T)}) - R_\varphi^* \leq \left((1-\alpha\beta)^T\|\widehat{\boldsymbol{w}}_{(0)} - \boldsymbol{w}^*\|^2 + \frac{4}{\beta^2 n}\left((1+\delta)v_X + \varepsilon^*\right)^2\right)v_X$$

which via the first inequality of this proof using $\Psi_{s,\gamma}$, yields the desired result. □

### A.3 Explanation of root-finding function `getroot`

We have uploaded some demonstrative software for the root-finding sub-routine needed for computation of $\Psi_{s,\gamma}$ to a public repository.[2] The core routine is captured in a function called `getroot`, using a very simple strategy, which we describe below.

The cubic polynomials considered are equations of the form

$$au^3 + bu^2 + cu + d = 0. \tag{20}$$

Recall the discriminant $\Delta$ given in (7). There are a few basic settings to consider, as below.

- If $\Delta < 0$, then there is only one real root (the rest are complex).
- If $\Delta = 0$, then all roots are real, but we have multiple roots.

[2] Available at: https://github.com/feedbackward/catcube.

- If $\Delta > 0$, then all roots are real, and distinct.

For the case of $\Delta < 0$, the traditional solution approach is as follows. Defining two new quantities

$$\Delta_0 := b^2 - 3ac$$
$$\Delta_1 := 2b^3 - 9abc + 27a^2d$$

the key value for computing roots is the following

$$C = \left( \frac{\Delta_1 \pm \sqrt{\Delta_1^2 - 4\Delta_0^3}}{2} \right)^{1/3}.$$

Assuming that $C$ is known, then roots are computed as

$$u^* = -\frac{1}{3a}\left(b + C + \frac{\Delta_0}{C}\right).$$

Naturally, if $C$ is real, then so is the resulting $u^*$. Computationally, how do we go about getting a real version? This is extremely straightforward. Let's take the addition case. Consider the condition

$$\Delta_1 + \sqrt{\Delta_1^2 - 4\Delta_0^3} \geq 0.$$

If this condition holds, then we can just compute as-is. If this condition fails to hold, then taking the cube root in many programming languages will lead to a complex number. To get a real number when the above condition fails, just compute

$$C = (-1)\left( \frac{|\Delta_1 + \sqrt{\Delta_1^2 - 4\Delta_0^3}|}{2} \right)^{1/3}.$$

With a real-valued $C$ in hand, $u^*$ immediately follows.

Next, consider the case of $\Delta = 0$. This case is very simple. This scenario also sub-divides, based on the value of $\Delta_0$. If $\Delta_0 = 0$, then the root is a "triple" root, and takes the form

$$u_T^* = -\frac{b}{3a}.$$

If $\Delta_0 \neq 0$, then we have two roots, a "double" root $u_D^*$ and a "single" root $u_S^*$, with the forms

$$u_D^* = \frac{9ad - bc}{2\Delta_0}$$
$$u_S^* = \frac{4abc - 9a^2d - b^3}{a\Delta_0}.$$

Finally, consider the case of $\Delta > 0$. For elegant computations, we make use of the trigonometric method pioneered by F. Viète. The starting point is a trigonometric identity, as follows:

$$\cos 3x = 4\cos^3 x - 3\cos x. \tag{21}$$

To prove this is straightforward. Making use of elementary trigonometric identities, observe first that

$$\cos 3x = \cos(2x + x) = \cos 2x \cos x - \sin 2x \sin x.$$

Looking at each of the terms individually,

$$\begin{aligned}\cos 2x \cos x &= (2\cos^2 x - 1)\cos x \\ &= 2\cos^3 x - \cos x \\ \sin 2x \sin x &= 2 \sin x \cos x \sin x \\ &= 2 \cos x \sin^2 x \\ &= 2 \cos x (1 - \cos^2 x)\end{aligned}$$

Taking the difference of the two new forms gives the desired identity (21). With this identity now at our disposal, we proceed with cleaning up the cubic equation. Dividing out $a$, and replacing $u$ with $v - b/(3a)$, note that this cleans up to

$$v^3 + pv + q = 0 \tag{22}$$

where

$$\begin{aligned}p &= \frac{3ac - b^2}{3a^2} \\ q &= \frac{2b^3 - 9abc + 27a^2 d}{27a^3}.\end{aligned}$$

Note that since we are assuming $\Delta > 0$ for the original cubic equation (20), which implies three distinct real roots for (20), it follows that (22) also has three distinct real roots. This can only happen when its discriminant is positive, which is to say when

$$-4p^3 - 27q^2 > 0. \tag{23}$$

Note that this implies

$$p^3 < -\frac{27q^2}{4} \leq 0.$$

This implies that $p < 0$, otherwise the cube of $p$ would necessarily be non-negative. Moving forward, considering the trigonometric identity (21), the desired form of our cubic equation is $4z^3 - 3z = e$, where $|e| \leq 1$ so that it falls in the range of the cosine function. To aid computations, let us introduce a couple more variables and coefficients. Set $k$ such that $p = -3k^2$, multiply by 4, and replace $v$ with $rz$, where $r$ is a coefficient to be defined shortly. Doing so, we have

$$\begin{aligned}0 &= 4(rz)^3 + 4(-3k^2)rz + 4q \\ &= 4z^3 - \frac{12k^3}{r^2}z + \frac{4q}{r^3}.\end{aligned}$$

Setting $r = 2k$, we can clean up into the following equation

$$4z^3 - 3z = -\frac{q}{2k^3}, \tag{24}$$

which is the desired form, as long as the right-hand side has absolute value no greater than unity. Fortunately, this is immediately true from our assumptions. To see this, first observe

$$\left(\frac{q}{2k^3}\right)^2 = \frac{-27q^2}{4p^3} = \frac{27q^2}{4|p|^3}$$

and recall that from (23) and the fact that $p < 0$, it follows that

$$0 < -27q^2 - 4p^3 = -27q^2 + 4|p|^3$$

which implies

$$1 > \frac{27q^2}{4|p|^3} = \left(\frac{q}{2k^3}\right)^2.$$

Thus, we have that in our current case of $\Delta > 0$, the right-hand side of (24) indeed falls on the interval $(-1, 1)$. As such, this means that there exists an angle $x^*$ such that plugging $\cos x^*$ into the polynomial (24), we have

$$4\cos^3 x^* - 3\cos x^* = -\frac{q}{2k^3}.$$

Then using the key identity (21), it follows that

$$\cos(3x^*) = -\frac{q}{2k^3}, \text{ implying } x^* = \frac{1}{3}\arccos\left(-\frac{q}{2k^3}\right).$$

So, we have that $\cos x^*$ solves (24). Note that since this function has period $2\pi$, it holds that

$$\cos(3x^*) = \cos(2\pi + 3x^*) = \cos(2\pi - 3x^*)$$

which after plugging in to (21), yields

$$\cos\left(3\left(\frac{2\pi}{3} \pm x^*\right)\right) = 4\cos^3\left(\frac{2\pi}{3} \pm x^*\right) - 3\cos\left(\frac{2\pi}{3} \pm x^*\right) = -\frac{q}{2k^3}.$$

That is to say, the following values are all solutions to (24):

$$z_1^* = \cos(x^*), \quad z_2^* = \cos\left(\frac{2\pi}{3} + x^*\right), \quad z_3^* = \cos\left(\frac{2\pi}{3} - x^*\right).$$

With these values in hand, all that remains is to backtrack to the roots of the original cubic polynomial of interest. For any $z_j^*$, this is done as

$$y_j^* = 2kz_j^*$$
$$u_j^* = y_j^* - \frac{b}{3a}$$

where $j = 1, 2, 3$. To summarize the case of finding roots when $\Delta > 0$, the basic computational procedure is as below.

1. From original polynomial, $(a, b, c, d) \mapsto (p, q)$.
2. From new polynomial, $(p, q) \mapsto k \mapsto e$, where $e := -q/(2k^3)$.
3. From final polynomial, $e \mapsto x^* \mapsto (z_1^*, z_2^*, z_3^*)$.
4. Backtrack over the roots as $z_j^* \mapsto y_j^* \mapsto u_j^*$ for $j = 1, 2, 3$.
5. Return $(u_1^*, u_2^*, u_3^*)$ as roots of (20).

This concludes our exposition of the content of `getroot`.

## References


[1] Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

[2] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: a nonasymptotic theory of independence*. Oxford University Press.

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

[5] Holland, M. J. and Ikeda, K. (2017). Efficient learning with robust gradient descent. *arXiv preprint arXiv:1706.00182*.

[6] Horn, R. A. and Johnson, C. R. (2013). *Matrix Analysis*. Cambridge University Press, 2nd edition.

[7] Kolmogorov, A. N. (1993). $\varepsilon$-entropy and $\varepsilon$-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.

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