Proceedings of Machine Learning Research vol 65:1–33, 2017

Ignoring Is a Bliss: Learning with Large Noise Through Reweighting-Minimization

Daniel Vainsencher DANIEL.VAINSENCHER@GMAIL.COM
Voleon

Shie Mannor SHIE@EE.TECHNION.AC.IL
Faculty of Electrical Engineering,
Technion Israel Institute of Technology

Huan Xu HUAN.XU@ISYE.GATECH.EDU
School of Industrial and Systems Engineering,
Georgia Institute of Technology

Abstract

We consider learning in the presence of arbitrary noise that can overwhelm the signal in terms of magnitude on a fraction of data points observed (aka outliers). Standard approaches based on minimizing empirical loss can fail miserably and lead to arbitrary bad solutions in this setting. We propose an approach that iterates between finding a solution with minimal empirical loss and reweighting the data, reinforcing data points where the previous solution works well. We show that our approach can handle arbitrarily large noise, is robust as having a non-trivial breakdown point, and converges linearly under certain conditions.

The intuitive idea of our approach is to automatically exclude "difficult" data points from model fitting. More importantly (and perhaps surprisingly), we validate this intuition by establishing guarantees for generalization and iteration complexity that *essentially ignore the presence of outliers*.

1. Introduction

This paper is about learning with gross noise. Gross noise exists widely in real applications, due to various reasons: unreliable acquisition, inadvertent mixing of sources, or occurrence of rare events that are validly measured but too scarce to estimate. Many common learning algorithms are not robust to gross noise – they can be disproportionately skewed if even a small proportion of training data points are affected by gross noise. Popular quadratic-loss based methods such as least squares regression and ℓ_2-SVM are especially brittle, but even methods based on more robust loss functions such as ℓ_1-SVM and Huber loss regression are susceptible to gross noise as well (Tukey, 1960, 1962; Huber et al., 1964), see Section 5 for a more detailed survey.

We propose a framework based on (re)-weighting the data. In particular, we propose to minimize a joint loss function that *chooses a weight vector over the training samples* and selects the model which best fits the weighted training set. The intuition is that by selecting a weight vector that minimizes the loss, the algorithm automatically reduces the effect or even completely ignores data points with excessive loss, thus making the algorithm less susceptible to a few gross outliers. To avoid overfitting "easy" data – the obtained weight concentrating on a few data points with very low loss – we introduce a regularization term *on the weight*. This *regularized weighting* approach is general and readily applicable to *any* learning problem that is formulated as minimizing a certain loss. *Regularized weighting* was initially proposed in Vainsencher et al. (2013) in a more general context of multiple-model estimation, where several models are learned simultaneously and data are associated to models according to the success of the models on the data. Here we consider the specialized case of learning just a single model while excluding outliers (or other points with a large noise) and therefore we are able to obtain stronger results in terms of convergence, statistical guarantees and robustness. We give the formulation and explore the properties of the weights it chooses in in Section 2.

Algorithmically, we solve the formulation by alternating minimization between *reweighting* the data (i.e., computing the weight) and optimizing over the weighted data (i.e., selecting the model with *minimal* loss). We therefore term the proposed approach Reweighting-Minimization (RM). Since the reweighting step is efficient, the cost of this approach is dominated by the number of iterations required. We describe the algorithm in detail and analyze the cost of a single iteration in Section 2.1 and in Section 3 show the number of iterations required is often logarithmic in the precision required.

Our main contribution is establishing formally that *regularized weighting* is an effective approach addressing the main issues of learning with gross noise. The cornerstone of our theoretical analysis is that the reweighting step possesses three desirable properties, shown in Proposition 1: (1) Weights are close to uniform. (2) Excessively large losses receive weight zero. And (3) weights depend continuously on the losses. Based on these structural properties of the weights, we then show the following advantages for the proposed reweighted minimization framework:

Sample complexity: Due to the first two properties the weighted loss is a weighted average of bounded variables, where the weight is close to uniform, and the range of those variables is close to a *typical* loss and rather than a maximal loss. This leads to a novel sample complexity result that is *exponentially* less sensitive to large values of losses (e.g., outliers) than previously known results for similar estimators. Indeed, our proof technique is rather general and hence of independent interest: it can be easily adapted to other robust estimators, such as to the well studied *Least Trimmed Squares* (Rousseeuw, 1984).

Iteration complexity: The third property allows us to show that for sufficiently regular estimation problems, the method converges to a local optimum linearly. The analysis enables the translation of additional assumptions about the outliers into conclusions about lack of local minima and hence global convergence.

Robustness of estimator to outliers: For some classes of problems our approach converts base estimators into robust ones, in the sense of having a non-zero breakdown point. This holds for location estimation as a special case of results in Vainsencher et al. (2013). For Sparse LTS, Alfons et al. (2013) presented an argument that can be generalized to regularized models outside regression and also from LTS to $\mathcal{L}_\beta$, but is not the focus of this paper.

This paper is organized as follows. In Section 2 we present the formulation of the proposed reweighted minimization framework, and provide the algorithmic approach to solve the formulation. Moreover, we establish structural properties of the obtained weight vectors, which serves as the corner stone of the theoretic analysis of the framework. We show in Section 3 that the presented algorithm converges to a local optimum linearly under mild technical conditions. Then, in Section 4 we establish sample complexity results for our formulation. As we discussed above, the sample complexity results is about *typical* losses, and we show that we obtain results exponentially less sensitive to outliers, which validates the robustness of the framework. We then discuss and compare with relevant literature in Section 5 before concluding the paper in Section 6. All proofs and simulatoins results are deferred to the appendix.

TECHNICAL PRELUDE

In analyzing this method for reweighting data, we find the following language useful. We represent data explicitly as an empirical distribution, and use a normalized Euclidean norm. This makes some quantities critical to our formulation conveniently independent of the sample size (and applicable to the limit of infinite data); we explain some unfamiliar consequences here along with other notation.

We denote by n the number of data points in a sample $(x_i)_{i=1}^n$ where $x_i \in \mathcal{X}$. We represent the sample by an empirical distribution μ assigning equal probability to each x_i. We assume a loss function $\ell : \mathcal{M} \times \mathcal{X} \to \mathbb{R}$ so that $\ell(m, x)$ gives the cost of using model m on data point x. We denote by ℓ^m the losses of a model $m \in \mathcal{M}$ on the data μ and by $\mathbf{w}$ the weights we assign to individual data points in μ. Hence $\ell^m, \mathbf{w}$ can be viewed as vectors in $\mathbb{R}^n$ or in a functional analysis point of view, as functions in the space $L^2(\mu)$. From the latter we need to take only its norm, different from the standard Euclidean norm only in normalization: $\|\mathbf{a}\|_\mu^2 = n^{-1}\sum_{i=1}^n a_i^2$. This choice induces a normalized inner product $\langle \mathbf{a}, \mathbf{b} \rangle_\mu = n^{-1}\sum_{i=1}^n a_i b_i$. Under this inner product, the average of vector $\mathbf{a}$ can be written as $\langle \mathbf{1}_n, \mathbf{a} \rangle_\mu$ where $\mathbf{1}_n$ is the n dimensional vector equal to 1 on all coordinates; generalizing, weighted averages will replace $\mathbf{1}_n$ by vectors from the normalized simplex $\triangle^\mu = \{\mathbf{w} \in \mathbb{R}^n : \langle \mathbf{w}, \mathbf{1}_n \rangle_\mu = 1 \text{ and } \mathbf{w}_i \geq 0, \forall i\}$.

When there is no risk of confusion we omit the subscript μ. We denote by $P_{\triangle^\mu}$ the projection operator into the simplex $\triangle^\mu$ with regard to $\|\cdot\|_\mu$. We denote by $B_d(x, r)$ the ball of radius r in metric d around point x. We denote by χ_A the indicator function that is 1 on A and 0 elsewhere, and by $\circ$ the elementwise product. For example, $\chi_A \circ \ell$ is a version of ℓ that is zero on data outside A. For a vector x we denote $[x]_+ = \max\{0, x\}$ elementwise.

2. Formulation

In this section we formalize the setting, specify our formulation and give a concrete example to illustrate the intuition. We are given a data space $\mathcal{X}$, a space of models $\mathcal{M}$, and a subroutine for finding a model that minimizes the weighted losses on the sample. Our formulation finds jointly a model $m \in \mathcal{M}$ and a weighting vector $\mathbf{w}$ by minimizing the following loss whose solutions have weights as in Eq. (2):

$$\mathcal{L}_\beta(m; \mu) = \min_{\mathbf{w} \in \triangle^\mu} \left\{ \langle \mathbf{w}, \ell^m \rangle_\mu + \beta \|\mathbf{w} - \mathbf{1}_n\|_\mu^2 \right\}, \tag{1}$$

where $\beta > 0$ and $\ell^m = (\ell(m, x_i))_{i=1}^n \in \mathbb{R}^n$ is the vector of losses of model m on the data represented by empirical distribution μ. Recalling our notation, we note that $\mathbf{w} \in \triangle^\mu$ means its entries are in $[0, n]$ and average 1. Therefore the first term $\langle \mathbf{w}, \ell^m \rangle_\mu$ is a weighted average of losses, and we explore $\mathcal{L}_\beta$ through the weights $\mathbf{w}$ that apply. As a simple example, since $\mathbf{1}_n \in \triangle^\mu$, our loss $\mathcal{L}_\beta(m; \mu)$ always bounds the average loss $\langle \mathbf{1}_n, \ell^m \rangle_\mu$ from below.

The first term induces a preference (higher weights) for lower losses in minimizing $\mathbf{w}$; further properties of $\mathbf{w}$ are determined by the weight regularization term $\beta \|\mathbf{w} - \mathbf{1}_n\|_\mu^2$. This second term keeps $\mathbf{w}$ close to a uniform distribution in a sense we formalize below, and prove in Appendix A.

Proposition 1 *Fix a model $m \in \mathcal{M}$. Then the minimizing $\mathbf{w}$ in $\mathcal{L}_\beta$ fulfills:*

(P1) Projection structure:

$$\mathbf{w} = P_{\triangle^\mu}\left(-\frac{\boldsymbol{\ell}^m}{2\beta}\right) = \left[-\boldsymbol{\ell}^m/\left(2\beta\right) + a\mathbf{1}_n\right]_+ \tag{2}$$

where $a \in \mathbb{R}$ is uniquely determined by the sum constraint on $\mathbf{w}$.

For any $\rho_l \leq \min_i \ell(m, x_i)$ and choose ρ_h such that $I = \{i : \ell(m, x_i) \leq \rho_h\}$, i.e., the set of indices of data points with losses upper-bounded by ρ_h, has $p = |I|/n \geq 2/3$. Then additionally, for $\beta = c(\rho_h - \rho_l)$ where $c \geq 1$, the following properties hold as well:

(P2) Uniform boundedness: $\mathbf{w}_i \leq p^{-1} + (2c)^{-1} \leq 2$, for all i.

(P3) Non-trivial weight for typical loss: $\mathbf{w}_i \geq \left(2 - p^{-1}\right)p^{-1} - (2c)^{-1} \geq 1/4$, for all $i \in I$.

(P4) Ignoring extreme outliers: $\mathbf{w}_i = 0$, whenever $\ell(m, x_i) \geq \rho_h + 3\beta$.

Moreover, denoting $L(\mu, p, m) = |I|^{-1}\sum_{i\in I}\ell(m, x_i)$ the average of losses in $[\rho_l, \rho_h]$, we have that $\mathcal{L}_\beta(m) - \rho_l \geq (L(\mu, p, m) - \rho_l)/6$ for such β.

Proposition 1 is a fundamental result for the proposed method, exposing the structure of weights especially for sufficiently large β. This structural result is the stepping stone for algorithmic results and statistical results presented in Sections 3 and 4. In the remainder of this section, we use the result to explore how $\mathcal{L}_\beta$ behaves.

The projection structure property (P1) of the proposition makes it easy to analyze the behavior of the weights. One immediate consequence is that the mapping of losses to weights is Lipschitz continuous with parameter $(2\beta)^{-1}$, hence more stable for large β. As another example, it is easy to see that if β equals the difference between the smallest loss and the second smallest loss, then all weight will be assigned to the single example with the smallest loss, causing extreme overfitting. Taking β larger than the range of the smallest 2/3 of losses avoids this, hence is assumed by the remaining properties P2-P4. Also, P1 states that computing $\mathbf{w}$ is essentially performing ℓ_1 ball projection, which can be done in $O(n)$ time as shown in Duchi et al. (2008).

The uniform boundedness property (P2) shows that for large enough β, the range of weights is reduced from $[0, n]$ to the near-uniform $[0, 2]$; in particular no small subset of data is given dominant weights. For sufficiently high β, for example taking $p = 1$ and $c = 10$ (P3) shows that any examples with loss within β of the minimal loss do receive non-trivial weight. From (P4) we see that losses worse than the last by a few β are assigned weight zero hence ignored as outliers. Hence while β governs an important tradeoff in our method, its effects are well understood on a wide range.

Lastly, Proposition 1 shows that $\mathcal{L}_\beta$ is sandwiched between the average loss and another well known loss. In LTS assigned weights are defined to be uniform over a proportion p of the data having the smallest losses and zero elsewhere. (P2) and (P4) show our $\mathbf{w}$ behaves similarly, and indeed the last part of Proposition 1 shows that $\mathcal{L}_\beta$ upper bounds the LTS loss.

To illustrate our formulation, take ridge regression as a concrete example. We let

$$\ell(m, x_i) = \left(\langle m, x_i\rangle - y_i\right)^2 + \lambda\|m\|_2^2$$

with $\mathcal{M} = \mathbb{R}^d$; $\mathcal{X} = \mathbb{R}^{d+1}$ where a single example consists of $(x_i, y_i) \in \mathcal{X}$. Note that the regularization of the model is included in the loss, and should be distinguished from regularization of the weight in our formulations. The subroutine to minimize the weighted loss can be implemented by solving ridge regression over the data scaled according to the weight.

Algorithm 1 The RM alternating minimization algorithm for regularized weighting.

```
1. Input: data $\mu$, $\beta$.
2. Initialization: $m^0 \in \mathcal{M}$
3. While not converged, for time step $s \in (1, 2, \dots)$:
    (a) $R$ step: Find $\mathbf{w}^s = \arg\min_{\mathbf{w}\in\triangle^{\mu}} \left\langle \mathbf{w}, \ell^{m^{s-1}} \right\rangle_{\mu} + \beta \left\| \mathbf{w} - \mathbf{1}_n \right\|_{\mu}^{2}$.
    (b) $M$ step: Find $m^s \in \arg\min_{m' \in \mathcal{M}} \left\langle \mathbf{w}^s, \ell^{m'} \right\rangle_{\mu}$.
4. Output $m^s$.
```

2.1. Implementation

To improve $\mathcal{L}_\beta(m)$, we iterate between Reweighting (R) steps and Minimization (M) steps; as in the Expectation Maximization algorithm, this is an implementation of alternating minimization. As detailed in Algorithm 1, the R steps fix the model m and update $\mathbf{w}$ according to Eq. (1) and the M steps fix the weights and find the model minimizing the weighted average loss using the given subroutine. We have no strong advice for the initialization of m^0; this is problem dependent, and due to non-convexity might affect the model found. The solution for the uniformly weighted problem seems to work well in practice. In the remainder of this section we describe the computational complexity of a single iteration of Algorithm 1.

For convenience, we abuse notations and denote by ℓ^s the loss of model m^s found in step s. The space complexity of RM algorithm (ignoring the input μ and output m^s) is $O(n)$ (the space needed to store $\ell^{s-1}, \mathbf{w}^s$), which are dominated by the size of the input. We now focus on the time complexity. Each R step can be solved in linear time in expectation (see Lemma 13 in the appendix for details), which is dominated by the $O(nd)$ time required to read the data and compute ℓ^{s-1}. Thus, order-wise speaking, the runtime of the RM algorithm is the product of the time of an M step and the number of iterations required, analyzed in Section 3.

3. Linear convergence

In this section we analyze the iteration complexity of the RM procedure. Our main result of the section, presented in Theorem 5, provides a set of sufficient conditions for RM to converge to a local optimum linearly. We remark that throughout this section, the problem type is fixed, including the loss ℓ and a metric space of models $(\mathcal{M}, d_{\mathcal{M}})$, as well as a training sample represented by μ. Hence we drop the subscript μ to reduce clutter, but it should be understood that many terms discussed depend on μ. We define an operator $T_\beta : \mathcal{M} \mapsto \mathcal{M}$ such that $T_\beta(m)$ is the resulting model when starting at a model m and applying one R step followed by one M step, while β is the tradeoff parameter used. Our analysis is based on the following question: given a class of models $M \subseteq \mathcal{M}$ that is closed under T_β, i.e., $T_\beta(M) \subseteq M$,[1] what are the conditions to ensure that $T_\beta^s(m) \overset{s}{\to} \arg\min_{m\in M} \mathcal{L}_\beta(m)$? I.e., repeated applications of T_β converge to the minimizer

[1] Here, $T_\beta(M)$ stands for the image of M under T_β.

of $\mathcal{L}_\beta$ in M? This is important for two reasons: for small M, it leads to rate of convergence to local optimum, and for large M it shows how the right β can preclude both local minima and the influence of outliers.

The proof relies on the fact that if similar models induce similar weights and therefore similar minimization problems, such that the resulting models are even closer, then T_β is a contraction and it generates a sequence that converges fast. The central property of our formulation that we exploit here is that when β is large, reweighting maps similar models (i.e., models with similar loss vectors) to similar weight vectors. Notice that whether T_β is a contraction depends on multiple factors including loss, data, models and β; we introduce the concept of *scope* to allow our analysis to be tightened by localizing it in a sense we specify below.

Definition 2 *Let $M \subset \mathcal{M}$ be a subset of models closed w.r.t. $d_\mathcal{M}$ and $A \subset \mathcal{X}$ be a subset of possible data. We call (M, A) a* scope *at β if M is closed over T_β, and the optimal weights for every $m \in M$ are zero on data outside A.*

We analyze convergence of T_β over a scope where M is chosen as an appropriate neighborhood of the local minimizer m^*_β, and A can be chosen to exclude outliers according to Proposition 1.

To ensure that T_β converges, we require that the M step is stable *vis a vis* small changes in weights. This condition is not always true. For example, when the M step is underdetermined, as in ordinary least squares regression when $n \ll d$ where d is the dimensionality, an arbitrary small change of weights may lead to significant change of the models obtained.

Definition 3 *We say that a problem is g-*determined *for a given $A \subset \mathcal{X}$ if for every $\mathbf{w}^1, \mathbf{w}^2$ supported on A, taking $m^i \in \arg\min_{m \in \mathcal{M}} \left\langle \mathbf{w}^i, \ell^m \right\rangle_\mu$:*

$$d_\mathcal{M}\left(m^1, m^2\right) \leq g \left\| \mathbf{w}^1 - \mathbf{w}^2 \right\|_\mu .$$

The above mentioned linear regression case can be made g determined by adding ridge regularization.

The second property we require is that *uniformly* over M, similar models will have losses that are similar over the data restricted in A.

Definition 4 *We say that a problem has f-*Lipschitz *losses for a given $A \subset \mathcal{X}, M \subset \mathcal{M}$ if for all $m^1, m^2 \in M$ and their corresponding loss vectors ℓ^1, ℓ^2,*

$$\left\| \chi_A \circ \left(\ell^1 - \ell^2\right) \right\|_\mu \leq f d_\mathcal{M}\left(m^1, m^2\right),$$

where χ_A is the indicator function on A, and $\chi_A \circ \ell$ for $\ell \in \mathbb{R}^n$ is a vector $\ell' \in \mathbb{R}^n$ that masks all entries corresponding to data points outside of A, i.e., $\ell'_i = \ell_i$ if $x_i \in A$, and $\ell'_i = 0$ otherwise.

For convergence to the unique minimizer in M it suffices to take β sufficiently large relative to the quantities f and g, as the following theorem shows. The proof uses the Banach fixed point theorem, and is given in Appendix B.

Theorem 5 *If for a given scope (M, A) at β a problem is g determined with f Lipschitz losses, where*

$$\beta > gf,$$

then the restriction of $\mathcal{L}_\beta$ to M, has a unique minimizer m^*_β. Moreover, with this fixed β, RM iterations $(m^s)_{s\in\mathbb{N}}$ converge to m^*_β linearly: for any $m^0 \in M$,

$$d_{\mathcal{M}}\left(m^s, m^*_\beta\right) \leq \frac{1}{2^s} d_{\mathcal{M}}\left(m^0, m^*_\beta\right).$$

Thus, when conditions of Theorem 5 hold, then the RM procedure will converge to the optimum solution in (M, A) linearly if m_0 is initialized in (M, A). If indeed $M = \mathcal{M}$ and $A = \mathcal{X}$, then convergence to the global optimum is guaranteed.

3.1. Examples

We illustrate the results with two classical estimation problems: the location estimation and the linear regression. We discuss the location estimation here, and defer the linear regression example to the appendix.

In the location estimation problem, the squared Euclidean distance is the loss. Our main result is as follows.

Proposition 6 *Given a location estimation problem and consider a scope (M, A) such that $A \subset B(c, r)$ and $M = A$, then the problem is g-determined with $g \leq r$, and has f Lipschitz losses with $f \leq 4r$.*

Thus, if $\beta > 4r^2$, each RM iteration reduces the distance to local optimum by at least half. To prove the theorem, we establish below two general lemmas for location estimation, combining which implies the theorem immediately.

Proposition 6 results from the following two lemmas, establishing f-Lipschitz loss and g-determinedness properties respectively.

Recall that $\chi_A(\cdot)$ is the indicator function, which equivalently can be represented as a vector in $\mathbb{R}^n$ such that its i-th entry is 1 if $x_i \in A$, and 0 otherwise. Thus, for $\ell \in \mathbb{R}^n$, $\chi_A \circ \ell$ is a vector given by the element wise product, which results in a vector ℓ' that masks all entries of ℓ outside of A. The inner product $\langle \chi_A, \ell \rangle_\mu = \frac{1}{n}\sum_i \chi_A(x_i)\ell_i$ where ℓ_i is the i-th entry of ℓ.

Lemma 7 *Let $m^1, m^2 \in \mathcal{X}$ be location estimates, ℓ^1, ℓ^2 the corresponding loss functions, and $A \subset \mathcal{X}$. We denote the average of the estimates $m_a = \left(m^1 + m^2\right)/2$ and ℓ^a the corresponding loss. Then:*

$$\frac{\left\|\chi_A \circ \left(\ell^1 - \ell^2\right)\right\|_\mu}{\left\|m^2 - m^1\right\|_2} \leq 2\sqrt{\langle \chi_A, \ell^a \rangle_\mu}.$$

Thus, if $M = A \subset B(c, r)$, and $m_1, m_2 \in M$, then $\left\|x - \frac{m^1+m^2}{2}\right\|_2 \leq 2r$ for $x \in A$, then $\langle \chi_A, \ell^a \rangle_\mu \leq 4r^2$. Then the problem has f Lipschitz losses for $f = 4r$.

Lemma 8 *Let $\mathbf{w}^1, \mathbf{w}^2$ be supported on A, and let m^1, m^2 be the minimizers for the corresponding weighted problems. Then for any $s \in \mathbb{R}^d$, denoting l^s the loss vector corresponding to it, we have*

$$\frac{\left\|m^1 - m^2\right\|_2}{\left\|\mathbf{w}^1 - \mathbf{w}^2\right\|_\mu} \leq \sqrt{\langle \chi_A, \ell^s \rangle_\mu}.$$

Thus, if $A \subset B(c, r)$, then take $s = c$, then $\sqrt{\langle \chi_A, \ell^s \rangle_\mu} \leq r$, and generally, $fg \leq 4r^2$. Proposition 6 follows immediately.

3.2. Localized convergence analysis

In this section we illustrate, with the following example, that it is often beneficial to refine the analysis by allowing β to decrease over time.

Example 1 *Consider location estimation in $\mathbb{R}$ (hence $\mathcal{M} = \mathcal{X} = \mathbb{R}$) using the squared Euclidean distance as the loss. Suppose that there are 100 data points: 90 of them are in $[-1, 1]$ and the remaining 10 are in $[99, 101]$.*

The question of interest is to identify appropriate range of β to ensure that Algorithm 1 converge linearly for this example. Indeed, we provide two results. The first uses Theorem 5 in a straightforward way, which shows that RM will converge to the global optimal solution for β that is relatively large ($\beta > (102)^2$ specifically). Choosing β too large may be undesirable in rejecting outliers, and our second result, based on applying Theorem 5 iteratively for a sequence of decreasing β, shows that the algorithm can converge to the global optimal solution for β that is much smaller, and consequently the solution completely ignores the outliers.

We start with the first result. Clearly, for any weights, the optimal model belongs to $[-1, 101]$. Thus, for every iteration except the first we can limit our analysis to $M = A = [-1, 101]$, which is a (trivial) scope at any β. Now we apply Lemmas 7 and 8 from Appendix 3.1 to find valid $g = 51$ and $f = 4 \cdot 51$. Theorem 5 therefore guarantees linear convergence in M for any $\beta \geq (102)^2$.

Note that the above argument suggests to choose a relatively large β, which may not be preferable. More specifically, since $\beta \geq (102)^2$, by P2 of Proposition 1, weights are upper bounded by 2, then every model after the second is in $[-1, 21]$. From this point on, the points near 100 are farther than those near 0 and since weights are decreasing in loss according to Eq. (2), we can conclude that every model after the third is in the range $[-1, 11]$. This analysis uses a large β and a scope that is trivial in the sense that it excludes no data. Consequently, the final model obtained assigns non-zero weights to the points near 100, though their weights are lower than those for inliers.

Thus, it is desirable to choose a smaller β to ignore outliers, and still be able to apply Theorem 5 to control convergence. This can be achieved via the following adaptive strategy: if we were to start afresh from $m \in [-1, 11]$, and choose $\beta' = (12)^2$, by P4 of Proposition 1 the weight given to points near 100 would be zero, hence every model henceforth would be in $[-1, 1]$.To analyze the convergence to $m^*_{\beta'}$, we take $A' = M' = [-1, 11]$, and obtain a tighter scope at β'. Lemmas 7 and 8 suggest $g' = 6, f' = 24$ hence $g' \cdot f' \leq \beta'$, so Theorem 5 again shows linear convergence, this time to a solution that completely ignores the outliers.

In this example decreasing β over time allows us to ignore outliers. Such an approach indeed can be applied more generally. Note that if $A' \subset A$ and $M' \subset M$ and for A, M a problem is g determined (has f Lipschitz losses) then it is g' determined (has f' Lipschitz losses) with $g' \leq g$ ($f' \leq f$), enabling fast convergence also with the corresponding smaller β' to $m^*_{\beta'}$. Hence we might apply Algorithm 1 as an inner loop, converging as described for a fixed A, M, β for a few iterations, then conclude a tighter scope applies, and run Algorithm 1 with the corresponding smaller β. As analysing scopes is dependent on the specific losses, we leave such algorithms to future work.

4. Sample complexity for generalization

One advantage of the proposed RW approach, compared to traditional approaches, is that it leads to favorable sample complexity results for generalization when outliers exist, where generalization is measured with the gap between training and testing performance on $\mathcal{L}_\beta$ (Recall $\mathcal{L}_\beta$ is sandwiched between the average loss and the average trimmed loss divided by six). Classical methods considers the average loss, which includes the loss of outliers. Moreover, standard analysis on sample complexity focuses on deriving the sample size such that the *average loss* well approximates the expected loss. This can be expensive when outliers are present, as we show in the below example .

We illustrate with an example the cost of outliers for the classical approach. Consider the problem of location estimation in $\mathbb{R}$ using the square loss, where the data are generated by a Gaussian mixture

$$0.99\mathcal{N}\left(0,1\right)+0.01\mathcal{N}\left(10^5,1\right).$$

For a model close to 0, about 99% of points incur losses in $(0,10)$, and the remaining one percent are around 10^{10}. The expected loss will be of order of 10^8, and the deviation of the average loss from the expected loss is dominated by the random number of outliers in the sample. Thus, to ensure that the deviation is significantly smaller than a typical loss (i.e., to be order of $\epsilon \approx 1$), we need a sample of size of $(B/\varepsilon)^2 \approx 10^{20}$ where B is an upper bound on the losses. Some recent work (discussed in Section 5) on the sample complexity of the average loss manages to replace the upper bound B in the sample complexity with the 2nd or 4th order moment conditions, but since the moments themselves scale with the magnitude of the outliers super-linearly, the sample complexity results obtained remain sensitive to the magnitude of even a small fixed proportion of outliers.

In contrast, the proposed approach leads to analysis of generalization which depends on the "typical" loss, and is almost independent *vis a vis* B the upper bound of the loss. Recall that generalization is about bounding the performance gap between training and testing. Embedding this in our context, let μ and μ' be two independent copies of empirical distributions of n i.i.d. draws from an unknown distribution ν, then we say that $\mathcal{L}_\beta$ generalizes when provably $\mathcal{L}_\beta\left(m;\mu'\right) \approx \mathcal{L}_\beta\left(m;\mu\right)$ over all m in a class of useful models defined below. Since $\mathcal{L}_\beta$ tends to ignore outliers, outliers can only affect generalization if their proportion in the sample is much larger than expected, which happens very rarely. Indeed, it can be shown that B the bound on outlier losses only affects sample complexity logarithmically: for an appropriately chosen, d dimensional set of models, we show

$$O\left(\left(\beta/\varepsilon\right)^2 d\log\left(B/\varepsilon\right)\right) \tag{3}$$

data points suffice for $\mathcal{L}_\beta$ to generalize. Notice the bound depends polynomial only on parameter β, which is set to a typical loss per Proposition 1. This improvement is significant when $\beta \ll B$, hence only for those models *having low typical losses*. To illustrate, consider the above example. Here a "good" model such as $m \in [-1,1]$ has more than 68% of points with losses smaller than $\beta = 10$ so for $\varepsilon \approx 1$ the polynomial term in Eq. (3) is under 100. In contrast, for a "bad" model such as $m > 50000$, their 68th percentile of losses is close to 10^{10}, which implies that β is close to B, thus rendering the advantages of Eq. (3) vacuous.

To address this challenge, we show that by applying a peeling technique, we are able to relate the generalization ability of $\mathcal{L}_\beta$ with only good and yet achievable models, of the following form:

Definition 9 *The class of models that for measure σ have at least p fraction of their losses below b is denoted by*

$$\mathcal{M}_p^{\sigma,b} = \left\{m \in \mathcal{M} : P_{x\sim\sigma}\left(\ell\left(m,x\right)-\rho_l^m < b\right) > p\right\},$$

where $\rho_l^m = \min_{x\in\mathcal{X}} \ell\left(m,x\right)$.

The result below shows that $\mathcal{L}_\beta$ is highly concentrated around *its own* expectation, i.e., it generalizes uniformly over $\mathcal{M}^{\mu,\beta}_{2/3+2s}$, where s is a small "safety buffer" .

To present the theorem, some notations are in order. We assume throughout that the losses of every model are bounded above such that $\max_{x\in\mathcal{X}} \ell(m,x) - \rho_l^m \leq B$. We treat the class of all models $\mathcal{M}$ as a metric space with the uniform metric induced by ℓ over $\mathcal{X}$: $d_{\mathcal{X}}(m,m') = \sup_{x\in\mathcal{X}} |\ell(m,x) - \ell(m',x)|$. We define $h(x) = x\log x$. Recall the standard definition of ϵ-cover number (Van der Vaart and Wellner, 1996).

Definition 10 *Let $\mathcal{S}$ be a metric space equipped with metric d, and let $S \subset \mathcal{S}$. Fix $\epsilon > 0$, a ϵ cover of S with respect to d is a set of balls with diameter ϵ wrt d (not necessarily disjoint), such that their union is a superset of S. The ϵ cover number of S, denoted by (S,d) is the smallest integer n such that their exists an ϵ cover of S wrt d that contains no more than n balls.*

Theorem 11 *Assume that $\forall\varepsilon > 0$, $(\mathcal{M}, d_{\mathcal{X}})$ has ε cover number upper bounded by $(CB/\varepsilon)^d$. Let μ, μ' be empirical distributions of n elements distributed according to ν and fix $\beta > \varepsilon > 0$. Taking $s = \frac{\varepsilon}{16\beta}$, set $p \geq 2/3 + n^{-1} + s$. Suppose that*

$$n > \max\left\{2s^{-2}\log\left(B/\left(24\beta\right)\right), 3h\left(4s^{-2}+3\right)\right\}.$$

Then with probability $1 - \exp\left(\log 20 + d\log\left(CB/\varepsilon\right) - \frac{n}{8}\left(\varepsilon/\left(24\beta\right)\right)^2\right)$, we have

$$\forall m \in \mathcal{M}^{\nu,\beta-\varepsilon}_{p+s}: \quad \left|\mathcal{L}_\beta\left(m;\mu\right) - \mathbb{E}_{\mu'}\mathcal{L}_\beta\left(m;\mu'\right)\right| \leq 3\varepsilon.$$

Remark 12 *Theorem 11 states that a sample of size $O\left(\left(\frac{\beta}{\varepsilon}\right)^2\left(d\log\left(\frac{B}{\varepsilon}\right) + \log\delta^{-1}\right)\right)$ suffices to generalize. This sample complexity depends on ν (which is not directly accessible except via μ) through the set of "good" models $\mathcal{M}^{\nu,\beta}_{p+s}$. If $m \in \mathcal{M}^{\mu,\beta}_{p+2s}$, it is also in $\mathcal{M}^{\nu,\beta}_{p+s}$ with probability at least as high as that of the theorem holding.*

We illustrate Theorem 11 revisiting the example from early in this section. The guarantee of Theorem 11 is that with high probability, uniformly over models m from a set, $\mathcal{L}_\beta(m,\mu)$ is very close to its expectation over independent samples μ' created in the same way as μ. In other words, however we find m, the measure of quality $\mathcal{L}_\beta(m,\mu)$ we propose for m is essentially not dependent on the randomness in the sample μ.

More concretely, let m be any point in $\mathbb{R}$, the corresponding vector of losses are the squared differences between m and x_i. The choice of $\beta = 10, \epsilon = 1, p = 2/3 + n^{-1} + s$ implies $s = 1/160$. This choice of β implies (By proposition 1) that we will give weight zero to points from one of the two sources, as the sources are too far to both be within 4β of m. If a model is near the mean of the unignored source, weights will be near uniform over points from it. The theorem applies to $m \in \mathcal{M}^{\nu,\beta-\varepsilon}_{p+s}$: all models within distance 3 of a fraction at least $p+s$ of the weight of ν. This excludes the vicinity of the outlier source, as there is not sufficient weight there, hence the guarantee applies approximately to models in the interval [-2,2], defining a set of "good" models. To apply this theorem or any that assumes the losses are bounded by B, we need roughly $B \geq 10^{10}$. The sample size required is dominated by $2s^{-2}\log(B/24\beta)$, so is of the order of 10^6, rather than the 10^{20} mentioned in the beginning of the section for concentration of a mean of losses.

5. Related work

In this section we discuss related work in literature. The central idea of this paper, namely Regularized Weighting, was originally formulated in Vainsencher et al. (2013) for learning *multiple* models (which includes clustering as a special case). Instead, in this paper we focus on the (special) case of learning *one* model with outlying observations, and hence obtain stronger structural and theoretical results. In particular, Vainsencher et al. (2013) did not establish results about convergence of the computation algorithm, and its sample complexity results hold only when weights are given to points with near maximal loss. In contrast, we show any point with loss significantly higher than typical will be excluded.

This paper is motivated from two (related) main considerations: meaningful sample complexity with very large B, and robustness of statistical procedure to outliers. The majority of research in statistical learning theory to establish sample complexity (e.g., Anthony and Bartlett, 1999) relies on the existence of a uniform upper bound B on all losses. This results in either obtaining loose sample complexity bounds for large B, or making an assumption that B is small which becomes unrealistic in the presence of heavy tailed noise, let alone outliers. As such, there has been a recent surge in developing methods and analysis to achieve meaningful statistical guarantees without assuming a small B. For example, Mendelson (2014) developed a refined analysis of the Empirical Risk Minimization (ERM) procedure without resort to concentration of measure technique, and hence is able to derive performance bounds where B is replaced by a norm of (possibly unbounded) noises. Audibert and Catoni (2011), using the PAC-Bayesian analysis, also bypasses the requirement of a bounded B inherent to concentration of measure technique, and established risk bounds under assumption of boundedness of certain *moments* of the noise, for ERM, ridge regression, as well as for a novel estimator based on truncating difference of losses which has better deviations. Hsu and Sabato (2016) achieved a similar goal, using a generalized median of mean approach, which is more computationally efficient compared with truncating difference of losses. Essentially, these works consider the following task: how to control the large deviation from the expected behavior with (exponentially) high probability, where only low-order moments such as variance are given. Admittedly this is a fundamental problem, it does not directly address the outlier setup, for the following reasons: (1) all data points are assumed to be iid; (2) more importantly, the performance guarantees obtained still heavily depends on the magnitude of outliers. In particular, when a fixed percentage of (large) outliers exist, then any moments may still scale super-linearly with the size of outliers, making these results vacuous. The outlier case is in sharp contrast to the heavy noise scenario which is the focus of these works, where the moments can scale much slower than the upper bound of the loss.

The study of statistical procedures robust to the effect of outliers dates back to the 50's by Box (1953), followed by Tukey (1960, 1962), Huber et al. (1964) and Hampel (1968), among others. A powerful paradigm termed M-estimators was developed by Huber et al. (1964), in which asymptotic properties of a class of statistical estimators (including notably the maximum likelihood estimators) are established, along with initial theory for constructing robust regression procedures. One M-estimator that is particularly close to our method is the classical Least Trimmed Squares (LTS) (Rousseeuw, 1984) estimator, which defines the loss of the model to be the average of a fixed proportion of its smallest losses on the data. This formulation is conceptually close to ours, to the extent that our proof can be easily adapted to provide the first (to our knowledge) finite sample results for LTS in the presence of outliers, despite extensive study of this and related methods.

The main advantage of the proposed RM algorithm over LTS is algorithmic: the local alternating minimization algorithm for LTS (see concentration steps in Rousseeuw and Van Driessen (2000)), is only known (to our knowledge) to terminate after a finite (but possibly exponentially large) number of steps; see Nguyen and Welsch (2010) for a review of attempts at efficient solution methods.

Recently, several works studied M-estimators from the perspective of high dimensional statistics, using the unified framework of *restricted strong convexity*, and established statistical and optimization properties of such estimators (Loh and Wainwright, 2015, 2014). Yet, the M-estimators studied in those papers are restricted to finite sample versions of functions that are convex in the population limit, which essentially studies formulations that are “close to being convex”, where this approximate convexity is established through assumptions on the formulation *and the data distribution*. Loh (2015) extended this analysis to the case where such approximate convexity is only required to hold at a neighborhood of the optimal solution. Even with such extension, this framework, albeit with nice iteration complexity, may be restrictive. For example, it is not applicable to the above discussed LTS estimator. Moreover, these work concentrated on the regression problem, and are not easily adapted to other problems of interest, such as Principal Component Analysis. In comparison, our approach aims are more general: our assumptions are minimal, and our approach works for a wide spectrum of statistical problem, provided that for each model and a given data point, the loss can be evaluated.

Aslan et al. (2012) extended the M-estimator framework, aiming to simultaneously achieve computation efficiency and robustness to outliers. They analyzed a variational representation of M-estimators, which includes many standard robust methods as special cases, and proposed computationally efficient relaxation strategies based this representation. The approximation algorithm runs in polynomial time, but requires solving a convex problem where computing a gradient costs $O\left(n^3\right)$ time, and hence may not scale. The authors further proposed a more scalable alternating minimization procedure (similar in spirit to ours), but did not discuss the rate of convergence. Moreover, the proposed method only achieves modest robustness guarantees: the authors showed that the proposed method will not breakdown if there is *one* arbitrary outlier. Beyond this case, no robustness guarantees are offered. Note that some variants of their formulation appear similar to ours, with the key difference that weights are not constrained to form a distribution.

The above reviewed work all focus on the linear regression case. A general and practical approach to robust estimation is l_1 penalized outlier correction (Giannakis et al., 2011; Mateos and Giannakis, 2012a,b) which models the observed data x as the sum of a sparse perturbation term o and an adjusted data term. This method alternates between computing o by a shrinkage operator, and applying the original estimation problem to the adjusted data term to find a model, which leads to scalable algorithms. Yet, convergence to global optimality is not guaranteed. Moreover, neither the rate for local convergence, nor the sample complexity are established, which is different from our proposed approach.

Finally, we remark that there are a few published works on the sample complexity of methods robust to outliers, all of which concentrate on the robust PCA task. For example Coudron and Lerman (2012) discusses the sample complexity of an algorithm for robust PCA where the data are sub-Gaussian; and Xu et al. (2013) propose algorithms for PCA robust to outliers, have sample complexity bounds and are efficient, in a setting very different from ours.

6. Conclusion

In this paper we studied a *general* approach for learning under outliers, that iterates between finding a solution with good empirical performance and re-weighting the data via regularized optimization. Intuitively, this will encourage the obtained solution to be biased towards easier data points and ignoring outliers, to a controlled extent. We validate this intuition by establishing guarantees for generalization and iteration complexity that essentially ignores the presence of outliers for the proposed approach. To the best of knowledge, this is the first scalable, general and robust estimation procedure with finite sample guarantees and linear convergence. Moreover, our analysis for estimating sample complexity are general and can be adapted to other robust estimators, to establish sample complexity *that ignore data points with large losses*.

The regularized re-weighting formulation proposed in this paper is tailored for mitigating the effect of large losses. An immediate direction to explore for future research is whether the weight uniformity can be traded off against other quantities, such as large effect on the optimal model, which would encourage algorithmic stability.

Acknowledgments

We thank the reviewers for constructive comments which greatly help improving the paper. The research of SM was partially supported by the Israel Science Foundation under contract 1380/16.

Appendix A. Proof of Proposition 1.

We begin by proving P1 Proposition 1 below and illustrate the property through a simple example. Then in Subsection A.1, we prove the rest Proposition 1 through a sequence of lemmas.

Lemma 13 *Let ℓ be a vector of losses, the optimal weights (solution of R step) can be computed in $O(n)$ expected time by finding $a \in \mathbb{R}$ such that:*

$$\mathbf{w} = P_{\triangle}\left(-\ell\,(2\beta)^{-1}\right) = \left[-\ell\,(2\beta)^{-1} + a\mathbf{1}_n\right]_+ . \tag{4}$$

Proof For a fixed model, problem (1) becomes

$$\min_{\mathbf{w}\in\triangle} \langle \mathbf{w}, \ell \rangle + \beta \left\|\mathbf{w} - \mathbf{1}_n\right\|^2 ,$$

which is minimizing a quadratic function over the simplex. The unconstrained optimum is at $\mathbf{1}_n - \ell\,(2\beta)^{-1}$ and the squared distance of any solution from this unconstrained optimum coincides with the sub-optimality of the solution in terms of the objective value. Thus, the optimal solution constrained in the simplex is is the projection of the unconstrained optimum to the simplex – i.e., Equation (4), since it minimizes the Euclidean distance. We can easily find the optimal a in time $O(n \log n)$ using a binary search; Duchi et al. (2008) remove this logarithmic factor using a trick similar to that for $O(n)$ median calculation. ∎

For a fixed ℓ, for any i, j, $|\mathbf{w}_i - \mathbf{w}_j|$ decreases at a rate of $(2\beta)^{-1}$. Thus, $\mathbf{w}$ tends to uniform if β is chosen large enough. This effect, formalized in the above lemma and Proposition 1, can be seen in Fig. 2, which applies to the example displayed in Fig. 1. We illustrate the effect of reweighting the data in Fig. 3: the weighted means of data induced by both initial models ignore the extreme outliers.

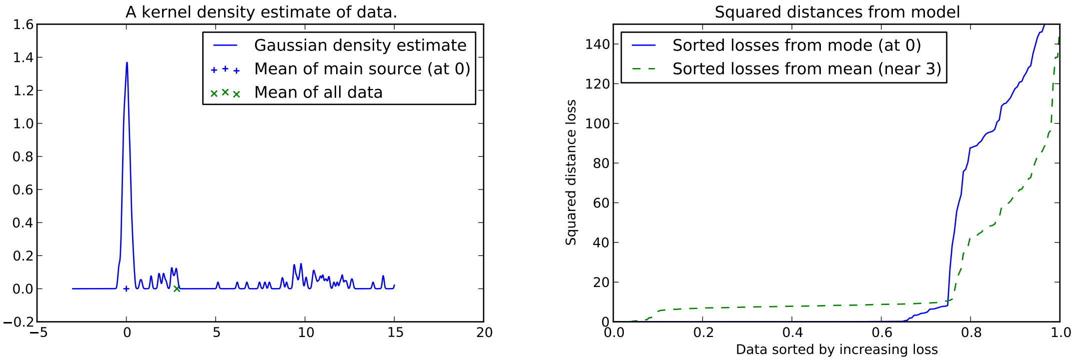

Figure 1: A mixture of data is generated, where most data are concentrated around 0 except for some outliers. Two models are considered, namely the mode (located at 0 and represented using"+") and the empirical mean (located near 3 and represented using"x"). The left figure illustrates a kernel density estimate using the data. The right figure shows the sorted losses of both models on the data. Observe that the mode is more precise on most of the data, but has larger losses on outliers.

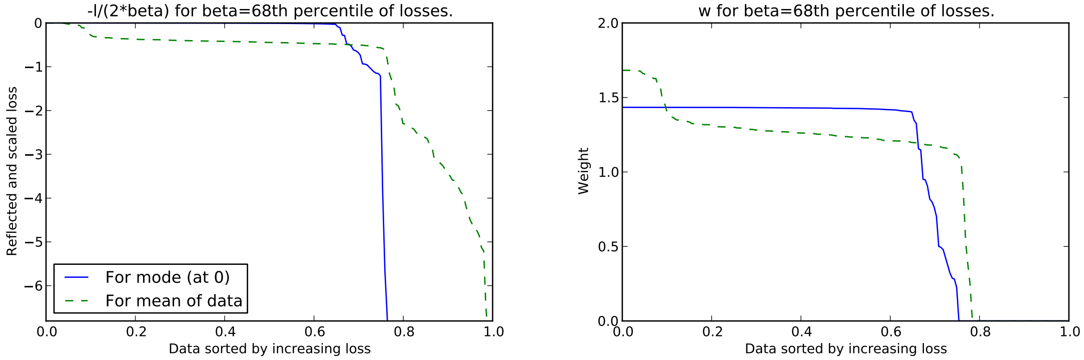

Figure 2: For the two models in Figure 1; Left: their losses are reflected about 0 and scaled by $(2\beta)^{-1}$ on the left; Right: the consequent projection into the simplex augments them by a and truncates negative weights (see Lemma 13).

A.1. Structural results

In this section we prove the remaining claims in Proposition 1 via establishing a list of lemmas progressively.

Lemma 14 *Assume that proportion $p \in [0,1]$ of all losses are in the interval $[\rho_l, \rho_h]$, and that $\beta = c(\rho_h - \rho_l)$. Let $\mathbf{w}(x) = \left[-\ell(x)(2\beta)^{-1} + a\right]_+$ be the optimal weights, then we have $a \leq p^{-1} + \rho_h (2\beta)^{-1}$. Furthermore, if $\ell(x) > \rho_h + 2p^{-1}\beta$, then $\mathbf{w}(x) = 0$.*

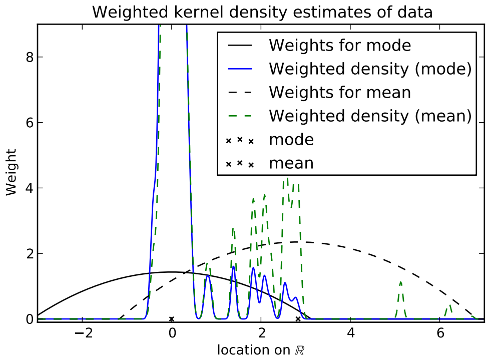

Figure 3: The weights obtained, and the weight adjusted density of data. Compare with Figure 1, we observe that both models result in ignoring the outliers at the extreme right.

Proof Losses in the interval receive weights of at least $a - \rho_h (2\beta)^{-1}$, then $1 = \langle 1, \mathbf{w} \rangle \geq \left(a - \rho_h (2\beta)^{-1}\right) p \Rightarrow a \leq p^{-1} + \rho_h (2\beta)^{-1}$, completing the first part. Using this upper bound on a, we find

$$\begin{aligned}\mathbf{w}(x) &= \left[-\ell(x)(2\beta)^{-1} + a\right]_+ \\ &\leq \left[-\ell(x)(2\beta)^{-1} + p^{-1} + \rho_h (2\beta)^{-1}\right]_+ \\ &= \left[\left(2p^{-1}\beta + \rho_h - \ell(x)\right)(2\beta)^{-1}\right]_+ .\end{aligned}$$

Thus if $\ell(x) > \rho_h + 2p^{-1}\beta$, we have $\mathbf{w}(x) = 0$, which established the second part. ■

Lemma 15 *Under the conditions of Lemma 14, and further assume that losses are lower bounded by $\ell \geq \rho_l$. Then $\mathbf{w}(x) \leq p^{-1} + (2c)^{-1}$.*

Proof The lemma holds by the following:

$$\begin{aligned}\mathbf{w} &= \left[-\ell (2\beta)^{-1} + a\right]_+ \\ &\leq \left[-\rho_l (2\beta)^{-1} + a\right]_+ && (\rho_l \text{ is a lower bound on loss}) \\ &\leq \left[-\rho_l (2\beta)^{-1} + p^{-1} + \rho_h (2\beta)^{-1}\right]_+ && (\text{Lemma } 14) \\ &= p^{-1} + (\rho_h - \rho_l)(2\beta)^{-1} && (\text{non negative}) \\ &= p^{-1} + (2c)^{-1} && (\text{definition of } c) .\end{aligned}$$

■

Lemma 16 *Under the conditions of Lemma 15, and further assume that* $p \geq 2/3$ *and* $c \geq 1$*. Then points with losses in the interval* $[\rho_l, \rho_h]$ *receive weight at least*

$$\left(2 - p^{-1}\right) p^{-1} - (2c)^{-1} \geq 1/4.$$

Proof From Lemma 14 we have that $-\rho_h (2\beta)^{-1} + a \leq p^{-1}$. Since $[\cdot]_+$ affects only negative entries, we conclude that losses greater than ρ_h receive weight at most

$$\left[-\rho_h (2\beta)^{-1} + a\right]_+ \leq p^{-1}.$$

Then denoting $I = \{i : \ell_i \in [\rho_l, \rho_h]\ \}$, we have the following bound

$$1 = \langle 1, \mathbf{w} \rangle \leq p|I|^{-1} \sum_{i \in I} \left[-\ell (2\beta)^{-1} + a\right]_+ + (1-p)\, p^{-1}, \tag{5}$$

where the first term and the second terms bounds the weight associated with small and large losses respectively. This leads to

$$\begin{aligned} 2 - p^{-1} =& 1 - (1-p)\, p^{-1} \\ \leq& p|I|^{-1} \sum_{i \in I} \left[-\ell (2\beta)^{-1} + a\right]_+ && \text{(Equation (5))} \\ \leq& p|I|^{-1} \sum_{i \in I} \left[-\rho_l (2\beta)^{-1} + a\right]_+ && (\text{def } \rho_l) \\ =& p \left[-\rho_l (2\beta)^{-1} + a\right]_+ && (\text{def } p) \\ =& p \left(a - \rho_l (2\beta)^{-1}\right) \end{aligned}$$

with the last equality holds because $\left[-\rho_l (2\beta)^{-1} + a\right]_+$ is an upper bound on all elements of $\mathbf{w}$, and hence must be positive.

Thus, by $2 - p^{-1} \leq p \left(a - \rho_l (2\beta)^{-1}\right) \iff \left(2 - p^{-1}\right) p^{-1} + \rho_l (2\beta)^{-1} \leq a$, we obtain a lower-bound of weights for losses under ρ_h by the following using Lemma 14

$$\left[\left(2 - p^{-1}\right) p^{-1} + (\rho_l - \rho_h)(2\beta)^{-1}\right]_+ .$$

By the definition of c, and the assumption $c \geq 1$, we have $(\rho_l - \rho_h)(2\beta)^{-1} = -(2c)^{-1} \geq -1/2$. The term $\left(2 - p^{-1}\right) p^{-1}$ is concave in p^{-1}, hence is minimized on the extrema of its range; by assumption $p \in [2/3, 1]$, then $p^{-1} \in [1, 3/2]$, and the minimum is 3/4 at $p = 2/3$. Summing these bounds, the lower bound on weights for losses in the interval is $1/4$. ■

Now we prove Proposition 1.

Proof The conditions of Lemmas 14, 15 and 16 hold, then so do their conclusions. In particular for $i \in I$, $\mathbf{w}_i \geq 1/4$. Then

$$\begin{aligned}
\mathcal{L}_\beta \left(m; \mu\right) - \rho_l &= \left\langle \mathbf{w}, \ell^m \right\rangle_\mu + \beta \left\| \mathbf{w} - \mathbf{1}_n \right\|_\mu^2 - \rho_l \\
&\geq \left\langle \mathbf{w}, \ell^m \right\rangle_\mu - \rho_l \\
&= n^{-1} \sum_{i=1}^{n} \mathbf{w}_i \left(\ell_i^m - \rho_l\right) \\
&\geq n^{-1} \sum_{i \in I} \frac{1}{4} \left(\ell_i^m - \rho_l\right) && \text{(Lemma 16)} \\
&= p \left|I\right|^{-1} \sum_{i \in I} \left(\ell_i^m - \rho_l\right)/4 \\
&\geq \left|I\right|^{-1} \sum_{i \in I} \left(\ell_i^m - \rho_l\right)/6. && \text{(By assumption } p \geq 2/3\text{)}
\end{aligned}$$

■

Appendix B. Proof of linear convergence

We first establish the main property of the R step, that similar losses lead to similar weights, and more so for larger β, and moreover ignored data have no effect on the resulting weight. Next we use it to prove Theorem 5.

Lemma 17 *Let $\beta > 0$ and ℓ^1, ℓ^2 loss functions whose corresponding weights after an R step $\mathbf{w}^1, \mathbf{w}^2$ are supported on $A \subset \mathcal{X}$, then*

$$\left\| \mathbf{w}^1 - \mathbf{w}^2 \right\| \leq (2\beta)^{-1} \left\| \chi_A \circ \left(\ell^1 - \ell^2\right) \right\|.$$

Proof We use the following natural consequence of Equation (2): losses in ℓ that receive weight zero may be increased arbitrarily without affecting the corresponding w, as they will keep receiving zero weight. We so construct (at the end of this proof) $\bar{\ell}^1, \bar{\ell}^2$ that outside A equal one another, and on A equal the corresponding ℓ^j. Then:

$$\begin{aligned}
\left\| \mathbf{w}_1 - \mathbf{w}_2 \right\|_\mu^2 &= \left\| P_{\triangle^\mu}\left(-\tfrac{\ell^1}{2\beta}\right) - P_{\triangle^\mu}\left(-\tfrac{\ell^2}{2\beta}\right) \right\|_\mu^2 \\
&= \left\| P_{\triangle^\mu}\left(-\tfrac{\bar{\ell}^1}{2\beta}\right) - P_{\triangle^\mu}\left(-\tfrac{\bar{\ell}^2}{2\beta}\right) \right\|_\mu^2 \\
&\leq \left\| \tfrac{\bar{\ell}^1}{2\beta} - \tfrac{\bar{\ell}^2}{2\beta} \right\|_\mu^2 && (P_{\triangle^\mu} \text{ is a contraction}) \\
&= (2\beta)^{-2} \left\| \chi_A \cdot \left(\ell^1 - \ell^2\right) \right\|_\mu^2.
\end{aligned}$$

To construct the modified $\bar{\ell}$ we use the explicit form for weights:

$$P_{\triangle^\mu}\left(-\frac{\ell}{2\beta}\right) = \left[-\frac{\ell}{2\beta} + a\right]_+,$$

with a chosen so that $\|\mathbf{w}\|_\mu = 1$, hence write

$$\mathbf{w}^1 = \left[-\frac{\ell^1}{2\beta} + a^1\right]_+ ; \mathbf{w}^2 = \left[-\frac{\ell^2}{2\beta} + a^2\right]_+ .$$

We define

$$\bar{\ell}^j(x) = \begin{cases} \ell^j(x) & x \in A \\ 2\beta\bar{a} & \text{otherwise} \end{cases}$$

where $\bar{a} = \max\left\{a^1, a^2\right\}$. We note two facts. First, all the losses that received weight in $\mathbf{w}^j$ remain unchanged. Second, $-\frac{2\beta\bar{a}}{2\beta} + a^j \leq 0$ so for each loss in ℓ^j that received no weight in $\mathbf{w}^j$ remains unweighted in $\bar{\ell}^j$. We conclude that

$$\left\| \left[-\frac{\bar{\ell}^j}{2\beta} + a^j\right]_+ \right\|_\mu = 1$$

with the same a^j, and $P_{\triangle^\mu}\bar{\ell}^j = \mathbf{w}^j$ as wanted. ■

We now prove Theorem 5.

Proof We consider two starting points m_p^1, m_p^2, corresponding to the losses ℓ^1, ℓ^2 respectively. These losses induce the respective weight vectors $\mathbf{w}^1, \mathbf{w}^2$, which in turn induce the optimal models m^1, m^2 for the next iterations, which induce the losses in the next iteration $\ell^{1\prime}, \ell^{2\prime}$.

Then applying the assumptions and Lemma 17:

$$\begin{array}{lll} & d_{\mathcal{M}}\left(m_p^1, m_p^2\right) & \\ \geq & f^{-1}\left\|\chi_A \circ \left(\ell^1 - \ell^2\right)\right\|_\mu & \text{(Lipschitz losses)} \\ \geq & f^{-1}(2\beta)\left\|\mathbf{w}^1 - \mathbf{w}^2\right\|_\mu & \text{(Lemma 17)} \\ \geq & f^{-1}(2\beta)\, g^{-1} d_{\mathcal{M}}\left(m^1, m^2\right), & (g \text{ determined}) \end{array}$$

which leads to

$$\frac{d_{\mathcal{M}}\left(m_p^1, m_p^2\right)}{d_{\mathcal{M}}\left(m^1, m^2\right)} \geq 2\beta/(fg).$$

Then when $\beta \geq g(\mu, A) f(\mu, M, A)$, $d_{\mathcal{M}}$ is reduced by half at each iteration. By the Banach fixed point theorem, RM iterations restricted to M then have a unique fixed point we denote by m_β^*. Take $m_p^2 = m_\beta^*$, then we have the outputs of each iteration never leave $B_{d_{\mathcal{M}}}\left(m_\beta^*, r_0\right)$; hence they converge to m_β^* with regard to $d_{\mathcal{M}}$. Because RM iterations are monotone decreasing in $\mathcal{L}_\beta$, m_β^* must be optimal within M. Otherwise, taking m_p^1 as the optimal within M leads to a contradiction. ■

B.1. Proofs of results in Section 3.1

Proof of Lemma 7

$$\begin{aligned}
\left\|\chi_A \circ \left(\ell^1 - \ell^2\right)\right\|_\mu^2 =& \mathbb{E}_{x\sim\mu}\chi_A\left(x\right)\left(\left\|x-m^1\right\|_2^2 - \left\|x-m^2\right\|_2^2\right)^2 \\
=& \mathbb{E}_{x\sim\mu}\chi_A\left(x\right)\left(\left\langle x-m^1, x-m^1\right\rangle - \left\langle x-m^2, x-m^2\right\rangle\right)^2 \\
=& 4\mathbb{E}_{x\sim\mu}\chi_A\left(x\right)\left\langle m^2-m^1, x-\frac{m^1+m^2}{2}\right\rangle^2 \\
=& 4\mathbb{E}_{x\sim\mu}\chi_A\left(x\right)Tr\left(\left(m^2-m^1\right)^\top\left(x-\frac{m^1+m^2}{2}\right)\left(x-\frac{m^1+m^2}{2}\right)^\top\left(m^2-m^1\right)\right) \\
=& 4\mathbb{E}_{x\sim\mu}\chi_A\left(x\right)Tr\left(\left(m^2-m^1\right)\left(m^2-m^1\right)^\top\left(x-\frac{m^1+m^2}{2}\right)\left(x-\frac{m^1+m^2}{2}\right)^\top\right) \\
=& 4Tr\left(\left(m^2-m^1\right)\left(m^2-m^1\right)^\top\left(\mathbb{E}_{x\sim\mu}\chi_A\left(x\right)\left(x-\frac{m^1+m^2}{2}\right)\left(x-\frac{m^1+m^2}{2}\right)^\top\right)\right) \\
\leq& 4\left\|m^2-m^1\right\|_2^2\left\|\mathbb{E}_{x\sim\mu}\chi_A\left(x\right)\left(x-\frac{m^1+m^2}{2}\right)\left(x-\frac{m^1+m^2}{2}\right)^\top\right\|_{2,2} \\
& \qquad\text{(Von Neumann trace ineq.)} \\
\leq& 4\left\|m^2-m^1\right\|_2^2\mathbb{E}_{x\sim\mu}\chi_A\left(x\right)\left\|\left(x-\frac{m^1+m^2}{2}\right)\left(x-\frac{m^1+m^2}{2}\right)^\top\right\|_{2,2} \\
& \qquad\text{(Jensen's ineq.)} \\
=& 4\left\|m^2-m^1\right\|_2^2\mathbb{E}_{x\sim\mu}\chi_A\left(x\right)\left\|x-\frac{m^1+m^2}{2}\right\|_2^2 \\
=& 4\left\|m^2-m^1\right\|_2^2\left\langle\chi_A, \ell^a\right\rangle_\mu .
\end{aligned}$$

■

Proof of Lemma 8

$$
\begin{aligned}
\left\|m^1 - m^2\right\|_2 &= \left\|\mathbb{E}_{x\sim\mu}\mathbf{w}^1(x)\,x - \mathbb{E}_{x\sim\mu}\mathbf{w}^2(x)\,x\right\|_2 \\
&= \left\|\mathbb{E}_{x\sim\mu}\mathbf{w}^1(x)\,(x-s) - \left(\mathbb{E}_{x\sim\mu}\mathbf{w}^2(x)\,(x-s)\right)\right\|_2 \\
&= \left\|\mathbb{E}_{x\sim\mu}(x-s)\cdot\left(\mathbf{w}^1-\mathbf{w}^2\right)(x)\right\|_2 \\
&= \left\|\mathbb{E}_{x\sim\mu}\chi_A(x)\,(x-s)\cdot\left(\mathbf{w}^1-\mathbf{w}^2\right)(x)\right\|_2 \\
&\qquad \left(\mathbf{w}^1,\mathbf{w}^2 \text{ are supported on } A\right) \\
&\leq \mathbb{E}_{x\sim\mu}\chi_A(x)\left\|x-s\right\|_2\left|\mathbf{w}^1(x)-\mathbf{w}^2(x)\right| \\
&\qquad \text{(Jensen's inequality)} \\
&\leq \sqrt{\mathbb{E}_{x\sim\mu}\chi_A^2(x)\left\|x-s\right\|_2^2}\sqrt{\mathbb{E}_{x\sim\mu}\left|\mathbf{w}^1(x)-\mathbf{w}^2(x)\right|^2} \\
&\qquad \text{(Cauchy Schwartz inequality)} \\
&= \sqrt{\left\langle\chi_A,\ell^s\right\rangle_\mu}\left\|\mathbf{w}^1-\mathbf{w}^2\right\|_\mu .
\end{aligned}
$$

■

B.2. Linear Convergence for Ridge Regression

In regression, each data point is a pair $(x,y)\in\mathcal{X}$, and similarly the empirical distribution μ is also on the pairs (x,y). Thus weight and loss functions are also defined over pairs (x,y). For a model m, we consider the ridge regression loss function $\ell^m(x,y) = \lambda\langle m,m\rangle + \left(\langle m,x\rangle - y\right)^2$. For simplicity, we keep λ fixed throughout. We use the Euclidean norm as the metric between models so $d_{\mathcal{M}}(m,m') = \|m-m'\|_2$. We denote $\xi\in\mathbb{R}^n$ the feature vector norms $\|x_i\|_2$ over the data set μ, and $\zeta\in\mathbb{R}^n$ the scaled norms $y_i\|x_i\|_2$. The below two lemmas establish the f-Lipschitz and g-determinedness property respectively.

Lemma 18 *Let m^1,m^2 be two regressors, $m^a = \left(m^1+m^2\right)/2$ their average, and let ℓ^1,ℓ^2,ℓ^a their corresponding loss functions, $A\subset\mathcal{X}$. Then:*

$$
\frac{\left\|\chi_A\circ\left(\ell^1-\ell^2\right)\right\|_\mu^2}{\left\|m^1-m^2\right\|_2^2} \leq 4\left\|\chi_A\circ\xi^2\right\|_\mu\left\|\chi_A\circ\ell^a\right\|_\mu .
$$

Lemma 19 *Let $(x,y)\sim\mu$ in $\mathbb{R}^{d+1}$ and $\mathbf{w}^1,\mathbf{w}^2\in\triangle^\mu$ be weights supported on $A\subset\mathcal{X}$. Assume that*

$$
\left\|\mathbf{w}^1-\mathbf{w}^2\right\|_\mu\left\|\chi_A\circ\xi^2+\lambda\right\|_\mu < \lambda/2,
$$

and m^1,m^2 be models minimizing the weighted problems then for every p,q such that $p^{-1}+q^{-1}=1$:

$$
\frac{\left\|m^1-m^2\right\|_2}{\left\|\mathbf{w}^1-\mathbf{w}^2\right\|_\mu} \quad\leq\quad \lambda^{-1}\left(\left\|\chi_A\circ\zeta\right\|_\mu + 2\lambda^{-1}\left\|\chi_A\circ\xi^2+\lambda\right\|_\mu\left\|\mathbf{w}^1\right\|_{L^p(\mu)}\left\|\chi_A\circ\zeta\right\|_{L^q(\mu)}\right).
$$

Here $\|\mathbf{z}\|_{L^p_{(\mu)}} \triangleq \left(\frac{1}{n}\sum_{i=1}^n (z_i)^p\right)^{\frac{1}{p}}$, for any $\mathbf{z}\in\mathbb{R}^n$, and z_i being its i-th entry. $\|\mathbf{z}\|_{L^q_{(\mu)}}$ is defined similarly.

To prove Lemma 18, we establish the following result first.

Lemma 20 *For $\ell = (\langle m, x\rangle - y)^2$, we have*

$$\ell^1 - \ell^2 = 2\left\langle m^1 - m^2, x\right\rangle \sqrt{\ell^a}.$$

Proof By definition,

$$\begin{aligned}\ell^1 - \ell^2 &= \left(\left\langle m^1, x\right\rangle - y\right)^2 - \left(\left\langle m^2, x\right\rangle - y\right)^2\\ &= \left\langle m^1, x\right\rangle^2 - 2y\left\langle m^1, x\right\rangle + y^2 - \left(\left\langle m^2, x\right\rangle^2 - 2y\left\langle m^2, x\right\rangle + y^2\right)\\ &= \left\langle m^1, x\right\rangle^2 - \left\langle m^2, x\right\rangle^2 + 2y\left\langle m^2 - m^1, x\right\rangle.\end{aligned}$$

Completing the square reveals:

$$\begin{aligned}\left\langle m^1, x\right\rangle^2 - \left\langle m^2, x\right\rangle^2 &= \left(\left\langle m^1, x\right\rangle - \left\langle m^2, x\right\rangle\right)\left(\left\langle m^1, x\right\rangle + \left\langle m^2, x\right\rangle\right)\\ &= \left\langle m^1 - m^2, x\right\rangle\left\langle m^1 + m^2, x\right\rangle,\end{aligned}$$

which leads to

$$\begin{aligned}&\ell^1 - \ell^2\\ =&\left\langle m^1 - m^2, x\right\rangle\left\langle m^1 + m^2, x\right\rangle + 2y\left\langle m^2 - m^1, x\right\rangle\\ =&2\left\langle m^1 - m^2, x\right\rangle\left(\left\langle\left(m^1 + m^2\right)/2, x\right\rangle - y\right)\\ =&2\left\langle m^1 - m^2, x\right\rangle\sqrt{\ell^a},\end{aligned}$$

where the last equality holds from the definition of ℓ^a. ■

Now we proceed to the proof of Lemma 18.

Proof of Lemma 18

$$\begin{aligned}\left\|\chi_A \circ \left(\ell^1 - \ell^2\right)\right\|_\mu^2 =&\mathbb{E}_{(x,y)\sim\mu}\chi_A(x,y)\left(\ell^1(x,y) - \ell^2(x,y)\right)^2\\ =&4\mathbb{E}_{(x,y)\sim\mu}\chi_A(x,y)\left\langle m^1 - m^2, x\right\rangle^2 \ell^a(x,y)\\ &\qquad\text{(by Lemma 20)}\\ =&4\mathbb{E}_{(x,y)\sim\mu}\chi_A(x,y)\left(Tr\left(\left(m^1 - m^2\right)\left(m^1 - m^2\right)^\top xx^\top\right)\ell^a(x,y)\right)\\ \le&4\mathbb{E}_{(x,y)\sim\mu}\left\|m^1 - m^2\right\|_2^2\chi_A(x,y)\left\|x\right\|_2^2\ell^a(x,y)\\ &\qquad\text{(Von Neumann Trace Ineq.)}\\ =&4\left\|m^1 - m^2\right\|_2^2\left\langle\chi_A \circ \xi^2, \chi_A \circ \ell^a\right\rangle_\mu\\ \le&4\left\|m^1 - m^2\right\|_2^2\left\|\chi_A \circ \xi^2\right\|_\mu\left\|\chi_A \circ \ell^a\right\|_\mu\\ &\qquad\text{(Cauchy Schwartz)}.\end{aligned}$$

■

Proof of Lemma 19 To characterize the difference between the m^i, we use the fact that they are the solutions that minimizes the weighted errors. The weighted errors in prediction

$$\begin{aligned}\langle \mathbf{w}, \ell^m \rangle_\mu =& \mathbb{E}_{(x,y)\sim\mu}\left(\lambda \langle m, m\rangle + \mathbf{w}(x,y)\cdot(\langle m, x\rangle - y)^2\right)\\ =& m^\top \underbrace{\mathbb{E}_{(x,y)\sim\mu}\mathbf{w}(x,y)\cdot\left(xx^\top + \lambda I\right)}_{H} m\\ &- 2\underbrace{\mathbb{E}_{(x,y)\sim\mu}\mathbf{w}(x,y)\cdot yx^\top}_{b} m + \underbrace{\mathbb{E}_{(x,y)\sim\mu}\mathbf{w}(x,y)\cdot y^2}_{c}\\ =& m^\top H m - 2b^\top m + c\end{aligned}$$

which is a convex quadratic. Notice that H are full rank, then the optimum is achieved at $m = H^{-1}b$. Now consider models m, m' induced by “similar” $\mathbf{w}, \mathbf{w}'$, i.e., $\|\mathbf{w} - \mathbf{w}'\|_\mu$ is small. Thus, m, m' are given by $H^{-1}b$ and $(H-D)^{-1}(b-d)$ respectively, where:

$$\begin{aligned}D &= \mathbb{E}_{(x,y)\sim\mu}\left(\mathbf{w} - \mathbf{w}'\right)(x,y)\cdot\left(xx^\top + \lambda I\right),\\ d &= \mathbb{E}_{(x,y)\sim\mu}\left(\mathbf{w} - \mathbf{w}'\right)(x,y)\cdot yx^\top.\end{aligned}$$

Then our goal is to bound $\left\|H^{-1}b - (H-D)^{-1}(b-d)\right\|_2^2$, in terms of $\left\|H^{-1}\right\|_{2,2}, \|D\|_{2,2}, \|d\|_2$ and $\|b-d\|_2$. First we bound these quantities:

$$\begin{aligned}\|d\|_2 =& \left\|\mathbb{E}_{(x,y)\sim\mu}\left(\mathbf{w} - \mathbf{w}'\right)(x,y)\cdot yx^\top\right\|_2\\ \le& \mathbb{E}_{(x,y)\sim\mu}\left\|\left(\mathbf{w} - \mathbf{w}'\right)(x,y)\cdot yx^\top\right\|_2 && \text{(Jensen's ineq.)}\\ =& \mathbb{E}_{(x,y)\sim\mu}\left\|\left(\mathbf{w} - \mathbf{w}'\right)(x,y)\cdot\chi_A(x,y)\cdot yx^\top\right\|_2\\ =& \mathbb{E}_{(x,y)\sim\mu}\left|\left(\mathbf{w} - \mathbf{w}'\right)(x,y)\right|\cdot\chi_A(x,y)\cdot y\|x\|_2\\ \le& \left\|\mathbf{w} - \mathbf{w}'\right\|_\mu \|\chi_A \circ \zeta\|_\mu && \text{(Cauchy Schwartz)}\end{aligned}$$

and similarly

$$\begin{aligned}\|D\|_{2,2} =& \left\|\mathbb{E}_{(x,y)\sim\mu}\left(\mathbf{w} - \mathbf{w}'\right)(x,y)\cdot\left(xx^\top + \lambda I\right)\right\|_{2,2}\\ \le& \mathbb{E}_{(x,y)\sim\mu}\left\|\left(\mathbf{w} - \mathbf{w}'\right)(x,y)\cdot\left(xx^\top + \lambda I\right)\right\|_{2,2} && \text{(Jensen's ineq.)}\\ =& \mathbb{E}_{(x,y)\sim\mu}\left\|\left(\mathbf{w} - \mathbf{w}'\right)(x,y)\cdot\chi_A(x,y)\cdot\left(xx^\top + \lambda I\right)\right\|_{2,2}\\ =& \mathbb{E}_{(x,y)\sim\mu}\left|\left(\mathbf{w} - \mathbf{w}'\right)(x,y)\right|\cdot\chi_A(x,y)\cdot\left\|xx^\top + \lambda I\right\|_{2,2}\\ =& \mathbb{E}_{(x,y)\sim\mu}\left|\left(\mathbf{w} - \mathbf{w}'\right)(x,y)\right|\cdot\chi_A(x,y)\cdot\left(\|x\|_2^2 + \lambda\right)\\ \le& \left\|\mathbf{w} - \mathbf{w}'\right\|_\mu \left\|\chi_A \circ \left(\xi^2 + \lambda\right)\right\|_\mu && \text{(Cauchy Schwartz)}\end{aligned}$$

and

$$\begin{aligned}
\|b-d\|_2 &= \left\|\mathbb{E}_{(x,y)\sim\mu}\mathbf{w}'(x,y)\cdot yx^\top\right\|_2 \\
&\le \mathbb{E}_{(x,y)\sim\mu}\left|\mathbf{w}'(x,y)\right|\cdot\chi_A(x,y)\cdot y\,\|x\|_2 && \text{(Jensen's ineq.)} \\
&\le \left\|\mathbf{w}'\right\|_{L^p(\mu)}\|\chi_A\circ\zeta\|_{L^q(\mu)}\,. && \text{(Hölder)}
\end{aligned}$$

Lastly, $\left\|H^{-1}\right\|_{2,2}\le\lambda^{-1}$.

Using just the triangle inequality and induced norms we find that

$$\begin{aligned}
\left\|H^{-1}b-(H-D)^{-1}(b-d)\right\|_2 &= \left\|H^{-1}b-H^{-1}(b-d)+H^{-1}(b-d)-(H-D)^{-1}(b-d)\right\|_2 \\
&\le \left\|H^{-1}b-H^{-1}(b-d)\right\|_2+\left\|H^{-1}(b-d)-(H-D)^{-1}(b-d)\right\|_2 \\
&= \left\|H^{-1}d\right\|_2+\left\|\left(H^{-1}-(H-D)^{-1}\right)(b-d)\right\|_2 \\
&\le \left\|H^{-1}\right\|_{2,2}\|d\|_2+\left\|H^{-1}-(H-D)^{-1}\right\|_{2,2}\|b-d\|_2 \\
&\le \lambda^{-1}\left\|\mathbf{w}-\mathbf{w}'\right\|_\mu\|\chi_A\circ\zeta\|_\mu \\
&\quad+\left\|H^{-1}-(H-D)^{-1}\right\|_{2,2}\left\|\mathbf{w}'\right\|_{L^p(\mu)}\|\chi_A\circ\zeta\|_{L^q(\mu)}\,.
\end{aligned}$$

To complete the proof, we use the assumption that $\theta_1=\|\mathbf{w}-\mathbf{w}'\|_\mu\left\|\chi_A\circ\left(\xi^2+\lambda\right)\right\|_\mu<\lambda/2$, and apply Lemma 21 below on the last term. Then writing $\theta_2=\|\mathbf{w}'\|_{L^p(\mu)}\|\chi_A\circ\zeta\|_{L^q(\mu)}$ the upper bound is:

$$\begin{aligned}
&\left\|H^{-1}b-(H-D)^{-1}(b-d)\right\|_2 \\
\le&\lambda^{-1}\left\|\mathbf{w}'-\mathbf{w}\right\|_\mu\left\|\chi_A\|yx\|_2\right\|_\mu+\frac{\lambda^{-1}\theta_1}{1-\lambda^{-1}\theta_1}\lambda^{-1}\theta_2 \\
\le&\lambda^{-1}\left\|\mathbf{w}'-\mathbf{w}\right\|_\mu\left\|\chi_A\|yx\|_2\right\|_\mu+2\lambda^{-1}\theta_1\lambda^{-1}\theta_2 \\
=&\lambda^{-1}\left\|\mathbf{w}'-\mathbf{w}\right\|_\mu\left(\|\chi_A\circ\zeta\|_\mu+2\left\|\chi_A\circ\left(\xi^2+\lambda\right)\right\|_\mu\lambda^{-1}\left\|\mathbf{w}'\right\|_{L^p(\mu)}\|\chi_A\circ\zeta\|_{L^q(\mu)}\right).
\end{aligned}$$

■

Lemma 21 *Let $H\succ\gamma I$, $\|D\|_{2,2}\le\varepsilon<\gamma$. Then*

$$\left\|H^{-1}-(H-D)^{-1}\right\|_{2,2}\le\frac{\gamma^{-2}\varepsilon}{1-\gamma^{-1}\varepsilon}$$

Proof We transform this expression using a Taylor expansion:

$$\begin{aligned}
&\left(H^{-1}-(H-D)^{-1}\right)\\
=&\left(H^{-1}H-(H-D)^{-1}H\right)H^{-1}\\
=&\left(I-\left(I-H^{-1}D\right)^{-1}\right)H^{-1}\\
=&\left(I-\sum_{i=0}^{\infty}\left(H^{-1}D\right)^{i}\right)H^{-1} && \text{(Taylor)}\\
=&-\left(\sum_{i=1}^{\infty}\left(H^{-1}D\right)^{i}\right)H^{-1},
\end{aligned}$$

which is applicable because $\left\|H^{-1}D\right\|_{2,2}\leq\left\|H^{-1}\right\|_{2,2}\|D\|_{2,2}\leq\gamma^{-1}\varepsilon<1$ implies that $I-H^{-1}D$ is invertible. Induced norms are sub-multiplicative, which leads to

$$\begin{aligned}
\left\|H^{-1}-(H-D)^{-1}\right\|_{2,2}=&\left\|-\left(\sum_{i=1}^{\infty}\left(H^{-1}D\right)^{i}\right)H^{-1}\right\|_{2,2}\\
\leq&\left\|\sum_{i=1}^{\infty}\left(H^{-1}D\right)^{i}\right\|_{2,2}\left\|H^{-1}\right\|_{2,2}\\
\leq&\sum_{i=1}^{\infty}\left\|\left(H^{-1}D\right)^{i}\right\|_{2,2}\left\|H^{-1}\right\|_{2,2} && \text{(Triangle ineq.)}\\
\leq&\left(\sum_{i=1}^{\infty}\gamma^{-1}\varepsilon\right)\gamma^{-1}\\
=&\frac{\gamma^{-2}\varepsilon}{1-\gamma^{-1}\varepsilon}.
\end{aligned}$$

■

Appendix C. Sample complexity theory

C.1. Bounded differences of $\mathcal{L}_\beta$

Bounded differences methods are a flexible toolset for proving that the distance of a random variable from its expected values is subgaussian. Here we prove that $\mathcal{L}_\beta(m;\mu)$ has bounded differences using two different approaches. These results will be used in Section C.2.

Definition 22 *A function $f: A^n\to\mathbb{R}$ has a bounded differences if for any x,x' that are identical except on one of their n entries, $|f(x)-f(x')|<a$ holds.*

Lemma 23 *Let the losses of m be bounded in an interval of length B, then $\mathcal{L}_\beta(m;\mu)$ has B bounded differences with regard to μ.*

Lemma 23 is straightforward, as $\mathcal{L}_\beta\left(m;\cdot\right)$ is in the interval because $\left\|\mathbf{w}-\mathbf{1}_n\right\|_\mu \geq 0$ and $\mathbf{1}_n$ is an allowed choice for the weight vector $\mathbf{w}$. The next result on bounded difference is more involved.

Proposition 24 *Let $n > 9$, let m satisfy the assumptions of Proposition 1 with $p \geq 2/3$ such that the smallest $pn+1$ entries belong to $[\rho_l, \rho_h]$. Then $\mathcal{L}_\beta\left(m;\mu\right)$ has $24\beta/n$ bounded differences with regard to μ.*

Proof Let ℓ' differ from ℓ by a single entry (the first, without loss of generality). Note that due to the slightly strengthened assumption here (i.e., $pn+1$ entries, as opposed to pn entries, of ℓ belong to $[\rho_l, \rho_h]$), ℓ' also satisfies the assumptions of Proposition 1. The key step is to show that the modified $\mathbf{w}'$ is very close to $\mathbf{w}$, by showing that $|a-a'| \leq 3n^{-1}$.

Step 1: to show $|a-a'| \leq 3n^{-1}$. Denote for convenience $\lambda = -\ell\left(2\beta\right)^{-1}$, and similarly λ' for ℓ', then $\mathbf{w} = \left[\lambda + a\right]_+$ and $\mathbf{w}' = \left[\lambda' + a\right]_+$. If $\mathbf{w}_1 = \mathbf{w}'_1$ then $a = a'$ (otherwise if $a < a'$ then $\mathbf{w}_i < \mathbf{w}'_i$ for all $i \neq 1$ since $\ell_i = \ell'_i$ which is a contradiction, and so is the case of $a > a'$) and Step 1 is true. Otherwise, denote $z = \left[\lambda'_1 + a'\right]_+ - \left[\lambda_1 + a\right]_+$, then by the constraint that $\langle \mathbf{w}, 1\rangle = 1$, we know also that $z = \sum_{i\neq 1}\left(\left[\lambda_i + a\right]_+ - \left[\lambda'_i + a'\right]_+\right)$. Since $\lambda'_i = \lambda_i$ for $i \neq 1$, we can write $z = \sum_{i\neq 1}\left(\left[\lambda_i + a\right]_+ - \left[\lambda_i + a'\right]_+\right)$ in which it is clear that any non zero summand $\left[\lambda_i + a\right]_+ - \left[\lambda_i + a'\right]_+$ must have the same sign as $a - a'$ and therefore z.

Now recall that by Proposition 1, for at least 2/3 fractions of indices $i \neq 1$, both $\mathbf{w}_i > 0$ and $\mathbf{w}'_i > 0$, denote this subset of entries I. Then by the argument of the previous paragraph, $|z| = \sum_{i\neq 1}\left(\left|\left[\lambda_i + a\right]_+ - \left[\lambda_i + a'\right]_+\right|\right) \geq \sum_{i\in I}\left(\left|\left[\lambda_i + a\right]_+ - \left[\lambda_i + a'\right]_+\right|\right) = \sum_{i\in I}\left(|a-a'|\right) \geq \left(|a-a'|\right)2n/3$, and hence $3\left|z\right|n^{-1}/2 \geq |a-a'|$.

Since z is a difference between weights we bound $|z|$ using Proposition 1: for our $p, c, \mathbf{w}_i, \mathbf{w}'_i \in \left[0, p^{-1} + \left(2c\right)^{-1}\right] \subset \left[0, 3/2 + 2^{-1}\right] = [0,2]$ and then $z \in [0,2]$ (we will reuse these facts). Then $3n^{-1} \geq |a-a'| = |\mathbf{w}_i - \mathbf{w}'_i|$ for $i \neq 1$.

Step 2: complete the proof of the proposition.

$$
\begin{aligned}
& n\left(\langle \mathbf{w}, \ell\rangle + \beta\left\|\mathbf{w}-1\right\|^2 - \left(\langle \mathbf{w}', \ell'\rangle + \beta\left\|\mathbf{w}'-1\right\|^2\right)\right) \\
=& \sum_{i=1}^{n}\left(\mathbf{w}_i\ell_i - \mathbf{w}'_i\ell'_i\right) + \beta\left(\sum_{i=1}^{n}\left(\left(\mathbf{w}_i - 1\right)^2 - \left(\mathbf{w}'_i - 1\right)^2\right)\right) \\
=& \underbrace{\mathbf{w}_1\ell_1 - \mathbf{w}'_1\ell'_1}_{a} + \sum_{i=2}^{n}\underbrace{\left(\mathbf{w}_i - \mathbf{w}'_i\right)\ell_i}_{b} \\
& + \beta\left(\underbrace{\left(\mathbf{w}_1 - 1\right)^2 - \left(\mathbf{w}'_1 - 1\right)^2}_{c} + \sum_{i=2}^{n}\underbrace{\left(\mathbf{w}_i - 1\right)^2 - \left(\mathbf{w}'_i - 1\right)^2}_{d}\right).
\end{aligned}
$$

To bound the loss terms a, b, we first recall that by Proposition 1, for any $\ell_i \geq \rho_h + 3\beta$, the weights $\mathbf{w}_i, \mathbf{w}'_i$ (hence also the products $\mathbf{w}_i\ell_i$) are exactly 0. Denote $J = \{i \neq 1 : \ell_i \leq \rho_h + 3\beta\}$, then $z = \sum_{i\in J}\left(\mathbf{w}_i - \mathbf{w}'_i\right)$. We trivially rewrite the contribution of terms b as $\sum_{i\neq 1}\left(\mathbf{w}_i - \mathbf{w}'_i\right)\ell_i = \sum_{i\in J}\left(\mathbf{w}_i - \mathbf{w}'_i\right)\ell_i = z\sum_{i\in J}\left(\left(\mathbf{w}_i - \mathbf{w}'_i\right)/z\right)\ell_i$. In this form it is clear that $\ell^* \triangleq \sum_{i\in J}\left(\left(\mathbf{w}_i - \mathbf{w}'_i\right)/z\right)\ell_i$ is a convex combination of elements from $[\rho_l, \rho_h + 3\beta]$, therefore also belong to this interval. Term b is succinctly written $z\ell^*$.

If both of $\mathbf{w}_1, \mathbf{w}'_1$ are zero, then $\mathbf{w} = \mathbf{w}'$ and the total difference is 0. We may then assume without loss of generality that $\mathbf{w}_1 \geq \mathbf{w}'_1$, and conclude that $\mathbf{w}_1 > 0$ and $\ell_1 < \rho_h + 3\beta$. We rewrite term a: $\mathbf{w}_1 \ell_1 - \mathbf{w}'_1 \ell'_1 = (\mathbf{w}'_1 - z) \ell_1 - \mathbf{w}'_1 \ell'_1 = \mathbf{w}'_1 (\ell_1 - \ell'_1) - z\ell_1$, and thus $a + b$ can be written as $z(\ell^* - \ell_1) + \mathbf{w}'_1 (\ell_1 - \ell'_1)$ for convenient bounding.

Now recall from part 1 that $|z|, \mathbf{w}'_i \leq 2$. Then $|z(\ell^* - \ell_1)| \leq 2(\rho_h - \rho_l + 3\beta)$. If $\mathbf{w}'_1 = 0$, then clearly $\mathbf{w}'_1 (\ell_1 - \ell'_1) = 0$. Otherwise, $\ell'_1 \leq \rho_h + 3\beta$, then $\mathbf{w}'_1 (\ell_1 - \ell'_1) \leq 2(\rho_h - \rho_l + 3\beta)$ also. Then the total contribution of a, b terms is bounded by $4(\rho_h - \rho_l + 3\beta) \leq 16\beta$.

Since $\mathbf{w}_1 \in [0, 2]$, $(\mathbf{w}_1 - 1)^2 \in [0, 1]$, and the same holds for $\mathbf{w}'_1$, then term c is upper bounded by 1. To bound term d, we use the following equality

$$(x-1)^2 - (x+h-1)^2 = h(2 - 2x - h)$$

and with the triangle inequality we find

$$\left|(\mathbf{w}_i - 1)^2 - (\mathbf{w}'_i - 1)^2\right| \leq \left|\mathbf{w}'_i - \mathbf{w}_i\right| \left(2|1 - \mathbf{w}_i| + \left|\mathbf{w}'_i - \mathbf{w}_i\right|\right) \leq 3n^{-1}\left(2 + 3n^{-1}\right),$$

where the second inequality holds because $i \neq 1$, and hence we can apply the tighter bound $|\mathbf{w}_i - \mathbf{w}'_i| \leq 3n^{-1}$ and note that $|\mathbf{w}_i - 1| \leq 1$.

Sum up all terms we have

$$\begin{aligned} n\left|\langle \mathbf{w}, \ell \rangle + \beta \|\mathbf{w} - 1\|^2 - \left(\langle \mathbf{w}', \ell' \rangle + \beta \|\mathbf{w}' - 1\|^2\right)\right| &\leq \left(16 + 1 + 3\left(2 + 3n^{-1}\right)\right)\beta \\ &= \left(23 + 9n^{-1}\right)\beta \\ &\leq 24\beta \end{aligned}$$

where the last inequality holds by $n > 9$. ■

C.2. Proof of sample complexity

The sample complexity proof follows a well known path of proving concentration for a single model, then applying a union bound over a discretization of the space of models. Hence the first part is to prove that for a fixed model m, the random variable $\mathcal{L}_\beta(m; \mu)$ (with regard to the sampling of μ) is highly concentrated around its expected value $\mathbb{E}_{\mu'} \mathcal{L}_\beta(m; \mu')$.

Recall that the losses of every model are bounded above such that $\max_{x \in \mathcal{X}} \ell(m, x) - \rho_l^m \leq B$.

Lemma 25 *Let μ, μ' be empirical distributions of n independent elements drawn according to ν. Fix $\beta, s > 0$, let $m \in \mathcal{M}$, and denote $p = P_{x \sim \nu}(\ell(m, x) < \rho_l^m + \beta)$, and assume that $p \geq 2/3 + n^{-1} + s$. If $\varepsilon \in (0, 48\sqrt{2}\beta s]$ and $n > \max\left\{2s^{-2}\log\left(B(24\beta)^{-1}\right), 3h\left(4s^{-2} + 3\right)\right\}$, then*

$$P\left(\left|\mathcal{L}_\beta(m; \mu) - \mathbb{E}_{\mu'}\mathcal{L}_\beta(m; \mu')\right| \geq \varepsilon\right) \leq 4\exp\left(-\frac{n\varepsilon^2}{8(24\beta)^2}\right).$$

The core of our argument is to apply a concentration inequality, Theorem 27 in Appendix C.3. Unlike the standard McDiarmid inequality, this uses methods developed in Kutin (2002) that apply even when a function has bounded differences that decrease with n only with high probability (w.h.p.) on an independent sample. More formally, we show that $\mathcal{L}_\beta$ is *strongly difference bounded*:

Definition 26 *A random variable $Z = f(X_1, \ldots, X_n)$ is* strongly difference bounded *by (b, b', δ) if f is b difference bounded and in addition the following holds: there is a "bad" subset $D \subset \Omega$, where $\delta = Pr(X \in D)$, and conditional on D not occurring, $f(X)$ is b' difference bounded.*

Then combining Lemma 23 and Proposition 24, $\mathcal{L}_\beta(m; \mu)$ is strongly difference bounded by $(B, 24\beta/n, \delta)$ with regard to μ, for the right β, n. We formalize the details below.

Proof We obtain concentration by Theorem 27, whose auxiliary requirements on the values of ε, n are assumed by this lemma. The main requirement is that $\mathcal{L}_\beta(m; \mu)$ is strongly difference bounded. According to Lemma $\mathcal{L}_\beta(m; \mu)$ is B difference bounded, then it suffices to show that $\mathcal{L}_\beta(m; \mu)$. has differences bounded by $24\beta/n$, except on an "bad event" that has probability at most $\exp(-Kn)$, and find the value of K.

Proposition 24 shows differences are bounded as we need if $n > 9$, $p' > 2/3$ and the $p'n + 1$ bottom losses are in an interval whose length is less than β. The assumption in this lemma on n is never smaller than 9.5, hence suffices for the proposition. Given the assumptions of this lemma on p, for Proposition 24 to hold, it is enough to show that $p' > p - s$ with probability $1 - \exp(-Kn)$.

The proportion of losses of a sample falling in a range is a binomial variable, and by the simple tail bound for binomials,

$$P\left(|\{i : \ell_i \in [\rho_l, \rho_l + \beta]\}| \, n^{-1} < p - s\right) \le \exp\left(-\frac{2(np - n(p - s))^2}{n}\right) = \exp\left(-2s^2 n\right)$$

then $K = 2s^2$ suffices. ■

After showing exponential concentration for a single model, we are ready to prove Theorem 11. The standard discretization argument is modified to account for the probability of the bad event occuring on any of the models.

Proof We denote $\mathcal{M}^\varepsilon$ a minimal ε cover of $\mathcal{M}$; by assumption the cardinality of $\mathcal{M}^\varepsilon$ is at most $(CB/\varepsilon)^d$. For a given model $m \in \mathcal{M}$, we denote $m^c \in \mathcal{M}^\varepsilon$ the closest element there, then $d_{\mathcal{X}}(m, m^c) \le \varepsilon$. By the triangle inequality,

$$\begin{aligned} \left|\mathcal{L}_\beta(m; \mu) - \mathbb{E}_{\mu'}\mathcal{L}_\beta\left(m; \mu'\right)\right| \le & \left|\mathcal{L}_\beta(m; \mu) - \mathcal{L}_\beta(m^c; \mu)\right| + \left|\mathcal{L}_\beta(m^c; \mu) - \mathbb{E}_{\mu'}\mathcal{L}_\beta\left(m^c; \mu'\right)\right| \\ & + \left|\mathbb{E}_{\mu'}\mathcal{L}_\beta\left(m^c; \mu'\right) - \mathbb{E}_{\mu'}\mathcal{L}_\beta\left(m; \mu'\right)\right|. \end{aligned}$$

We bound the first and third terms by ε deterministically. Let $\mathbf{w}, \ell$ correspond to m, and ℓ^c to m^c, then since $|\ell(m, \cdot) - \ell(m^c, \cdot)| \le \varepsilon$ we see that $\mathcal{L}_\beta(m^c; \mu) - \mathcal{L}_\beta(m; \mu) \le \langle \ell^c, \mathbf{w} \rangle_\mu + \|\mathbf{w} - \mathbf{1}_n\|_\mu^2 - \mathcal{L}_\beta(m; \mu) = \langle \ell^c - \ell, \mathbf{w} \rangle_\mu \le \varepsilon$ by the Hölder inequality. By symmetry we conclude $|\mathcal{L}_\beta(m; \mu) - \mathcal{L}_\beta(m^c; \mu)| \le \varepsilon$. This bound on the first term holds for any μ, so the bound on the third term follows by taking expectations.

It remains to bound the second term with high probability, for the following subset of the cover:

$$\mathcal{M}_{\textbf{discrete}} = \left\{m^c : m \in \mathcal{M}_{p+s}^{\nu, \beta - \varepsilon}\right\}.$$

For $m \in \mathcal{M}_{\textbf{discrete}}$, because $d_{\mathcal{X}}$ is the supremum norm, $P_{x \sim \nu}\left(\ell(m, x) - \rho_l^m < \beta\right) > p + s$.

The assumed p fulfills the relevant condition of Lemma 25 with β, and our choice of s gives $\varepsilon = 16\beta s \in \left(0, 48\sqrt{2}\beta s\right]$. Then for $n > \max\left\{2s^{-2}\log\left(B(24\beta)^{-1}\right), 3h\left(4s^{-2} + 3\right)\right\}$ we have that for a particular $m \in \mathcal{M}_{\mathbf{discrete}}$,

$$P\left(\left|\mathcal{L}_\beta\left(m;\mu\right)-\mathbb{E}_{\mu'}\mathcal{L}_\beta\left(m;\mu'\right)\right|>\varepsilon\right)\leq 4\exp\left(-\frac{n\varepsilon^2}{8\left(24\beta\right)^2}\right).$$

By a union bound, the maximal deviation over all $m \in \mathcal{M}_{\mathbf{discrete}}$ is under ε, except with probability at most:

$$\left(CB/\varepsilon\right)^d\left(4\exp\left(-\frac{n\varepsilon^2}{8\left(24\beta\right)^2}\right)\right) \quad = \quad \exp\left(\log 4+\log\left(CB/\varepsilon\right)^d-\frac{n\varepsilon^2}{8\left(24\beta\right)^2}\right).$$

■

C.3. Concentration for strongly difference bounded RVs with large range

The following theorem is proved in Kutin (2002) with

$$M\left(b,\lambda,K\right)=M_{kutin}\left(b,\lambda,K\right)=\max\left\{b\lambda^{-1},3h\left(6/K+3\right)\right\}.$$

Theorem 27 *(Theorem 1.9 from Kutin (2002)) Let $Z=f\left(X\right)$ be strongly difference bounded by $\left(b,\lambda/n,\exp\left(-Kn\right)\right)$. Then for any $\varepsilon\in\left(0,2\lambda\sqrt{K}\right]$ and*

$$n>M\left(b,\lambda,K\right)$$

then

$$\Pr\left(\left|Z-\mathbb{E}Z\right|\geq\varepsilon\right)\leq 4\exp\left(-\frac{n}{8}\left(\frac{\varepsilon}{\lambda}\right)^2\right).$$

Indeed, this theorem also holds with

$$M_{new}\left(b,\lambda,K\right)=\max\left\{4K^{-1}\log\left(b\lambda^{-1}\right),3h\left(8/K+3\right)\right\},$$

which has a much improved dependence on b. To establish this we recall two results from (Kutin, 2002).

Theorem 28 *(Theorem 3.3 from Kutin (2002)) Let $Z=f\left(X\right)$ be strongly difference bounded by $\left(b,c,\delta\right)$. Then for any $\varepsilon>0$ and any $\alpha>0$*

$$\Pr\left(\left|Z-\mathbb{E}Z\right|\geq\varepsilon\right)\leq 2\left(\exp\left(-\frac{\varepsilon^2}{2n\left(c+b\alpha\right)^2}\right)+\frac{n}{\alpha}\delta\right).$$

Lemma 29 *(Lemma 3.7 from Kutin (2002)) For any $z>0$, if $n>3h\left(z+3\right)$ then $n\ln^{-1}n>z$.*

We are now ready to prove Theorem 27 with the M_{new}.
Proof Substituting $(b,\lambda/n,\exp(-Kn))$ into Theorem 28, we have

$$\Pr\left(\left|Z-\mathbb{E}Z\right|\geq\varepsilon\right)\leq 2\left(\exp\left(-\frac{\varepsilon^2}{2n\left(\lambda/n+b\alpha\right)^2}\right)+\frac{n}{\alpha}\exp\left(-Kn\right)\right).$$

Now choose $\alpha = \lambda n^{-1} b^{-1}$ and simplify:

$$\begin{aligned}\Pr\left(|Z - \mathbb{E}Z| \geq \varepsilon\right) \leq & 2\left(\exp\left(-\frac{n\varepsilon^2}{8\lambda^2}\right) + \frac{bn^2}{\lambda}\exp\left(-Kn\right)\right) \\ = & 2\left(\exp\left(-\frac{n\varepsilon^2}{8\lambda^2}\right) + \exp\left(\log\left(\frac{bn^2}{\lambda}\right) - Kn\right)\right)\end{aligned}$$

For the second exponent to be dominated by the first, it suffices to have

$$\begin{aligned} -\frac{n\varepsilon^2}{8\lambda^2} > & \log\left(\frac{bn^2}{\lambda}\right) - Kn \\ \iff \qquad Kn > & \frac{n\varepsilon^2}{8\lambda^2} + \log\left(b\lambda^{-1}\right) + 2\log\left(n\right)\end{aligned}$$

which holds if the following holds:

$$Kn > \max\left(\frac{n}{4}\left(\frac{\varepsilon}{\lambda}\right)^2, 4\log\left(b\lambda^{-1}\right), 8\log\left(n\right)\right).$$

To assure the first, it suffice to have $8^{-1}\left(\frac{\varepsilon}{\lambda}\right)^2 < K/2 \iff \varepsilon < 2\lambda\sqrt{K}$. The second holds due to $n > 4K^{-1}\log\left(b\lambda^{-1}\right)$. For the last we use Lemma 29: if $n \geq 3h\left(z+3\right)$ then $\frac{n}{\ln n} > z \iff nz^{-1} > \ln n$, and take $z = \frac{8}{K}$. ■

Appendix D. Empirical illustration

We compare our approach to an existing robust method and a baseline on three common problems: location estimation, linear regression and dimensionality reduction. We briefly explain the general simulation setup. For each task, we generate most data from a natural source, and some smaller proportion of outliers from a different source. We vary the proportion of outliers and the magnitude of the perturbation in the outlier source. The three algorithms (ours, l_1 adjustments and plain least squares, explained in detail below) are each applied to a dataset for each pair of parameters for each problem type. We define the estimation error as the logarithm of the norm of some distance between the true majority source and the learned model. In other words, the task is to estimate the majority model, not model all data, and we compare error levels in decibel units. The results are plotted in Figures 4, 5 and 6, with darker shadow corresponding to higher errors. As a baseline for a problem, we take the estimation error scale using the non-robust estimators *in the absence of outliers*. Subtracting the baseline estimation error from the estimation error for a method indicates the additional error due to outliers. We average this additional error across data generation parameters to summarize algorithm performance in a single number for each problem.

The specific setups for each task are as follows. In the location estimation task, the data are generated from the mixture of two Gaussian distributions. The less likely one is treated as a noise source, and our goal is to identify the mean of the more likely one. In the linear regression task the labels of most data are generated by a linear model with additive noise, and the remainder are generated from a perturbed linear model. The norm of the perturbation can be much larger than that of the ground truth model. In the dimensionality reduction task, most points are from a 4 dimensional subspace of $\mathbb{R}^{10}$, and the remainder are from a 4 dimensional subspace sharing 3 basis elements with the first.

We compare three algorithms. The first one is the proposed RW algorithm (i.e., Algorithm 1). Here, we choose β by first solving the uniformly weighted problem, and then take β equal to the average of empirical losses. The second one – the alternative robust estimation method – is the approach proposed in Giannakis et al. (2011); Mateos and Giannakis (2012a,b) which also uses alternating minimization. In particular, each data point is modeled as the sum of an estimated value and an adjustment; the sum of the norms of the adjustments is taken as a penalty. At each iteration, the model is fit to the estimated data and the adjustment for each data point is updated so as to minimize the sum of the model loss and the l_1 penalty on the adjustments. Due to the sparsity encouraging property of l_1 regularization, data points that fit the model well are not affected by the adjustment at all, while adjustments for outliers can be large. In simple cases, this scheme corresponds to replacing the squared distance loss with the Huber loss. This approach is generic conceptually, although applying it to each individual problem type requires some customization. The third algorithm, as a baseline, minimizes the average of quadratic losses.

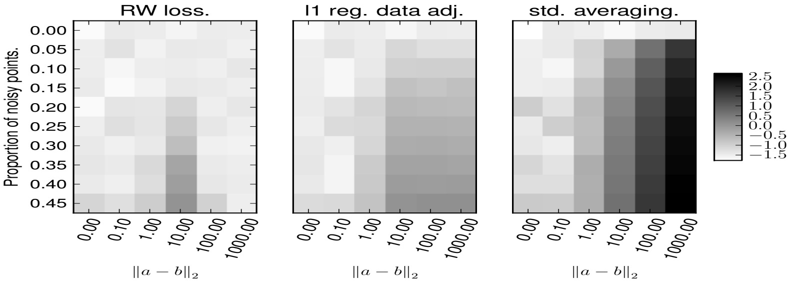

Figure 4: *Location estimation.* The majority source follows $\mathcal{N}(a, d^{-1}I)$, while the noisy points follow $\mathcal{N}(b, 100d^{-1}I)$. The goal is an estimate $\hat{a}$ of a. Color corresponds to error measured as $\log \|\hat{a}-a\|_2$; the average additional errors due to outliers are (0.19, 0.55, 1.47) for RM, the robust competitor and the mean, respectively.

In each of the problem types, the proposed method achieves smallest additional error due to outliers. Large perturbations cause severe error for standard linear regression, smaller error for the l_1 regularized correction method, and essentially no additional error for regularized weighting.

Finally, we comment on the implementation issue about tuning the β parameter. Our theorems identify a tradeoff between robustness and generalization by tuning β, which should be set according to the losses for non-outliers. In practice as a rule of thumb we suggest choosing β by first solving the uniformly weighted problem, and then take β equal to the average of empirical losses, which is also how we implement the algorithm in the simulation section.

References

Andreas Alfons, Christophe Croux, and Sarah Gelper. Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics*, 7(1):226–248, 2013.

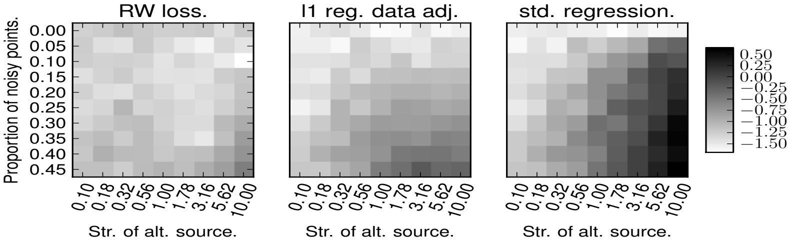

Figure 5: *Linear regression.* The data are generated as $x \sim \mathcal{N}(0, d^{-1/2}I)$ and $y \sim \langle m, x\rangle + s\langle m', x\rangle + \mathcal{N}(0, (10)^{-1/2})$ where $s = 0$ except with probability varying with the vertical axis. The value of s when non-zero varies with the horizontal axis. Color corresponds to error measured as $\log \|\hat{m} - m\|_2$; the average additional errors due to outliers are (0.0089, 0.15, 0.52) decibels respectively for RM, the robust competitor and least squares regression.

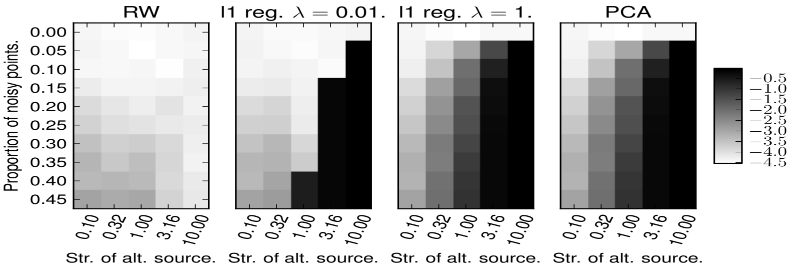

Figure 6: *Principal component analysis.* Each data point is generated according to $x = Az + sBz_1$, where $z \in \mathbb{R}^4$ follow $z \sim \mathcal{N}(0, I)$, $A \in \mathbb{R}^{10\times4}$ is orthonormal and $B \in \mathbb{R}^{10\times1}$ is perpendicular to the columns of A. The probability that $s \neq 0$ and the value of s are the vertical axis and the horizontal axis respectively. Color corresponds to the logarithm of the squared Euclidean norm of the sines of the 4 principal angles between the column space of A and the estimated 4 dimensional subspace. The average additional errors due to outliers are (0.41, 2.19, 2.93, 2.93) respectively for RM, the robust competitor at two parameter values, and standard PCA.

Martin Anthony and Peter L. Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 1999.

Ozlem Aslan, Dale Schuurmans, and Yao-liang Yu. A polynomial-time form of robust regression. In *Advances in Neural Information Processing Systems*, pages 2483–2491, 2012.

Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *The Annals of Statistics*, pages 2766–2794, 2011.

George EP Box. Non-normality and tests on variances. *Biometrika*, 40(3/4):318–335, 1953.

Matthew Coudron and Gilad Lerman. On the sample complexity of robust PCA. In *Advances in Neural Information Processing Systems*, pages 3221–3229, 2012.

John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l_1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 272–279. ACM, 2008.

Georgios B. Giannakis, Gonzalo Mateos, Shahrokh Farahmand, Vassilis Kekatos, and Hao Zhu. Uspacor: Universal sparsity-controlling outlier rejection. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 1952–1955. IEEE, 2011.

Frank Rudolf Hampel. *Contributions to the theory of robust estimation*. PhD thesis, University of California Berkeley, 1968.

Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, 17(18):1–40, 2016.

Peter J Huber et al. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

Samuel Kutin. Extensions to McDiarmid's inequality when differences are bounded with high probability. Technical Report TR-2002-04, Department of Computer Science, The University of Chicago, 2002.

Po-Ling Loh. Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *arXiv preprint arXiv:1501.00312*, 2015.

Po-Ling Loh and Martin J Wainwright. Support recovery without incoherence: A case for nonconvex regularization. *arXiv preprint arXiv:1412.5632*, 2014.

Po-Ling Loh and Martin J Wainwright. Regularized m-estimators with nonconvexity: statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research*, 16(1):559–616, 2015.

Gonzalo Mateos and Georgios B. Giannakis. Robust nonparametric regression via sparsity control with application to load curve data cleansing. *Signal Processing, IEEE Transactions on*, 60(4): 1571–1584, 2012a.

Gonzalo Mateos and Georgios B. Giannakis. Robust PCA as bilinear decomposition with outlier-sparsity regularization. *Signal Processing, IEEE Transactions on*, 60(10):5176–5190, 2012b.

Shahar Mendelson. Learning without concentration. In *Proceedings of The 27th Conference on Learning Theory*, pages 25–39, 2014.

TD Nguyen and R Welsch. Outlier detection and least trimmed squares approximation using semi-definite programming. *Computational Statistics & Data Analysis*, 54(12):3212–3226, 2010.

Peter J Rousseeuw. Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880, 1984.

Peter J Rousseeuw and Katrien Van Driessen. An algorithm for positive-breakdown regression based on concentration steps. In *Data Analysis*, pages 335–346. Springer, 2000.

J. Tukey. The future of data analysis. *Annals of Mathematical Statistics*, 33:1–67, 1962.

John W Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, 2:448–485, 1960.

Daniel Vainsencher, Shie Mannor, and Huan Xu. Learning multiple models via regularized weighting. In *Advances in Neural Information Processing Systems 26*, pages 1977–1985, 2013.

Aad W. Van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Verlag, 1996.

Huan Xu, Constantine Caramanis, and Shie Mannor. Outlier-robust PCA: The high-dimensional case. *IEEE transactions on information theory*, 59(1):546–572, 2013.