Classification using concentrated margin pursuit

Matthew J. Holland[*]

Osaka University
Yamada-oka 2-8, Suita, Osaka

Abstract

Training by empirical risk minimization is a popular strategy, but when unbounded surrogate functions are used for classification, many observations may be required in order to generalize well. Countless algorithms with robustness to the data distribution have been proposed, but typically leave a significant gap between the procedure for which appealing guarantees hold, and the procedure that can actually be *implemented.* To try and close this gap as cheaply as possible, we propose an algorithm which searches the hypothesis space in such a way that a pre-set "margin level" ends up being a highly robust estimator of the location of the margin distribution. The procedure is easily implemented using gradient descent, and admits finite-sample bounds on the excess risk. Empirical tests on real-world benchmark data reinforce the basic principles highlighted by the theory, and are suggestive of a promising new technique for classification.

1 Introduction

To effectively carry out any machine learning application, one requires procedures for statistical inference which are sufficiently reliable, and implementations of these procedures which are computationally efficient. This setting is often formulated using an expected loss, or risk $R(\boldsymbol{w}) \coloneqq \mathbf{E}\, l(\boldsymbol{w};\boldsymbol{z})$, where $\boldsymbol{w}$ is a "parameter" (typically a vector, function, set, etc.), and expectation is taken over the unknown data distribution. As far as performance is concerned, given observations $\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n$, should a learning procedure yield $\widehat{\boldsymbol{w}}$ such that the risk $R(\widehat{\boldsymbol{w}})$ realized is small with high confidence over the sampling distribution, this is taken as formal evidence for good generalization, up to computational costs and assumptions on the underlying distribution. When considering the efficiency of learning, the statistical side is relevant since R is always unknown, and the implementation is relevant since in practice the only $\widehat{\boldsymbol{w}}$ we will ever obtain is one that can be computed within budget.

Our primary focus in this paper is a binary classification algorithm which is both practical from a computational standpoint, and simultaneously admits strong statistical performance guarantees that depend on the sample size, data distribution, and the computational cost incurred.

Binary classification, in which $\boldsymbol{z} = (\boldsymbol{x}, y) \in \mathcal{X} \times \{-1, 1\}$ and $l(\boldsymbol{w};\boldsymbol{z}) = I\{\operatorname{sign}(h(\boldsymbol{x};\boldsymbol{w})) \neq y\}$, is the canonical supervised learning task [16, 11]. Empirical risk minimization (ERM), which admits any minimizer of $n^{-1}\sum_{i=1}^{n} l(\cdot;\boldsymbol{z}_i)$, is the canonical learning strategy [3, 1, 4]. Since optimizing the empirical mean of the zero-one loss is not computationally tractable, surrogate functions are typically minimized instead. This alleviates one computational issue, but opens the learner up to severe limitations in the statistical efficiency of ERM, which is

[*] Email: `matthew-h@ids.osaka-u.ac.jp`.

sensitive to the method of implementation [10]. In particular, when unbounded losses are utilized, as is common in classification (e.g., most convex potential functions), then unless one has a computational procedure that can deal with errant observations, it may take an infinitely large sample in order to guarantee a small risk [12].

Review of related work Many procedures offering some notion of distributional robustness have been proposed in the literature. Some methods include sub-routines for explicitly identifying and discarding outliers [22], but can incur a large bias under well-behaved data and smaller data sets. Methods using non-convex potential functions are provably robust to label noise [17, 13], but are computationally expensive and lack stability from sample to sample. When we have prior knowledge of how flipped labels may lead to outlying instances, novel loss functions have been proposed [19, 23], but the risk bounds given by Natarajan et al. [19] tacitly assume that the ERM solution can be obtained, and do not reflect any explicit computational procedure. A novel technique was studied by Brownlees et al. [6], who forge robust estimates of the risk to create a new objective to be minimized; the procedure has strong guarantees under both heavy-tailed and more docile data, but implementation is highly non-trivial, and extending their guarantees to a practical computational procedure appears intractable. Recent work on using robust estimates of the risk gradient in a steepest descent procedure offers statistical guarantees implementable procedures [14, 9], but with the downside of potentially large computational overhead as the dimension grows.

Our contributions To deal with the limitations of existing procedures highlighted above, the key idea here is to introduce a new convex loss that encourages the distribution of the *margin* induced by $\widehat{\boldsymbol{w}}$, namely $y\,h(\boldsymbol{x};\widehat{\boldsymbol{w}})$, to be tightly concentrated near a certain prescribed level. The procedure is easily implemented using gradient descent, admits formal performance guarantees reflecting computational cost and optimization error, and aside from the usual cost of gradient computation there is virtually no computational overhead. Our main contributions:

- A novel classification method based on a convex surrogate that enables control of the margin distribution. It is easily implemented, and can be adapted to stochastic subsampling for larger tasks.
- High-probability bounds on excess risk of the proposed procedure, reflecting both statistical and computational factors, under mild assumptions on the distribution of the loss gradient.
- Numerical experiments using real-world benchmark data sets, in which we analyze the trajectory of the margin distribution, the surrogate and classification risks, comparing these metrics with those for the well-known Pegasos algorithm [21].

2 Concentrated margin pursuit

As a starting point to motivate a general learning algorithm, recall the simplest support vector machine procedure for binary classification: a hyperplane $\{\boldsymbol{u} : \langle \boldsymbol{w}, \boldsymbol{u}\rangle = 0\}$ to be used as a decision boundary is determined, given sample $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, by minimizing the following objective:

$$\widehat{R}(\boldsymbol{w}) = \|\boldsymbol{w}\|^2 + \frac{1}{n}\sum_{i=1}^{n} \max\left(0, 1 - y_i \langle \boldsymbol{w}, \boldsymbol{x}_i\rangle\right). \tag{1}$$

Besides the traditional margin maximization interpretation, if we write $\mathcal{E}_{\text{bad}} \coloneqq \{1 > y \langle \boldsymbol{w}, \boldsymbol{x} \rangle\}$ for the "bad event" of falling below the prescribed margin level, any minimizer of (1) naturally seeks to make $\mathbf{E}_\mu (y \langle \boldsymbol{w}, \boldsymbol{x} \rangle | \mathcal{E}_{\text{bad}}) \, \mathbf{P}(\mathcal{E}_{\text{bad}})$ small, under norm constraints. What can we say about the distribution of the margin $y \langle \boldsymbol{w}, \boldsymbol{x} \rangle$? Fixing any candidate $\boldsymbol{w}$, then taking expectation over the sample, if $\boldsymbol{w}$ achieves risk of $R(\boldsymbol{w}) \coloneqq \mathbf{E} \, \widehat{R}(\boldsymbol{w}) = \|\boldsymbol{w}\|^2 + \operatorname{Err}(\boldsymbol{w})$, this immediately implies

$$1 - \mathbf{E}_\mu \, y \langle \boldsymbol{w}, \boldsymbol{x} \rangle \leq \operatorname{Err}(\boldsymbol{w}). \tag{2}$$

Unfortunately this is not immediately of any use since in practice $\boldsymbol{w}$ will never be pre-fixed, but rather will be the output of some learning algorithm, written $\widehat{\boldsymbol{w}}$, and will be sample-dependent. A more powerful kind of control on the distribution than is possible with (2) would be as follows: taking probability with respect to the sample, with accuracy $\varepsilon > 0$ and confidence $\delta \in (0, 1)$ we desire

$$\mathbf{P} \left\{ 1 - \mathbf{E}_\mu \, y \langle \widehat{\boldsymbol{w}}, \boldsymbol{x} \rangle \leq \varepsilon \right\} \geq 1 - \delta. \tag{3}$$

How can we weave this condition into a learning algorithm? To approximate this, we make use of the following even (symmetric about 0), convex, and continuously differentiable function:

$$\rho(u) \coloneqq \begin{cases} \frac{u^2}{2} - \frac{u^4}{24} & |u| \leq \sqrt{2}, \\ |u| \frac{2\sqrt{2}}{3} - \frac{1}{2} & |u| > \sqrt{2}. \end{cases} \tag{4}$$

This function is obtained by integration of the influence function used by Catoni and Giulini [8], behaving like the quadratic function around the origin, while enjoying a bounded first derivative. Writing $\rho_s(u) \coloneqq \rho(u/s)$ for re-scaling by $s > 0$ and fixing arbitrary $\boldsymbol{w}$, the Catoni-type estimate of the margin location is defined

$$\widehat{\gamma} \coloneqq \operatorname*{arg\,min}_{\gamma \in \mathbb{R}} \sum_{i=1}^{n} \rho_s \left(\gamma - y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle \right). \tag{5}$$

Intuitively, a better choice of $\boldsymbol{w}$ (for the classification task) will lead to a larger value of $\widehat{\gamma}$. More formally, for any candidate $\boldsymbol{w}$ it follows that

$$\widehat{\gamma} - \mathbf{E}_\mu \, y \langle \boldsymbol{w}, \boldsymbol{x} \rangle \leq \frac{c \operatorname{var}_\mu y \langle \boldsymbol{w}, \boldsymbol{x} \rangle}{s} + \frac{s \log(2\delta^{-1})}{n}$$

with probability no less than $1 - \delta$. Under the weak assumption of a bound $\operatorname{var}_\mu y \langle \boldsymbol{w}, \boldsymbol{x} \rangle \leq v$ this flips the randomization of (3), and says that over the random draw of the sample, the location of the margin distribution (reflected by $\mathbf{E}_\mu \, y \langle \boldsymbol{w}, \boldsymbol{x} \rangle$) will not fall very much below the data-dependent margin level specified by $\widehat{\gamma}$. To integrate this into our learning procedure, a natural modification to the loss function of (1) is given as

$$\widehat{R}_\gamma(\boldsymbol{w}) = \|\boldsymbol{w}\|^2 + \frac{s}{n} \sum_{i=1}^{n} \rho_s \left(\gamma - y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle \right). \tag{6}$$

The idea is that if $\widehat{\boldsymbol{w}}$ minimizes $\widehat{R}_\gamma(\cdot)$ for a sufficiently large value of γ, the resulting Catoni-type estimate of $\mathbf{E}_\mu \, y \langle \widehat{\boldsymbol{w}}, \boldsymbol{x} \rangle$ should be sharply concentrated in a region beyond unity, implying that the margin distribution realized by $\widehat{\boldsymbol{w}}$ should be located in a region conducive to off-sample generalization. This procedure can be readily implemented using a steepest descent type of optimizer: we give a general version of such a procedure in Algorithm 1.

The chief benefit of this procedure is that it enjoys strong learning guarantees (section 3), under weak assumptions on the data distribution, while still being easily implemented and useful in practice (section 4). We can thus effectively bridge the gap between the algorithm for which statistical guarantees hold, and the algorithm used by the practitioner.

Algorithm 1 Concentrated margin pursuit by steepest descent.

```
input: $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \in \mathbb{R}^d \times \{-1, 1\}$
parameters: $\widehat{\boldsymbol{w}}_{(0)} \in \mathbb{R}^d$, $\lambda \geq 0$, $\gamma \in \mathbb{R}$, $\alpha > 0$
scaling: $s \geq \sqrt{\dfrac{n \, \mathbf{E}_\mu \|\boldsymbol{x}\|^2}{2\lambda \log(\delta^{-1})}}$
for $t = 0, 1, \ldots, T-1$ do
    $\widehat{\boldsymbol{w}}_{(t+1)} \leftarrow (1 - \lambda\alpha) \, \widehat{\boldsymbol{w}}_{(t)} + \dfrac{\alpha}{n} \sum_{i=1}^{n} \rho'_s \left( \gamma - y_i \langle \widehat{\boldsymbol{w}}_{(t)}, \boldsymbol{x}_i \rangle \right) y_i \boldsymbol{x}_i$
    $\widehat{\boldsymbol{w}}_{(t+1)} \leftarrow \min \left( 1, \dfrac{(1/\sqrt{\lambda})}{\|\widehat{\boldsymbol{w}}_{(t+1)}\|} \right) \widehat{\boldsymbol{w}}_{(t+1)}$
end for
```

3 Theoretical analysis

Here we analyze some basic statistical and computational properties of the core procedure highlighted in section 2. First a summary of our notation is given, followed by representative results with explanatory remarks. All detailed proofs are relegated to the supplementary materials.

Notation Let μ denote the data distribution of random vector $(\boldsymbol{x}, y)$, here taking values on $\mathbb{R}^d \times \{-1, 1\}$. The data sample refers to n independent and identically distributed copies of $(\boldsymbol{x}, y)$, denoted $(\boldsymbol{x}_i, y_i)$, for $i \in [n]$. We will utilize a more general version of (5) as follows:

$$\widehat{\gamma}(h) = \operatorname*{arg\,min}_{\gamma \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \rho_s \left(\gamma - y_i \, h(\boldsymbol{x}_i) \right), \tag{7}$$

where the dependence on $s > 0$ is supressed in the notation, and $h : \mathbb{R}^d \to \mathbb{R}$ is a general classifier; setting $h(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$ yields the special case highlighted in section 2. Write $\mathcal{H}$ for the hypothesis class containing all h being considered, most generally $\mathcal{H} = \{h : \mathbf{E}_\mu |h(\boldsymbol{x})|^2 < \infty\}$ but special cases will also be of interest. For integer k, write $[k] := \{1, \ldots, k\}$ for all the positive integers from 1 to k. We shall be rather free in our usage of the symbol $\mathbf{P}$ as a generic probability measure; in most cases, this will be the product measure induced by the sample, but in other cases it will represent μ. The measure being used will be made clear from the context. Finally, we shall write $v_X := \mathbf{E}_\mu \|\boldsymbol{x}\|^2$.

To begin, we show that proper control of scale can be used to smoothly control the location parameter that of the margin distribution that is being used to guide the learning procedure.

Proposition 1 (Properties of location level). *For any $h \in \mathcal{H}$ and scale $s > 0$, the estimate $\widehat{\gamma}$ defined in (7) satisfies the following:*

1. *There exists $0 < s' < \infty$ such that for all $0 < s \leq s'$, we have $\widehat{\gamma}(h) = \operatorname{med}\{y_i \, h(\boldsymbol{x}_i)\}_{i \in [n]}$.*

2. *There exists a constant $c > 0$ such that for all $s > 0$,*

$$\left| \widehat{\gamma}(h) - \frac{1}{n} \sum_{i=1}^{n} y_i \, h(\boldsymbol{x}_i) \right| \leq \frac{c}{s^2}.$$

3. Scaling with $s^2 = (n \operatorname{var}_\mu y\, h(\boldsymbol{x}))/2\log(\delta^{-1})$*, we have that*

$$\mathbf{P}\left\{ \widehat{\gamma}(h) - \mathbf{E}_\mu\, y\, h(\boldsymbol{x}) > \sqrt{\frac{2 \operatorname{var}_\mu y\, h(\boldsymbol{x}) \log(\delta^{-1})}{n}} \right\} \leq \delta.$$

Remark 2. Statements 1–3 can be applied to any iterative step in Algorithm 1, and tell us how setting s relatively small will encourage $\widehat{\gamma}$ to be a median estimator, while a larger s will encourage it to be a mean estimator, with the bias shown explicitly to be dependent on s. This becomes important when the underlying margin distribution is asymmetric (since the median and mean can diverge), and provides a natural mechanism for direct control of how tolerant the algorithm is with respect to errors (how many, and how large).

Remark 3 (Existence of performance gaps). When a classifier h is fixed and $\widehat{\gamma}$ is computed, all the guarantees above are valid; however, it should be intuitively clear that this does not always hold in the opposite direction. That is, if the γ level is set too high given the hypothesis class $\mathcal{H}$ at hand, we cannot expect γ to represent a good approximation of the location of the margin $y\, h(\boldsymbol{x})$. This can be easily proven: there exists a set of classifiers $\mathcal{H}$ and distribution μ under which even a perfect optimizer of the new risk has a Catoni-type estimate smaller than γ (proven in supplementary materials).

Proceeding with our analysis, the ultimate evaluation metric of interest here is the classification risk (expectation of the zero-one loss), denoted

$$R(h) \coloneqq \mathbf{P}\{\operatorname{sign}(h(\boldsymbol{x})) \neq y\}, \quad R^* \coloneqq \inf_{h \in \mathcal{H}} R(h). \tag{8}$$

Using empirical estimates of the zero-one loss is not conducive to efficient learning algorithms, and our Algorithm 1 involves the minimization of a new loss $\widehat{R}_\gamma(\cdot)$, defined in equation (6). To ensure that good performance in this metric implies low classification risk, the first step is to ensure that the function is *calibrated* for classification, in the sense of Bartlett et al. [2]. To start, we shall focus on this loss without regularization; for $\gamma > 0$, writing $\varphi(u) \coloneqq s\, \rho_s(\gamma - u)$, this furnishes the surrogate risk

$$R_\varphi(h) \coloneqq \mathbf{E}_\mu\, \varphi\left(y\, h(\boldsymbol{x})\right), \quad R_\varphi^* \coloneqq \inf_{h \in \mathcal{H}} R_\varphi(h). \tag{9}$$

The basic idea is that if this loss φ is calibrated, then one can show that there exists a function Ψ, non-decreasing on the positive real line, such that

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

Our choice of ρ defined in (4) is particularly congenial because in addition to allowing for guarantees of the form highlighted in Prop. 1, it also enjoys the fact that (a) it is classification-calibrated, and (b) the resulting Ψ can be computed exactly, for arbitrary values of γ. This computation is elementary but tedious, and the details are relegated to the appendix. We summarize the basic facts in the following lemma.

Lemma 4. *Using* $\varphi(u) = \rho(\gamma - u)$ *as a loss function, with* ρ *as defined in (4), the function is classification calibrated such that for each* γ *there exists a function* $\Psi_\gamma : [0,1] \to \mathbb{R}_+$ *for which* $\Psi_\gamma(R(h) - R^*) \leq R_\varphi(h) - R_\varphi^*$*, taking the form*

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

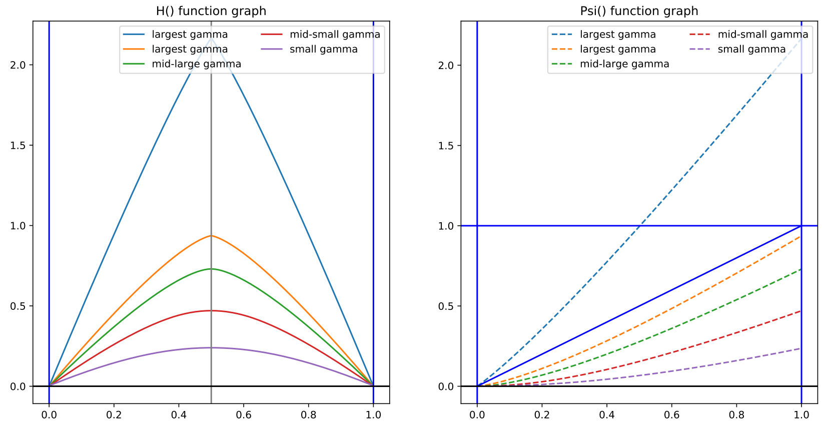

Figure 1: Graphs of $H_\gamma(\cdot)$ (left) and $\Psi_\gamma(\cdot)$ (right) over the unit interval for different settings of γ. Ranging from small to largest, these values are $\gamma = \sqrt{2}/2, \sqrt{2} - 0.4, \sqrt{2} - 0.11, \sqrt{2} + 0.11, 2\sqrt{2}$.

where $H_\gamma(\cdot)$ is a concave function defined on $[0,1]$, specified in the proof, plotted in Figure 1. Furthermore, given a sequence $(\widehat{h}_n)$ of sample-dependent $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\} \mapsto \widehat{h}_n$, we have that consistency in our surrogate is sufficient for consistency in classification risk:

$$\left\{\lim_{n\to\infty} R_\varphi(\widehat{h}) = R_\varphi^*\right\} \subseteq \left\{\lim_{n\to\infty} R(\widehat{h}) = R^*\right\}.$$

Finally, all these properties immediately extend to the general case of $\rho_s(\cdot)$, for any $s > 0$.

Remark 5 (Impact of γ level setting). One would naturally expect that all else equal, if a classifier achieves the same excess φ-risk for a larger value of γ, then the resulting excess classification risk should be smaller, or at least no larger. More concretely, we should expect that

$$\gamma \leq \gamma' \implies \Psi_\gamma^{-1}(a) \geq \Psi_{\gamma'}^{-1}(a), \quad a \in [0, \rho_s(\gamma)].$$

This monotonicity follows from properties of ρ (see Figure 1).

Now that we have a clear link between the risk being approximated by $\widehat{R}_\gamma$ defined in (6) and the classification risk, we can consider a practical implementation of a learning machine that minimizes this objective via steepest descent, which is precisely what Algorithm 1 does.

Next we pursue an excess risk bound for Algorithm 1, which consists of three simple steps: (a) re-scaling to control margin estimates and strong convexity of the loss, (b) a parameter update in the direction of steepest descent, and finally (c) a projection to the ℓ_2 ball of radius $1/\sqrt{\lambda}$. To keep the results concrete and interpretable, we focus on the model used in Algorithm 1, namely $h(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$. To make notation more transparent, we accordingly write $R(\boldsymbol{w})$ and $R_\varphi(\boldsymbol{w})$ to denote the respective risks under $\mathcal{H} = \{h : h(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle, \boldsymbol{w} \in \mathcal{W}\}$, where $\mathcal{W} = \{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\| \leq 1/\sqrt{\lambda}\}$. The basic assumptions on the data distribution are as follows:

A1 Assume that $R_\varphi(\boldsymbol{w})$ is κ-strongly convex on $\mathcal{W}$, with minimum[1] denoted by $\boldsymbol{w}^*$.

A2 Assume gradient $-\rho'(\gamma - y\langle \boldsymbol{w}, \boldsymbol{x} \rangle) y \, \boldsymbol{x}$ is sub-Gaussian.

With these assumptions in place, finite-sample risk bounds can be obtained.

[1] Assuming we can take the derivative under the integral, the smoothness of ρ implies differentiability of R_φ. Then using the compactness of $\mathcal{W}$, it follows that $\boldsymbol{w}^* \in \mathcal{W}$.

Theorem 6. *Under assumptions **A1**–**A2**, running Algorithm 1 for T iterations, the final output produced, written $\widehat{\boldsymbol{w}}_{(T)}$, for constant $c > 0$ and $\beta := 2\kappa v_X/(\kappa + v_X)$ satisfies*

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

with probability no less than $1 - 2\delta$ over the random draw of the sample, where the dominant term ε^* is defined

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

Remark 7 (Interpretation of bounds). There are several important terms and factors in the upper bound of Theorem 6. The first term is an optimization error term decreasing in T, followed by a statistical error term decreasing in sample size n. The former depends on the initial estimate $\widehat{\boldsymbol{w}}_{(0)}$, the step-size α, and the convexity of the surrogate risk through β. The statistical error term depends chiefly on the sample size, the dimension, and the second-order moments of the input $\boldsymbol{x}$. Finally, we note the extra d factor in ε^* is due to a covering number argument used to obtain a bound on the empirical gradient error that holds uniformly over $\boldsymbol{w} \in \mathcal{W}$. Does there exist another computational procedure, with the same rate of optimization error, and without this seemingly superfluous d factor in the statistical error? We pursue such analysis in future work.

4 Empirical analysis

The primary goal of our numerical experiments is to complement the theory of the previous section, by shedding light on how the classification procedure proposed in Algorithm 1 behaves over time, and how this depends on both parameter settings and the underlying learning task. In particular, we examine the trajectory of key performance metrics such as average classification error and our surrogate loss, as well as the location and shape of the margin distribution itself over many iterations.

Experimental setup In all the experiments discussed here, we consider binary classification on real-world data sets, modified to control for unbalanced ratios of positive and negative labels. Training for each data set is done using pair $(\boldsymbol{X}, \boldsymbol{y})$, where $\boldsymbol{X}$ is $n \times d$, and $\boldsymbol{y}$ is $n \times 1$, and testing is done on a disjoint subset. The train-test sequence is repeated over 25 trials, and all numerical performance metrics displayed henceforth should be assumed to be averages taken over all trials.

We use four data sets, denoted cov, digit5, protein, and sido, creating subsets under the following constraints: (1) Sample size n is no more than ten times the nominal dimension d, and (2) both the training and testing data sets have balanced ratios of labels (as close as possible to 50% each). Starting with cov ($n = 540$, $d = 54$, non-zero: 22%), this is the "Forest CoverType dataset" on the UC Irvine repository, converted into a binary task identifying class 1 against the rest. digit5 ($n = 5000$, $d = 784$, non-zero: 19%) is the MNIST hand-written digit data, converted into a binary task for the digit 5. protein ($n = 740$, $d = 74$, non-zero: 99%) is the protein homology dataset (KDD Cup 2004). sido ($n = 425$, $d = 4932$, non-zero: 11%) is the molecular descriptor data set (NIPS 2008 causality challenge), with binary-valued features.

As a well-known benchmark algorithm against which we can compare the behaviour and performance of the proposed Algorithm 1, we implement and run the well-known Pegasos algorithm of Shalev-Shwartz et al. [21]. For both methods, the initial value $\widehat{\boldsymbol{w}}_{(0)}$ is determined randomly in each trial. We explore multiple settings of Algorithm 1 described further below, but in all cases we take the stochastic optimization approach: instead of using all n training examples at each step, we randomly select one at a time for computing the update direction, and use a step size of $\alpha = (\sqrt{\lambda}(1+t))^{-1}$. Finally, for direct comparison with Pegasos, we set the margin level to $\gamma = 1$.

Trajectory comparison In Figure 2, we look at the trajectories of several important performance metrics, comparing our Algorithm 1 with the benchmark procedure. Reading left to right, in the first plot we have classification error on the training and test data. The second plot shows the mean, median, and 25th/75th percentiles of $\{y_i \langle \widehat{\boldsymbol{w}}_{(t)}, \boldsymbol{x}_i \rangle\}_i$ on the testing data for each time step $t = 0, 1, \ldots, T$, with colour filled in representing the analogous standard deviation. The third plot shows the same mean and median, but re-scaled by the standard deviation. The fourth plot shows the ℓ_2 norm of $\widehat{\boldsymbol{w}}_{(t)}$. For large t both procedures behave similarly, but there is a dramatic difference in the margin distribution early on, with the margin distribution being far more asymmetric and more widely dispersed in the Pegasos case, even under identical regularization. The smaller variance and higher degree of symmetry is precisely what we would expect given the definition of ρ in (4), since "overconfidence" is penalized both when the classification is correct *and* when it is incorrect. Algorithm 1 pushes the location of the distribution of $y \langle \widehat{\boldsymbol{w}}_{(T)}, \boldsymbol{x} \rangle$ towards γ in a stabler, more monotonic manner. Analogous trends can be observed across all the other data sets tested.

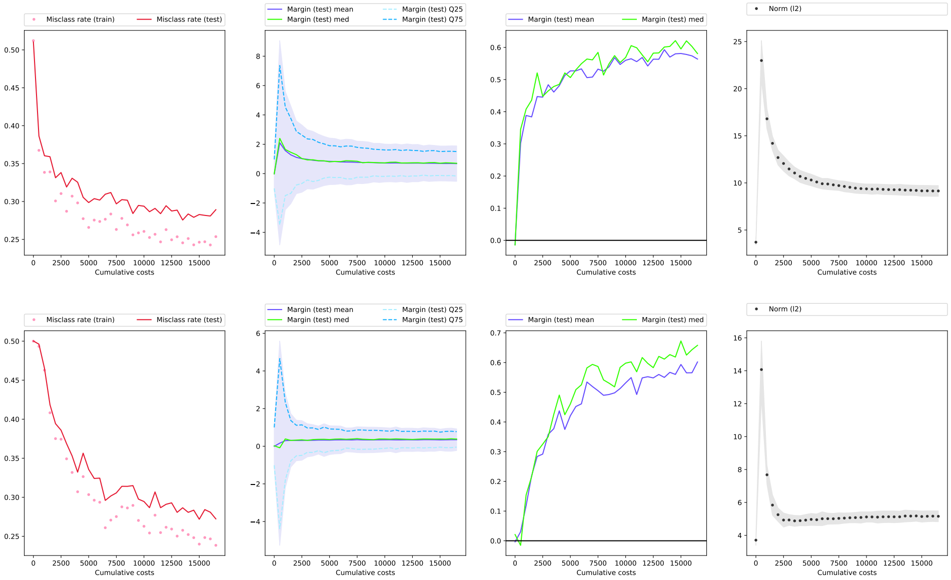

Figure 2: Performance metrics as a function of cost (gradients computed), for COV dataset. Top row: Pegasos. Bottom row: Algorithm 1 *without* scaling ($s = 1$). Regularization parameter λ (for both procedures) is set to the value at which Pegasos achieved the smallest off-sample error, here $\lambda = 10^{-3}$. In early stages of learning, the margin distribution for our algorithm has decidedly smaller variance and is notably much more symmetric.

Scaling and regularization In Figure 3, we look at the same performance metrics, but this time for a significantly under-regularized implementation of Algorithm 1, and examine the role that scaling plays. The case of no scaling just sets $s = 1$, while an upper bound on v_X for setting $s \geq \sqrt{n v_X / (2\lambda \log(\delta^{-1}))}$ is obtained using the 75th quantile of the deviation $\{|y_i \langle \widehat{\boldsymbol{w}}_{(t)}, \boldsymbol{x}_i \rangle - \gamma|\}_{i=1}^{n}$. Before re-scaling, the variance of the random update direction dominates, and the learner cannot find a good solution before the step size grows too small, highly inefficient. We can clearly see that re-scaling accelerates the regularization and stabilizes the updates in such a way that meaningful progress can be made in the critical early stages. Very similar trends hold over other data sets and comparably weak regularization, which is encouraging since *a priori* knowledge of good λ settings will not in general be available.

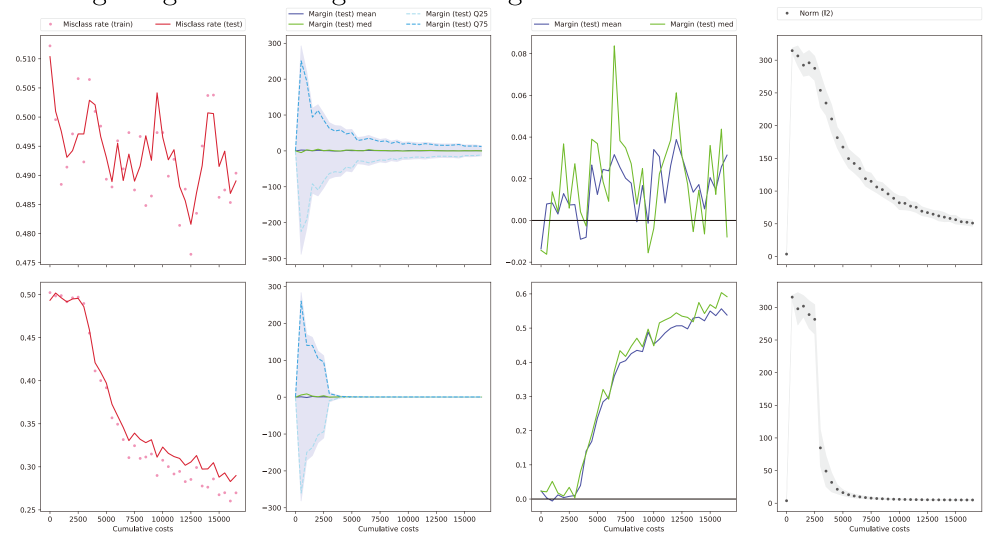

Figure 3: Performance metrics are just as introduced in Figure 2, except that both rows correspond to Algorithm 1, with much weaker regularization, here $\lambda = 10^{-5}$. Top row: without scaling. Bottom row: with scaling, but delayed for illustrative purposes, here re-scaling is done once at $t = 2500$. The impact on off-sample error and the margin distribution is stark, and is clearly effective at jump-starting an under-regularized procedure.

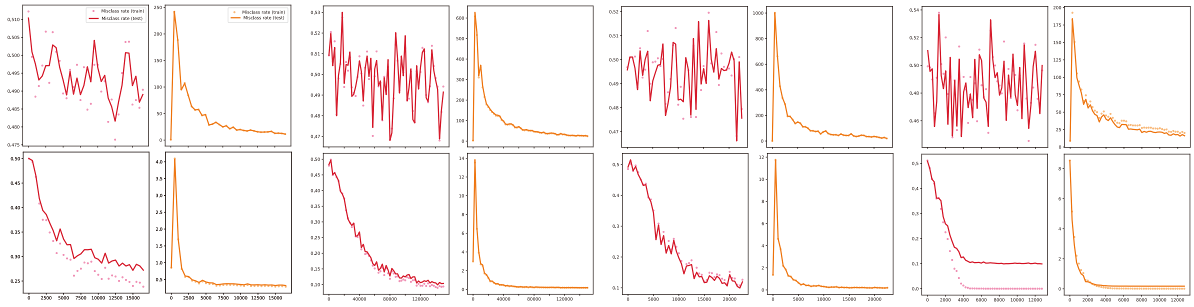

Figure 4: Comparison of classification error and surrogate error for Algorithm 1. Each column (two plots wide) corresponds to a distinct data set (from left to right: COV, DIGIT5, PROTEIN, SIDO). Top row: unscaled, just as in Figure 3. Bottom row: regularization just as in Figure 2. As we should expect, the bound on excess risk induced by Ψ_γ only becomes meaningful when the proxy R_φ is taken small enough.

Surrogate and classification error In Figure 4, we look at the relation between the classification error, and objective R_φ actually used by Algorithm 1 as a proxy, over all data sets. Good performance, as good or better than Pegasos, can readily be achieved either by appropriate regularization (λ setting) or via re-scaling. In the case of matching the λ settings of Algorithm 1 to the best choice for Pegasos, we see in Table 1 that the performance is effectively the same, if not better. Taken together with all the other figures, these results suggest that when the labels are in balanced proportions, our proposed algorithm provides a flexible and robust new procedure that can deal with data sets of varying sizes, sparsities, and distributions with little to no fine-tuning.

Table 1: Off-sample classification error (at *last* step $t = T$) when we naively match λ settings in Algorithm 1 to those of Pegasos, over all data sets. The "asymmetric" setting is Algorithm 1 but with $\gamma - y_i \langle \widehat{\boldsymbol{w}}_{(t)}, \boldsymbol{x} \rangle$ passed through the hinge loss before being passed to ρ.

Method	COV	DIGIT5	PROTEIN	SIDO
Pegasos (benchmark)	0.290	**0.071**	0.074	0.101
Ours, symmetric	**0.272**	0.074	0.089	**0.099**
Ours, asymmetric	0.278	**0.071**	**0.071**	0.102

5 Concluding remarks

In this paper, we introduced and analyzed a new learning algorithm which, via a new convex loss with re-scaling, lets us pursue stronger guarantees for the resulting margin distribution (and classifier) than are possible with the traditional hinge loss. This allows us to bridge the gap between inference and computation, since strong learning guarantees are available for Algorithm 1, which is readily implemented in practice. Empirical tests confirmed that the algorithm basically behaves as we would expect, and that even with naive parameter settings, appropriate re-scaling on the back end allows our procedure to match or exceed the performance of well-known competitors.

References

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

[2] 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.

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

[4] Bartlett, P. L. and Mendelson, S. (2003). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482.

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

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

[7] 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.

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

[9] Chen, Y., Su, L., and Xu, J. (2017). Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *arXiv preprint arXiv:1705.05491*.

[10] Daniely, A. and Shalev-Shwartz, S. (2014). Optimal learners for multiclass problems. In *27th Annual Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 287–316.

[11] Devroye, L., Györfi, L., and Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. Springer.

[12] Feldman, V. (2016). Generalization of ERM in stochastic convex optimization: The dimension strikes back. In *Advances in Neural Information Processing Systems 29*, pages 3576–3584.

[13] Ghosh, A., Manwani, N., and Sastry, P. (2015). Making risk minimization tolerant to label noise. *Neurocomputing*, 160:93–107.

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

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

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

[17] Long, P. M. and Servedio, R. A. (2010). Random classification noise defeats all convex potential boosters. *Machine learning*, 78(3):287–304.

[18] Luenberger, D. G. (1969). *Optimization by Vector Space Methods*. John Wiley & Sons.

[19] Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. (2013). Learning with noisy labels. In *Advances in Neural Information Processing Systems 26*, pages 1196–1204.

[20] Nesterov, Y. (2007). Gradient methods for minimizing composite objective function. Technical report, Université catholique de Louvain.

[21] Shalev-Shwartz, S., Singer, Y., Srebro, N., and Cotter, A. (2011). Pegasos: Primal Estimated sub-GrAdient SOlver for SVM. *Mathematical Programming*, 127(1):3–30.

[22] Takeda, A., Fujiwara, S., and Kanamori, T. (2014). Extended robust support vector machine based on financial risk minimization. *Neural Computation*, 26(11):2541–2569.

[23] Van Rooyen, B., Menon, A., and Williamson, R. C. (2015). Learning with symmetric label noise: The importance of being unhinged. In *Advances in Neural Information Processing Systems 28*, pages 10–18.

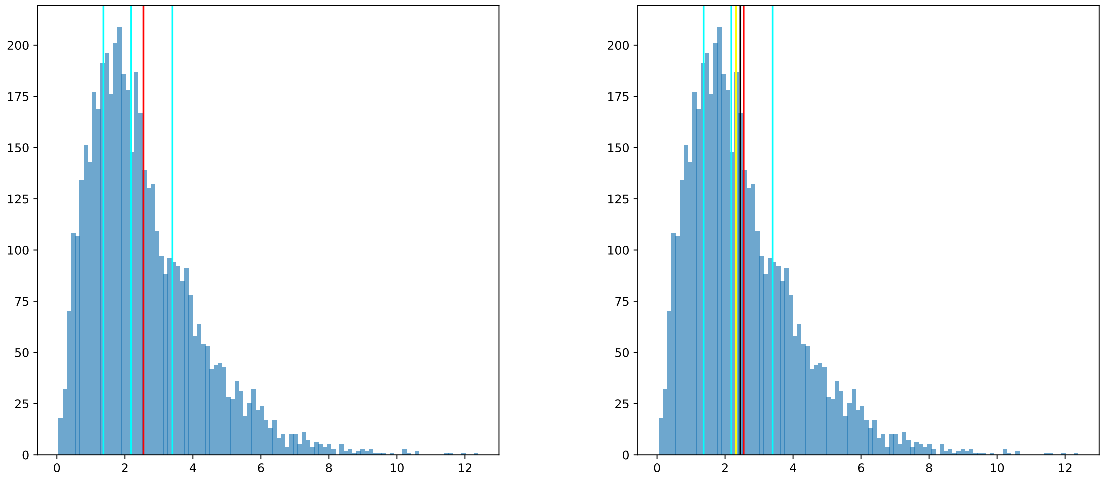

Figure 5: In the left-hand figure, we have plotted a histogram of a sample of 5000 points from the Gamma distribution (shape 2.5, scale 1.0), with the 25th, 50th, and 75th percentiles marked in cyan, and the mean marked in red. In the right-hand figure, we have marked the values of $\widehat{\gamma}$ when $s = 1.5$ (yellow) and $s = 3.0$ (black).

6 Appendix

Proofs of results in the main text

Proof of Prop. 1. We begin with a sufficient condition for γ 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 mth 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 [14]:

Lemma 8. *Let x be an arbitrary random variable with distribution ν. Assuming $\mathbf{E}_{\nu} |x|^3 < \infty$, it follows that defining*

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

the deviation can be controlled as

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

for constant $c > 0$.

Plugging in μ for ν, $y \, h(\boldsymbol{x})$ for x, and considering the analogous

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

it immediately follows that

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

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

For part 3, extending the results of Catoni [7], as long as ρ satisfies

$$-\log(1 - u + Cu^2) \leq \rho'(u) \leq \log(1 + u + Cu^2), \quad u \in \mathbb{R} \tag{10}$$

then exponential tails on the empirical estimator's deviation can be obtained; there is nothing particularly special about ρ in (4) besides computational convenience and ease of analysis. Given this inequality, Lemma 1 of Holland and Ikeda [14] implies that

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

For our setting, ρ defined in (4) indeed satisfies (10), with $C = 1/2$, which follows from Lemma 1 of Catoni and Giulini [8], where this function in analyzed in the context of robust vector mean estimates. Optimizing the upper bound with respect to $s > 0$ and plugging in $C = 1/2$ yields part 3. □

Proof of Remark 3. 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, $(-\infty, \alpha_l^*]$. (α_l^*, α_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 $\mathbf{E}_{\mu} \, I\{y \neq h(x)\}$, the optimal choice is to select $h(\cdot; \alpha_l^*)$, in which case

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

This gives us a lower bound on performance, namely

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

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

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

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

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

Since we know that

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

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

Proof of Lemma 4. We follow along with the now-standard framework set out by Bartlett et al. [2]. We define our loss function $\varphi(u) := \rho(\gamma - u)$ using ρ as specified in the hypothesis.

Next we put together the analytical machinery that will be used. First, the conditional expected φ-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 η denotes $\eta = \mathbf{P}\{y = 1\}$. A generalization of this quantity for arbitrary $\eta \in [0, 1]$ is constructed as

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

The optimal value that this takes is denoted by

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

Denote any optimal value using u^*, namely any

$$u^* \in \operatorname*{arg\,min}_{u \in \mathbb{R}} C_{\eta}(u).$$

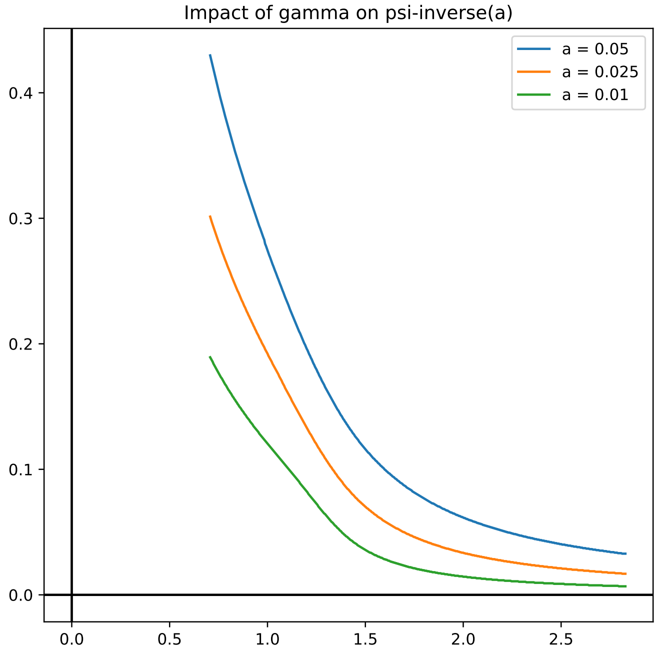

Figure 6: Graph of $\Psi_{\gamma}^{-1}(a)$ as a function of γ ranging between $\sqrt{2}/2$ and $2\sqrt{2}$, for fixed values of a. Computation is approximate, and done as follows. For each γ, we compute $\Psi_{\gamma}(u)$ for $u \in [0,1]$ over a uniformly spaced grid $0 = u_1 \leq u_2 < \cdots < u_K = 1$, with $K = 2500$. The approximate value is then given as $\Psi_{\gamma}^{-1}(a) = u_{k^*}$, where $k^* = \max\{k \in [K] : \Psi_{\gamma}(u_k) \leq a\}$.

If this value is indeed unique, then it makes sense to map $\eta \mapsto u^*(\eta)$. In relating R and R_{φ}, 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:

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

Using the "best (generalized) conditional φ-risk that can be achieved despite having different signs from $(2\eta - 1)$." Of course, the word "despite" becomes appropriate when we replace η 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 φ 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 φ-risk could be achieved despite having mis-labeled some instances, which would immediately imply a small R_{φ} but larger R. To ensure that this cannot happen, the condition put forward by Bartlett et al. [2] is very lucid: call φ *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 Ψ. It is this function that has the desirable properties that interest us. For one, via their Theorem 1, for any non-negative φ, 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 φ 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 φ, being a convex function on $\mathbb{R}$, is classification calibrated. Furthermore, via Theorem 2(2), the Ψ function takes a particularly simple form:

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

where $\varphi(0) = \rho(\gamma)$ gives us the expression from the hypothesis. All that remains in order to obtain Ψ then is to compute H 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

$$\frac{\rho'(u-\gamma)}{\rho'(u+\gamma)} = \frac{\eta - 1}{\eta}. \tag{11}$$

That is to say, for $\eta \in (0,1)$, any u^{*} satisfying (11) will be a minimizer in that $C_{\eta}(u^{*}) \leq C_{\eta}(u)$ for all u.

It should be clear that the value of γ 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 $H(0) = H(1) = 0$. 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 (11) 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 (11) is always satisfied by setting $u = 0$, for which case we have

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

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

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

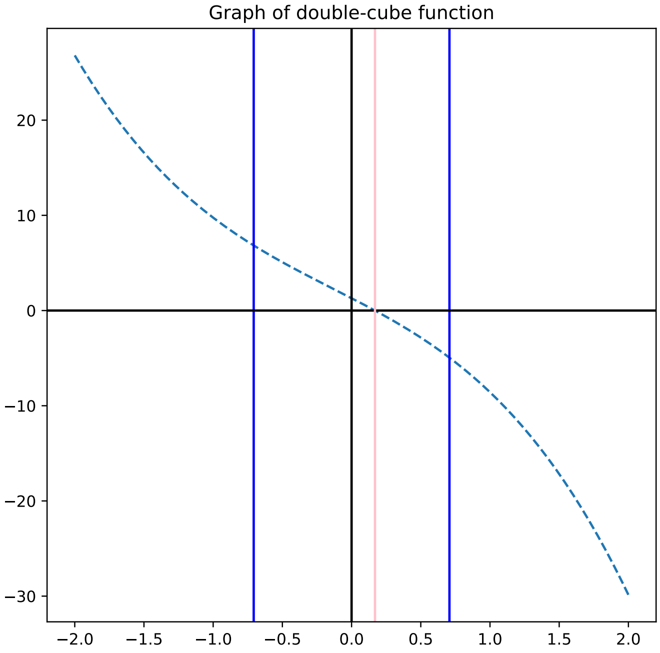

Figure 7: Graph of the third-degree polynomial used in the double-cube condition (12). Vertical blue lines denote ±γ (here $\gamma = \sqrt{2}/2$), and the vertical pink line denotes the root computed analytically.

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 $\rho'(u) = u - u^3/6$, rearrange equality (11), and solve for roots of the resulting cubic polynomial. The computations are quick, and writing

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

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

$$\begin{aligned} a &= 1 - \alpha \\ b &= -3\gamma(1 + \alpha) \\ c &= 3(1 - \alpha)(\gamma^2 - 2) \\ d &= (1 + \alpha)(6\gamma - \gamma^3) \\ P(u; a, b, c, d) &= 0. \end{aligned} \tag{12}$$

Call this (12) the *double-cube* condition (see Figure 7). The discriminant of an arbitrary cubic polynomial is defined

$$\Delta \coloneqq 18abcd - 4b^3d + b^2c^2 - 4ac^3 - 27a^2d^2, \tag{13}$$

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 $H(\eta) = C_{\eta}(u^{*}(\eta))$.

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

$$\begin{aligned} a &= 1 \\ b &= -3\gamma \\ c &= 3\gamma^2 - 6 \\ d &= -6\left(\gamma + \frac{\rho'(\sqrt{2})}{\alpha} - \frac{\gamma^3}{6}\right) \\ P(u; a, b, c, d) &= 0 \end{aligned} \tag{14}$$

and

$$\begin{aligned} a &= 1 \\ b &= 3\gamma \\ c &= 3\gamma^2 - 6 \\ d &= 6\left(\gamma + \alpha\rho'(\sqrt{2}) - \frac{\gamma^3}{6}\right) \\ P(u; a, b, c, d) &= 0 \end{aligned} \tag{15}$$

Graphs of these polynomials are plotted in Figure 8. Computationally determining which to use is straightforward. By the monotonicity of ρ', we can simply check the edge case $u = \operatorname{sign}(\eta - 1/2)\delta$. In the case of $\eta > 1/2$, noting that both the LHS and RHS are negative, if

$$\frac{\rho'(\delta - \gamma)}{\rho'(\delta + \gamma)} < \frac{\eta - 1}{\eta}, \tag{16}$$

then the solution must be larger than δ, 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

$$\frac{\rho'(-\delta - \gamma)}{\rho'(-\delta + \gamma)} > \frac{\eta - 1}{\eta}, \text{ or more cleanly, } \frac{\rho'(\delta - \gamma)}{\rho'(\delta + \gamma)} < \frac{\eta}{\eta - 1}, \tag{17}$$

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 γ values being considered[2], as follows:

[2] 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 (δ, γ) ($\eta > 1/2$ case) or $(-\gamma, -\delta)$ ($\eta < 1/2$ case).

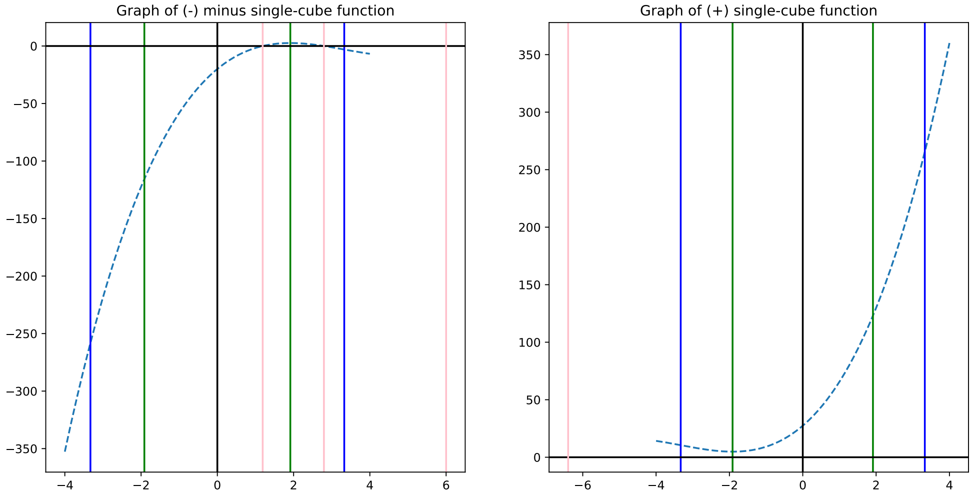

Figure 8: Graph of the third-degree polynomials used in the single-cube conditions. The left figure corresponds to condition (14), and the right figure corresponds to condition (15). The vertical blue lines are again $\pm\gamma$ (with $\gamma = 2\sqrt{2} + 1/2$ here), and the vertical green lines are $\pm\delta$.

- When $\eta > 1/2$:
  - If (16), then solve (14), take root falling in (δ, γ).
  - Else, solve (12).
- When $\eta < 1/2$:
  - If (17), then solve (15), take root falling in $(-\gamma, -\delta)$.
  - Else, solve (12).

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 $C_\eta(u^*) = H(\eta)$. It follows then that following the procedures outlined above, we can also compute $\Psi(a) = \varphi(0) - H((1 + a)/2)$ for arbitrary $a \in [-1, 1]$.

As for the consistency part of our hypothesis, note that by using Bartlett et al. [2, Lemma 1], we have that for any pair of sequences (a_n) and (b_n) where $a_n \downarrow 0$ and $\Psi(b_n) \leq a_n$ for all n, it follows that

$$0 = \lim_{n\to\infty} a_n \geq \lim_{n\to\infty} \Psi(b_n) = \Psi\left(\lim_{n\to\infty} b_n\right) \geq 0.$$

This implies $b_n \to 0$ as $n \to \infty$, and thus the sufficiency of the consistency condition given in the hypothesis. □

Proof of Theorem 6. By Lemma 4, we have that for any choice of $\boldsymbol{w} \in \mathcal{W}$,

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

To control the right-hand side, note that by strong convexity $\boldsymbol{w}^*$ is the unique minimum of R_φ, and so $R_\varphi(\boldsymbol{w}) - R_\varphi^* = R_\varphi(\boldsymbol{w}) - R_\varphi(\boldsymbol{w}^*)$. Since R_φ is smooth via Lemma 10 with coefficient v_X, using basic property 25 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. This can be broken up into computational and statistical elements, as follows. Writing the pre-projection version of $\widehat{\boldsymbol{w}}_{(t)}$ as $\widetilde{\boldsymbol{w}}_{(t)}$ for each $t = 0, \ldots, T-1$, we can readily bound this distance from above as

$$\begin{aligned}
\|\widehat{\boldsymbol{w}}_{(t+1)} - \boldsymbol{w}^*\| &= \|\Pi_{\mathcal{W}}(\widetilde{\boldsymbol{w}}_{(t+1)}) - \Pi_{\mathcal{W}}(\boldsymbol{w}^*)\| \\
&\leq \|\widetilde{\boldsymbol{w}}_{(t+1)} - \boldsymbol{w}^*\| \\
&= \|\widehat{\boldsymbol{w}}_{(t)} - \alpha \widehat{R}'_\varphi(\widehat{\boldsymbol{w}}_{(t)}) - \boldsymbol{w}^*\| \\
&\leq \|\widehat{\boldsymbol{w}}_{(t)} - \alpha R'_\varphi(\widehat{\boldsymbol{w}}_{(t)}) - \boldsymbol{w}^*\| + \alpha \|\widehat{R}'_\varphi(\widehat{\boldsymbol{w}}_{(t)}) - R'_\varphi(\widehat{\boldsymbol{w}}_{(t)})\|.
\end{aligned}$$

A few comments on these inequalities. The projection map $\Pi_{\mathcal{W}}$ is well-defined since $\mathcal{W}$ is closed and convex, and the contraction property used here is a standard result [18, Sec. 3.12, Thm. 3.12]. Here $\widehat{R}_\varphi$ denotes the empirical mean of R_φ based on the sample. By **A1** we have that $\|\boldsymbol{w}^*\| \leq 1/\sqrt{\lambda}$, so $\Pi_{\mathcal{W}}(\boldsymbol{w}^*) = \boldsymbol{w}^*$ trivially. After using the triangle inequality to get a bound on $\|\widetilde{\boldsymbol{w}}_{(t+1)} - \boldsymbol{w}^*\|$, the first summand of that bound is the ideal gradient descent update (one step) of the iterative procedure to minimize R_φ, given $\widehat{\boldsymbol{w}}_{(t)}$ output from the previous step. For small enough step size $0 < \alpha < 2/(\kappa + v_X)$, the update improves on the previous error as

$$\|\widehat{\boldsymbol{w}}_{(t)} - \alpha R'_\varphi(\widehat{\boldsymbol{w}}_{(t)}) - \boldsymbol{w}^*\|^2 \leq \left(1 - \frac{2\alpha\kappa v_X}{\kappa + v_X}\right) \|\widehat{\boldsymbol{w}}_{(t)} - \boldsymbol{w}^*\|^2.$$

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

$$\|\widehat{\boldsymbol{w}}_{(t+1)} - \boldsymbol{w}^*\| \leq \sqrt{1 - \alpha\beta} \|\widehat{\boldsymbol{w}}_{(t)} - \boldsymbol{w}^*\| + \alpha \|\widehat{R}'_\varphi(\widehat{\boldsymbol{w}}_{(t)}) - R'_\varphi(\widehat{\boldsymbol{w}}_{(t)})\|. \tag{18}$$

This deals with the computational error part. Now for the statistical error part, namely the accuracy of the $\widehat{R}'_\varphi \approx R'_\varphi$ approximation. Using assumption **A2**, the sub-Gaussianity we are assuming is that for some constant $c > 0$, the moment generating function can be bounded as

$$\mathbf{E} \exp(a \langle \boldsymbol{u}, \boldsymbol{g}(\boldsymbol{w}) - \mathbf{E}_\mu \boldsymbol{g}(\boldsymbol{w}) \rangle) \leq \exp(ca^2 \langle \boldsymbol{u}, \Sigma(\boldsymbol{w}) \boldsymbol{u} \rangle), \quad a \geq 0, \boldsymbol{w} \in \mathbb{R}^d$$

for all $\|\boldsymbol{u}\| = 1$, where $\Sigma(\boldsymbol{w})$ is the covariance matrix of $\boldsymbol{g}(\boldsymbol{w})$. Now, suppressing $\boldsymbol{w}$ from the notation for readability, noting that

$$\begin{aligned}
\Sigma &= \mathbf{E}_\mu (\boldsymbol{g} - \mathbf{E}_\mu \boldsymbol{g})(\boldsymbol{g} - \mathbf{E}_\mu \boldsymbol{g})^T \\
&= \mathbf{E}_\mu \boldsymbol{g}\boldsymbol{g}^T - (\mathbf{E}_\mu \boldsymbol{g})(\mathbf{E}_\mu \boldsymbol{g})^T
\end{aligned}$$

and by positive semi-definiteness of $(\mathbf{E}_\mu \boldsymbol{g})(\mathbf{E}_\mu \boldsymbol{g})^T$ since we have for all $\boldsymbol{u}$ that

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

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

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

With these inequalities, we can leverage Lemma 11 to prove that for any fixed $\boldsymbol{w}$, writing

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

and taking the empirical vector mean estimator as $\bar{\boldsymbol{g}}(\boldsymbol{w}) \coloneqq n^{-1}\sum_{i=1}^{n}\boldsymbol{g}_i(\boldsymbol{w})$, for any fixed $\boldsymbol{w}$ we have that the event

$$\mathcal{E}(\boldsymbol{w}) \coloneqq \left\{ \|\bar{\boldsymbol{g}}(\boldsymbol{w}) - R'_{\varphi}(\boldsymbol{w})\| > 2\sqrt{\frac{c\rho'(\sqrt{2})^2\,\mathbf{E}_{\mu}\,\|\boldsymbol{x}\boldsymbol{x}^T\|\log(\delta^{-1})}{n}} \right\} \tag{19}$$

has probability $\mathbf{P}\,\mathcal{E}(\boldsymbol{w}) \leq \delta$ (in the hypothesis statement, constant c absorbs all these coefficients).

In practice however, $\widehat{\boldsymbol{w}}_{(t)}$ for all $t > 0$ will be random, dependent on the sample, and thus in general there is typically no 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 ϵ-cover of $\mathcal{W}$. Since $\mathcal{W}$, a ball with radius $1/\sqrt{\lambda}$, is a compact subset of $\mathbb{R}^d$, the ϵ covering number is bounded as $N_\epsilon \leq (3/\epsilon\sqrt{\lambda})^d$. Denote the centers of the ϵ 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

$$\|\bar{\boldsymbol{g}}(\boldsymbol{w}) - R'_{\varphi}(\boldsymbol{w})\| \leq \|\bar{\boldsymbol{g}}(\boldsymbol{w}) - \bar{\boldsymbol{g}}(\widetilde{\boldsymbol{w}})\| + \|R'_{\varphi}(\boldsymbol{w}) - R'_{\varphi}(\widetilde{\boldsymbol{w}})\| + \|\bar{\boldsymbol{g}}(\widetilde{\boldsymbol{w}}) - R'_{\varphi}(\widetilde{\boldsymbol{w}})\|. \tag{20}$$

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 ρ', it follows immediately that we can bound

$$\|\boldsymbol{g}_i(\boldsymbol{w}) - \boldsymbol{g}_i(\widetilde{\boldsymbol{w}})\| \leq \|\boldsymbol{x}_i\|\|\boldsymbol{w} - \widetilde{\boldsymbol{w}}\|.$$

This implies that

$$\|\bar{\boldsymbol{g}}(\boldsymbol{w}) - \bar{\boldsymbol{g}}(\widetilde{\boldsymbol{w}})\| \leq \epsilon\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}_i\|^2 \leq \frac{\epsilon\,\mathbf{E}_{\mu}\,\|\boldsymbol{x}\|^2}{\delta} = \frac{\epsilon v_X}{\delta} \tag{21}$$

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 ϵ, this (21) represents a uniform bound.

For the second term, we just use the v_X-smoothness of R_φ, shown in Lemma 10. This implies

$$\|R'_{\varphi}(\boldsymbol{w}) - R'_{\varphi}(\widetilde{\boldsymbol{w}})\| \leq v_X\|\boldsymbol{w} - \widetilde{\boldsymbol{w}}\| \leq v_X\epsilon \tag{22}$$

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 (19). The critical fact of course is that making the dependence of $\widetilde{\boldsymbol{w}}$ on $\boldsymbol{w}$ explicit, we have

$$\sup_{\boldsymbol{w}\in\mathcal{W}}\left\|\bar{\boldsymbol{g}}(\widetilde{\boldsymbol{w}}(\boldsymbol{w})) - R'_{\varphi}(\widetilde{\boldsymbol{w}}(\boldsymbol{w}))\right\| = \max_{k\in[N_\epsilon]}\left\|\bar{\boldsymbol{g}}(\widetilde{\boldsymbol{w}}_k) - R'_{\varphi}(\widetilde{\boldsymbol{w}}_k)\right\|.$$

The good event of interest here is

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

and thus with a union bound we have that with probability no less than $1-\delta$, we can uniformly bound as

$$\|\bar{\boldsymbol{g}}(\widetilde{\boldsymbol{w}}(\boldsymbol{w})) - R'_{\varphi}(\widetilde{\boldsymbol{w}}(\boldsymbol{w}))\| \leq 2\sqrt{\frac{c\rho'(\sqrt{2})^2 \mathbf{E}_{\mu} \|\boldsymbol{x}\boldsymbol{x}^T\| \log(N_{\epsilon}\delta^{-1})}{n}}, \quad \forall\, \boldsymbol{w} \in \mathcal{W}. \tag{23}$$

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

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

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

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

where we have defined

$$\varepsilon^* := 2\sqrt{\frac{c\rho'(\sqrt{2})^2 \mathbf{E}_{\mu} \|\boldsymbol{x}\boldsymbol{x}^T\|(d\log(3\sqrt{n}(2\sqrt{\lambda}\delta)^{-1}) + \log(\delta^{-1}))}{n}}. \tag{24}$$

On the good event $\mathcal{E}_+$, then denoting $\varepsilon = \varepsilon^* + (1+\delta)v_X/\sqrt{n}$, for *all* steps t we can re-write (18) 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

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

which via the first inequality of this proof using Ψ_{γ}, yields the desired result. □

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

Lemma 9. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable, convex, and l-smooth. Then, we have*

$$f(\boldsymbol{u}) - f(\boldsymbol{v}) \leq \frac{l}{2}\|\boldsymbol{u}-\boldsymbol{v}\|^2 + \langle f'(\boldsymbol{v}), \boldsymbol{u}-\boldsymbol{v}\rangle \tag{25}$$

$$\|f'(\boldsymbol{u}) - f'(\boldsymbol{v})\|^2 \leq 2l\left(f(\boldsymbol{u}) - f(\boldsymbol{v}) - \langle f'(\boldsymbol{v}), \boldsymbol{u}-\boldsymbol{v}\rangle\right). \tag{26}$$

Proof. Given in chapter 2 of Nesterov [20]. □

Lemma 10. *The surrogate risk $R'_{\varphi}(h)$ defined in (9), for $h(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x}\rangle$, $\boldsymbol{w} \in \mathbb{R}^d$, is l-smooth with coefficient $l = \mathbf{E}_{\mu}\|\boldsymbol{x}\|^2$.*

Proof. Assuming the order of integration and differentiation can be reversed, one can write R'_{φ} as

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

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

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

where we utilized the property that ρ' is 1-Lipschitz. This implies that

$$\|R'_{\varphi}(\boldsymbol{w}_1) - R'_{\varphi}(\boldsymbol{w}_2)\| \leq l\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|, \quad \boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d \tag{27}$$

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

Lemma 11 (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 of Lemma 11. 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 [5]. 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 ℓ_2-induced matrix norm, equivalent to the spectral norm, i.e., the largest singular value of Σ_X [15]. 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

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

Plugging this in to our Chernoff equality,

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

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

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

which implies the desired result. □

Explanation of root-finding function `getroot` The basic strategy used for root finding is very straightforward. The cubic polynomials considered are equations of the form

$$au^3 + bu^2 + cu + d = 0. \tag{28}$$

Recall the discriminant Δ given in (13). 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.
- 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

$$\begin{aligned}\Delta_0 &:= b^2 - 3ac\\ \Delta_1 &:= 2b^3 - 9abc + 27a^2 d\end{aligned}$$

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 Δ_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

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

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:

$$\cos 3x = 4\cos^3 x - 3\cos x. \tag{29}$$

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 (29). 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

$$v^3 + pv + q = 0 \tag{30}$$

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 (28), which implies three distinct real roots for (28), it follows that (30) also has three distinct real roots. This can only happen when its discriminant is positive, which is to say when

$$-4p^3 - 27q^2 > 0. \tag{31}$$

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 (29), 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

$$4z^3 - 3z = -\frac{q}{2k^3}, \tag{32}$$

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 (31) 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 (32) 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 (32), we have

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

Then using the key identity (29), 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 (32). Note that since this function has period 2π, it holds that

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

which after plugging in to (29), 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 (32):

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

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

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

where $j = 1, 2, 3$. To summarize the case of finding roots when $\Delta > 0$, the basica 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 (28).