A Polynomial-time Form of Robust Regression

Yaoliang Yu, Özlem Aslan and Dale Schuurmans
Department of Computing Science, University of Alberta, Edmonton AB T6G 2E8, Canada
{yaoliang,ozlem,dale}@cs.ualberta.ca

Abstract

Despite the variety of robust regression methods that have been developed, current regression formulations are either NP-hard, or allow unbounded response to even a single leverage point. We present a general formulation for robust regression—Variational M-estimation—that unifies a number of robust regression methods while allowing a tractable approximation strategy. We develop an estimator that requires only polynomial-time, while achieving certain robustness and consistency guarantees. An experimental evaluation demonstrates the effectiveness of the new estimation approach compared to standard methods.

1 Introduction

It is well known that outliers have a detrimental effect on standard regression estimators. Even a single erroneous observation can arbitrarily affect the estimates produced by methods such as least squares. Unfortunately, outliers are prevalent in modern data analysis, as large data sets are automatically gathered without the benefit of manual oversight. Thus the need for regression estimators that are both scalable and robust is increasing.

Although the field of robust regression is well established, it has not considered computational complexity analysis to be one of its central concerns. Consequently, none of the standard regression estimators in the literature are both robust and tractable, even in a weak sense: it has been shown that standard robust regression formulations with non-zero breakdown are NP-hard [1, 2], while any estimator based on minimizing a convex loss cannot guarantee bounded response to even a single leverage point [3] (definitions given below). Surprisingly, there remain no standard regression formulations that guarantee both polynomial run-time with bounded response to even single outliers.

It is important to note that robustness and tractability can be achieved under restricted conditions. For example, if the domain is *bounded*, then any estimator based on minimizing a convex and Lipschitz-continuous loss achieves high breakdown [4]. Such results have been extended to kernel-based regression under the analogous assumption of a bounded kernel [5, 6]. Unfortunately, these results can no longer hold when the domain or kernel is *unbounded*: in such a case arbitrary leverage can occur [4, 7] and no (non-constant) convex loss, even Lipschitz-continuous, can ensure robustness against even a single outlier [3]. Our main motivation therefore is to extend these existing results to the case of an unbounded domain. Unfortunately, the inapplicability of convex losses in this situation means that computational tractability becomes a major challenge, and new computational strategies are required to achieve tractable robust estimators.

The main contribution of this paper is to develop a new robust regression strategy that can guarantee both polynomial run-time and bounded response to individual outliers, including leverage points. Although such an achievement is modest, it is based on two developments of interest. The first is a general formulation of adaptive M-estimation, *Variational M-estimation*, that unifies a number of robust regression formulations, including convex and bounded M-estimators with certain subset-selection estimators such as Least Trimmed Loss [7]. By incorporating Tikhonov regularization, these estimators can be extended to reproducing kernel Hilbert spaces (RKHSs). The second development is a convex relaxation scheme that ensures bounded outlier influence on the final estimator.

The overall estimation procedure is guaranteed to be tractable, robust to single outliers with unbounded leverage, and consistent under non-trivial conditions. An experimental evaluation of the proposed estimator demonstrates effective performance compared to standard robust estimators.

The closest previous works are [8], which formulated variational representations of certain robust losses, and [9], which formulated a convex relaxation of bounded loss minimization. Unfortunately, [8] did not offer a general characterization, while [9] did not prove their final estimator was robust, nor was any form of consistency established. The formulation we present in this paper generalizes [8] while the convex relaxation scheme we propose is simpler and tighter than [9]; we are thus able to establish non-trivial forms of both robustness and consistency while maintaining tractability.

There are many other notions of "robust" estimation in the machine learning literature that do not correspond to the specific notion being addressed in this paper. Work on "robust optimization" [10–12], for example, considers minimizing the worst case loss achieved given bounds on the maximum data deviation that will be considered. Such results are not relevant to the present investigation because we explicitly do not bound the magnitude of the outliers. Another notion of robustness is algorithmic stability under leave-one-out perturbation [13], which analyzes specific learning procedures rather than describing how a stable algorithm might be generally achieved.

2 Preliminaries

We start by considering the standard linear regression model

$$y \quad = \quad \mathbf{x}^T \boldsymbol{\theta}^* + u \tag{1}$$

where $\mathbf{x}$ is an $\mathbb{R}^p$-valued random variable, u is a real-valued random noise term, and $\boldsymbol{\theta}^* \in \Theta \subseteq \mathbb{R}^p$ is an unknown deterministic parameter vector. Assume we are given a sample of n independent identically distributed (i.i.d.) observations represented by an $n \times p$ matrix X and an $n \times 1$ vector $\mathbf{y}$, where each row $X_{i:}$ is drawn from some unknown marginal probability measure $P_{\mathbf{x}}$, and y_i are generated according to (1). Our task is to estimate the unknown deterministic parameter $\boldsymbol{\theta}^* \in \Theta$. Clearly, this is a well-studied problem in statistics and machine learning. If the noise distribution has a known density $p(\cdot)$, then a standard estimator is given by maximum likelihood

$$\hat{\boldsymbol{\theta}}_{ML} \quad \in \quad \arg\min_{\boldsymbol{\theta} \in \Theta} \tfrac{1}{n} \textstyle\sum_{i=1}^n - \log p(y_i - X_{i:}\boldsymbol{\theta}) \quad = \quad \arg\min_{\boldsymbol{\theta} \in \Theta} \tfrac{1}{n} \textstyle\sum_{i=1}^n - \log p(r_i), \tag{2}$$

where $r_i = y_i - X_{i:}\boldsymbol{\theta}$ is the ith residual. When the noise distribution is unknown, one can replace the negative log-likelihood with a *loss function* $\rho(\cdot)$ and use the estimator

$$\hat{\boldsymbol{\theta}}_{M} \quad \in \quad \arg\min_{\boldsymbol{\theta} \in \Theta} \tfrac{1}{n} \mathbf{1}^T \boldsymbol{\rho}(\mathbf{y} - X\boldsymbol{\theta}), \tag{3}$$

where $\boldsymbol{\rho}(\mathbf{r})$ denotes the vector of losses obtained by applying the loss componentwise to each residual, hence $\mathbf{1}^T \boldsymbol{\rho}(\mathbf{r}) = \sum_{i=1}^n \rho(r_i)$. Such a procedure is known as M-estimation in the robust statistics literature, and empirical risk minimization in the machine learning literature.[1]

Although uncommon in robust regression, it is conventional in machine learning to include a regularizer. In particular we will use Tikhonov ("ridge") regularization by adding a squared penalty

$$\hat{\boldsymbol{\theta}}_{MR} \quad \in \quad \arg\min_{\boldsymbol{\theta} \in \Theta} \tfrac{1}{n} \mathbf{1}^T \boldsymbol{\rho}(\mathbf{y} - X\boldsymbol{\theta}) + \tfrac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2 \qquad \text{for } \lambda \geq 0, \tag{4}$$

The significance of Tikhonov regularization is that it ensures $\hat{\boldsymbol{\theta}}_{MR} = X^T \boldsymbol{\alpha}$ for some $\boldsymbol{\alpha} \in \mathbb{R}^n$ [14]. More generally, under Tikhonov regularization, the regression problem can be conveniently expressed in a reproducing kernel Hilbert space (RKHS). If we let $\mathcal{H}$ denote the RKHS corresponding to positive semidefinite kernel $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, then $f(x) = \langle \kappa(x, \cdot), f \rangle_{\mathcal{H}}$ for any $f \in \mathcal{H}$ by the reproducing property [14, 15]. We consider the generalized regression model

$$y \quad = \quad f^*(x) + u \tag{5}$$

where x is an $\mathcal{X}$-valued random variable, u is a real-valued random noise term as above, and $f^* \in \mathcal{H}$ is an unknown deterministic function. Given a sample of n i.i.d. observations $(x_1, y_1), ..., (x_n, y_n)$,

[1] Generally one has to introduce an additional scale parameter σ and allow rescaling of the residuals via r_i/σ, to preserve parameter equivariance [3, 4]. However, we will initially assume a known scale.

where each x_i is drawn from some unknown marginal probability measure P_x, and y_i are generated according to (5),[2] the task is then to estimate the unknown deterministic function $f^* \in \mathcal{H}$. To do so we can express the estimator (4) more generally as

$$\hat{f}_{MR} \quad \in \quad \arg\min_{f \in \mathcal{H}} \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \rho(y_i - f(x_i)) + \tfrac{\lambda}{2} \|f\|_{\mathcal{H}}^2. \tag{6}$$

By the representer theorem [14], the solution to (6) can be expressed by $\hat{f}_{MR}(x) = \sum_{i=1}^{n} \alpha_i \kappa(x_i, x)$ for some $\boldsymbol{\alpha} \in \mathbb{R}^n$, and therefore (6) can be recovered by solving the finite dimensional problem

$$\hat{\boldsymbol{\alpha}}_{MR} \quad \in \quad \arg\min_{\boldsymbol{\alpha}} \tfrac{1}{n} \mathbf{1}^T \boldsymbol{\rho}(\mathbf{y} - K\boldsymbol{\alpha}) + \tfrac{\lambda}{2} \boldsymbol{\alpha}^T K \boldsymbol{\alpha} \quad \text{ such that } K_{ij} = \kappa(x_i, x_j). \tag{7}$$

Our interest is understanding the tractability, robustness and consistency aspects of such estimators.

Consistency: Much is known about the consistency properties of estimators expressed as regularized empirical risk minimizers. For example, the ML-estimator (2) and the M-estimator (3) are both known to be parameter consistent under general conditions [16].[3] The regularized M-estimator in RKHSs (6), is loss consistent under some general assumptions on the kernel, loss and training distribution.[4] Furthermore, a weak form of f-consistency has also established in [6]. For bounded kernel and bounded Lipschitz losses, one can similarly prove the loss consistency of the regularized M-estimator (6) (in RKHS). See Appendix C.1 of the supplement for more discussion.

Generally speaking, any estimator that can be expressed as a regularized empirical loss minimization is consistent under "reasonable" conditions. That is, one can consider regularized loss minimization to be a (generally) sound principle for formulating regression estimators, at least from the perspective of consistency. However, this is no longer the case when we consider robustness and tractability; here sharp distinctions begin to arise within this class of estimators.

Robustness: Although robustness is an intuitive notion, it has not been given a unique technical definition in the literature. Several definitions have been proposed, with distinct advantages and disadvantages [4]. Some standard definitions consider the asymptotic invariance of estimators to an infinitesimal but arbitrary perturbation of the underlying distribution, e.g. the influence function [4, 17]. Although these analyses can be useful, we will focus on finite sample notions of robustness since these are most related to concerns of computational tractability. In particular, we focus on the following definition related to the *finite sample breakdown point* [18, 19].

Definition 1 (Bounded Response)**.** *Assuming the parameter set Θ is metrizable, an estimator has* bounded response *if for any finite data sample its output remains in a bounded interior subset of the closed parameter set Θ (or respectively $\mathcal{H}$), no matter how a* single *observation pair is perturbed.*

This is a much weaker definition than having a non-zero breakdown point: a breakdown of ϵ requires that bounded response be guaranteed when any ϵ fraction of the data is perturbed arbitrarily. Bounded response is obviously a far more modest requirement. However, importantly, the definition of bounded response allows the possibility of arbitrary *leverage*; that is, no bound is imposed on the magnitude of a perturbed input (i.e. $\|\mathbf{x}_1\| \to \infty$ or $\kappa(x_1, x_1) \to \infty$). Surprisingly, we find that even such a weak robustness property is difficult to achieve while retaining computational tractability.

Computational Dilemma: The goals of robustness and computational tractability raise a dilemma: it is easy to achieve robustness (i.e. bounded response) or tractability (i.e. polynomial run-time) in a consistent estimator, but apparently not both.

Consider, for example, using a *convex* loss function. These are the best known class of functions that admit computationally efficient polynomial-time minimization [20] (see also [21]). It is sufficient that the objective be polynomial-time evaluable, along with its first and second derivatives,

[2] We are obviously assuming $\mathcal{X}$ is equipped with an appropriate σ-algebra, and $\mathbb{R}$ with the standard Borel σ-algebra, such that the joint distribution P over $\mathcal{X} \times \mathbb{R}$ is well defined and $\kappa(\cdot, \cdot)$ is measurable.

[3] In particular, let $M_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T \boldsymbol{\theta})$, let $M(\boldsymbol{\theta}) = \mathbb{E}(\rho(y_1 - \mathbf{x}_1^T \boldsymbol{\theta}))$, and equip the parameter space Θ with the uniform metric $\|\cdot\|_\Theta$. Then $\hat{\boldsymbol{\theta}}_M^{(n)} \to \boldsymbol{\theta}^*$, provided $\|M_n - M\|_\Theta \to 0$ in outer probability (adopted to avoid measurability issues) and $M(\boldsymbol{\theta}^*) > \sup_{\boldsymbol{\theta} \in G} M(\boldsymbol{\theta})$ for every open set G that contains $\boldsymbol{\theta}^*$. The latter assumption is satisfied in particular when $M : \Theta \mapsto \mathbb{R}$ is upper semicontinuous with a unique maximum at $\boldsymbol{\theta}^*$. It is also possible to derive asymptotic convergence rates for general M-estimators [16].

[4] Specifically, let $\rho^* = \inf_{f \in \mathcal{H}} E[\rho(y_1 - f(x_1))]$. Then [6] showed that $\frac{1}{n} \sum_{i=1}^{n} \rho(y_i - \hat{f}_{MR}(x_i)) \to \rho^*$ provided the regularization constant $\lambda_n \to 0$ and $\lambda_n^2 n \to \infty$, the loss ρ is convex and Lipschitz-continuous, and the RKHS $\mathcal{H}$ (induced by some bounded measurable kernel κ) is separable and dense in $L_1(\mathbb{P})$ (the space of $\mathbb{P}$-integrable functions) for all distributions $\mathbb{P}$ on $\mathcal{X}$. Also, $\mathcal{Y} \subset \mathbb{R}$ is required to be *closed* where $y \in \mathcal{Y}$.

and that the objective be *self-concordant* [20].[5] Since a Tikhonov regularizer is automatically self-concordant, the minimization problems outlined above can all be solved in polynomial time with Newton-type algorithms, provided $\rho(r)$, $\rho'(r)$, and $\rho''(r)$ can all be evaluated in polynomial time for a self-concordant ρ [22, Ch.9]. Standard loss functions, such as squared error or Huber's loss satisfy these conditions, hence the corresponding estimators are polynomial-time.

Unfortunately, loss minimization with a (non-constant) convex loss yields unbounded response to even a single outlier [3, Ch.5]. We extend this result to also account for regularization and RKHSs.

Theorem 1. *Empirical risk minimization based on a (non-constant) convex loss cannot have bounded response if the domain (or kernel) is unbounded, even under Tikhonov regularization.* (Proof given in Appendix B of the supplement.)

By contrast, consider the case of a (non-constant) *bounded* loss function.[6] Bounded loss functions are a common choice in robust regression because they not only ensure bounded response, trivially, they can also ensure a high breakdown point of $(n-p)/(2n)$ [3, Ch.5]. Unfortunately, estimators based on bounded losses are inherently intractable.

Theorem 2. *Bounded (non-constant) loss minimization is NP-hard.* (Proof given in Appendix E.)

These difficulties with empirical risk minimization have led the field of robust statistics to develop a variety of alternative estimators [4, Ch.7]. For example, [7] recommends subset-selection based regression estimators, such as Least Trimmed Loss:

$$\hat{\boldsymbol{\theta}}_{LTL} \quad \in \quad \arg\min_{\boldsymbol{\theta}\in\Theta} \sum_{i=1}^{n'} \rho(r_{[i]}). \tag{8}$$

Here $r_{[i]}$ denotes sorted residuals $r_{[1]} \leq \cdots \leq r_{[n]}$ and $n' < n$ is the number of terms to consider. Traditionally $\rho(r) = r^2$ is used. These estimators are known to have high breakdown [7],[7] and obviously demonstrate bounded response to single outliers. Unfortunately, (8) is NP-hard [1].

3 Variational M-estimation

To address the dilemma, we first adopt a general form of adaptive M-estimator that allows flexibility while allowing a general approximation strategy. The key construction is a variational representation of M-estimation that can express a number of standard robust (and non-robust) methods in a common framework. In particular, consider the following adaptive form of loss function

$$\rho(r) \quad = \quad \min_{0\leq\eta\leq 1} \eta\ell(r) + \psi(\eta). \tag{9}$$

where r is a residual value, ℓ is a closed *convex* base loss, η is an adaptive weight on the base loss, and ψ is a convex auxiliary function. The weight can choose to ignore the base loss if $\ell(r)$ is large, but this is balanced against a prior penalty $\psi(\eta)$. Different choices of base loss and auxiliary function will yield different results, and one can represent a wide variety of loss functions ρ in this way [8]. For example, any convex loss ρ can be trivially represented in the form (9) by setting $\ell = \rho$, and $\psi(\eta) = \delta_{\{1\}}(\eta)$.[8] *Bounded* loss functions can also be represented in this way, for example

$$\text{(Geman-McClure) [8]} \quad \rho(r) = \tfrac{r^2}{1+r^2} \qquad \ell(r) = r^2 \qquad \psi(\eta) = (\sqrt{\eta}-1)^2 \tag{10}$$

$$\text{(Geman-Reynolds) [8]} \quad \rho(r) = \tfrac{|r|}{1+|r|} \qquad \ell(r) = |r| \qquad \psi(\eta) = (\sqrt{\eta}-1)^2 \tag{11}$$

$$\text{(LeClerc) [8]} \quad \rho(r) = 1-\exp(-\ell(r)) \qquad \ell(\cdot) \text{ convex} \qquad \psi(\eta) = \eta\log\eta - \eta + 1 \tag{12}$$

$$\text{(Clipped-loss) [9]} \quad \rho(r) = \max(1, \ell(r)) \qquad \ell(\cdot) \text{ convex} \qquad \psi(\eta) = 1-\eta. \tag{13}$$

Appendix D in the supplement demonstrates how one can represent general functions ρ in the form (9), not just specific examples, significantly extending [8] with a general characterization.

[5] A function ρ is *self-concordant* if $|\rho'''(r)| \leq 2\rho''(r)^{3/2}$; see e.g. [22, Ch.9].

[6] A bounded function obviously cannot be convex over an unbounded domain unless it is constant.

[7] When n' approaches $n/2$ the breakdown of (8) approaches $1/2$ [7].

[8] We use $\delta_C(\eta)$ to denote the indicator for the point set C; i.e., $\delta_C(\eta) = 0$ if $\eta \in C$, otherwise $\delta_C(\eta) = \infty$.

Therefore, all of the previous forms of regularized empirical risk minimization, whether with a convex or bounded loss ρ, can be easily expressed using only convex base losses ℓ and convex auxiliary functions ψ, as follows

$$\hat{\boldsymbol{\theta}}_{VM} \in \arg\min_{\boldsymbol{\theta}\in\Theta} \min_{0\le\boldsymbol{\eta}\le 1} \boldsymbol{\eta}^T\boldsymbol{\ell}(\mathbf{y}-X\boldsymbol{\theta}) + \mathbf{1}^T\boldsymbol{\psi}(\boldsymbol{\eta}) + \tfrac{\lambda}{2}\|\boldsymbol{\eta}\|_1\|\boldsymbol{\theta}\|_2^2 \tag{14}$$

$$\hat{f}_{VM} \in \arg\min_{f\in\mathcal{H}} \min_{0\le\boldsymbol{\eta}\le 1} \textstyle\sum_{i=1}^n \{\eta_i\ell(y_i - f(x_i)) + \psi(\eta_i)\} + \tfrac{\lambda}{2}\|\boldsymbol{\eta}\|_1\|f\|_{\mathcal{H}}^2 \tag{15}$$

$$\hat{\boldsymbol{\alpha}}_{VM} \in \arg\min_{\boldsymbol{\alpha}} \min_{0\le\boldsymbol{\eta}\le 1} \boldsymbol{\eta}^T\boldsymbol{\ell}(\mathbf{y}-K\boldsymbol{\alpha}) + \mathbf{1}^T\boldsymbol{\psi}(\boldsymbol{\eta}) + \tfrac{\lambda}{2}\|\boldsymbol{\eta}\|_1\boldsymbol{\alpha}^T K\boldsymbol{\alpha}. \tag{16}$$

Note that we have added a regularizer $\|\boldsymbol{\eta}\|_1/n$, which increases robustness by encouraging η weights to prefer small values (but adaptively increase on indices with small loss). This particular form of regularization has two advantages: (i) it is a smooth function of $\boldsymbol{\eta}$ on $0 \le \boldsymbol{\eta} \le 1$ (since $\|\boldsymbol{\eta}\|_1 = \mathbf{1}^T\boldsymbol{\eta}$ in this case), and (ii) it enables a tight convex approximation strategy, as we will see below.

Note that other forms of robust regression can be expressed in a similar framework. For example, generalized M-estimation (GM-estimation) can be formulated simply by forcing each η_i to take on a specific value determined by $\|x_i\|$ or r_i [7], ignoring the auxilary function ψ. Least Trimmed Loss (8) can be expressed in the form (9) provided only that we add a shared constraint over $\boldsymbol{\eta}$:

$$\hat{\boldsymbol{\theta}}_{LTL} \in \arg\min_{\boldsymbol{\theta}\in\Theta} \min_{0\le\boldsymbol{\eta}\le 1:\mathbf{1}^T\boldsymbol{\eta}=n'} \boldsymbol{\eta}^T\boldsymbol{\ell}(\mathbf{r}) + \boldsymbol{\psi}(\boldsymbol{\eta}) \tag{17}$$

where $\psi(\eta_i) = 1 - \eta_i$ and $n' < n$ specifies the number of terms to consider in the sum of losses. Since $\boldsymbol{\eta} \in \{0,1\}^n$ at a solution (see e.g. [9]), (17) is equivalent to (8) if ψ is the clipped loss (13).

These formulations are all convex in the parameters given the auxiliary weights, and vice versa. However, they are not jointly convex in the optimization variables (i.e. in $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$, or in $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$). Therefore, one is not assured that the problems (14)–(16) have only global minima; in fact local minima exist and global minima cannot be easily found (or even verified).

4 Computationally Efficient Approximation

We present a general approximation strategy for the variational regression estimators above that can guarantee polynomial run-time while ensuring certain robustness and consistency properties. The approximation is significantly tighter than the existing work [9], which allows us to achieve stronger guarantees while providing better empirical performance. In developing our estimator we follow standard methodology from combinatorial optimization: given an intractable optimization problem, first formulate a (hopefully tight) convex relaxation that provides a lower bound on the objective, then round the relaxed minimizer back to the feasible space, hopefully verifying that the rounded solution preserves desirable properties, and finally re-optimize the rounded solution to refine the result; see e.g. [23].

To maintain generality, we formulate the approximate estimator in the RKHS setting. Consider (16). Although the problem is obviously convex in $\boldsymbol{\alpha}$ given $\boldsymbol{\eta}$, and vice versa, it is not jointly convex (recall the assumption that ℓ and ψ are both convex functions). This suggests that an obvious computational strategy for computing the estimator (16) is to alternate between $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ optimizations (or use heuristic methods [2]), but this cannot guarantee anything other than local solutions (and thus may not even achieve any of the desired theoretical properties associated with the estimator).

Reformulation: We first need to reformulate the problem to allow a tight relaxation. Let $\Delta(\boldsymbol{\eta})$ denote putting a vector $\boldsymbol{\eta}$ on the main diagonal of a square matrix, and let $\circ$ denote componentwise multiplication. Since ℓ is closed and convex by assumption, we know that $\ell(r) = \sup_\nu \nu r - \nu\ell^*(\nu)$, where ℓ^* is the Fenchel conjugate of ℓ [22]. This allows (16) to be reformulated as follows.

Lemma 1.
$$\min_{0\le\boldsymbol{\eta}\le 1} \min_{\boldsymbol{\alpha}} \boldsymbol{\eta}^T\boldsymbol{\ell}(\mathbf{y}-K\boldsymbol{\alpha}) + \mathbf{1}^T\boldsymbol{\psi}(\boldsymbol{\eta}) + \tfrac{\lambda}{2}\|\boldsymbol{\eta}\|_1\boldsymbol{\alpha}^T K\boldsymbol{\alpha} \tag{18}$$

$$= \min_{0\le\boldsymbol{\eta}\le 1} \sup_{\boldsymbol{\nu}} \mathbf{1}^T\boldsymbol{\psi}(\boldsymbol{\eta}) - \boldsymbol{\eta}^T(\boldsymbol{\ell}^*(\boldsymbol{\nu}) - \Delta(\mathbf{y})\boldsymbol{\nu}) - \tfrac{1}{2\lambda}\boldsymbol{\nu}^T\left(K \circ (\boldsymbol{\eta}\|\boldsymbol{\eta}\|_1^{-1}\boldsymbol{\eta}^T)\right)\boldsymbol{\nu}, \tag{19}$$

where the function evaluations are componentwise. (Proof given in Appendix A of the supplement.)

Although no relaxation has been introduced, the new form (19) has a more convenient structure.

Relaxation: Let $N = \boldsymbol{\eta}\|\boldsymbol{\eta}\|_1^{-1}\boldsymbol{\eta}^T$ and note that, since $0 \leq \boldsymbol{\eta} \leq 1$, N must satisfy a number of useful properties. We can summarize these by formulating a constraint set $N \in \mathcal{N}_{\boldsymbol{\eta}}$ given by:

$$\mathcal{N}_{\boldsymbol{\eta}} = \{N : N \succcurlyeq 0, N\mathbf{1} = \boldsymbol{\eta}, \mathrm{rank}(N) = 1\} \quad (20)$$

$$\mathcal{M}_{\boldsymbol{\eta}} = \{M : M \succcurlyeq 0, M\mathbf{1} = \boldsymbol{\eta}, \mathrm{tr}(M) \leq 1\}. \quad (21)$$

Unfortunately, the set $\mathcal{N}_{\boldsymbol{\eta}}$ is not convex because of the rank constraint. However, relaxing this constraint leads to a set $\mathcal{M}_{\boldsymbol{\eta}} \supseteq \mathcal{N}_{\boldsymbol{\eta}}$ which preserves much of the key structure, as we verify below.

Lemma 2. $$(19) = \min_{0\leq\boldsymbol{\eta}\leq 1}\min_{N\in\mathcal{N}_{\boldsymbol{\eta}}}\sup_{\boldsymbol{\nu}} \mathbf{1}^T\boldsymbol{\psi}(\boldsymbol{\eta}) - \boldsymbol{\eta}^T(\boldsymbol{\ell}^*(\boldsymbol{\nu}) - \Delta(\mathbf{y})\boldsymbol{\nu}) - \tfrac{1}{2\lambda}\boldsymbol{\nu}^T(K\circ N)\boldsymbol{\nu} \quad (22)$$

$$\geq \min_{0\leq\boldsymbol{\eta}\leq 1}\min_{M\in\mathcal{M}_{\boldsymbol{\eta}}}\sup_{\boldsymbol{\nu}} \mathbf{1}^T\boldsymbol{\psi}(\boldsymbol{\eta}) - \boldsymbol{\eta}^T(\boldsymbol{\ell}^*(\boldsymbol{\nu}) - \Delta(\mathbf{y})\boldsymbol{\nu}) - \tfrac{1}{2\lambda}\boldsymbol{\nu}^T(K\circ M)\boldsymbol{\nu}. \quad (23)$$

using the fact that $\mathcal{N}_{\boldsymbol{\eta}} \subseteq \mathcal{M}_{\boldsymbol{\eta}}$. (Proof given in Appendix A of the supplement.)

Crucially, the constraint set $\{(\boldsymbol{\eta}, M) : 0 \leq \boldsymbol{\eta} \leq 1, M \in \mathcal{M}_{\boldsymbol{\eta}}\}$ is jointly convex in $\boldsymbol{\eta}$ and M, thus (23) is a convex-concave min-max problem. To see why, note that the inner objective function is jointly convex in $\boldsymbol{\eta}$ and M, and concave in $\boldsymbol{\nu}$. Since a pointwise maximum of convex functions is convex, the problem is convex in $(\boldsymbol{\eta}, M)$ [22, Ch.3]. We conclude that all local minima in $(\boldsymbol{\eta}, M)$ are global. Therefore, (23) provides the foundation for an efficiently solvable relaxation.

Rounding: Unfortunately the solution to M in (23) does not allow direct recovery of an estimator $\boldsymbol{\alpha}$ achieving the same objective value in (18), unless M satisfies $\mathrm{rank}(M) = 1$. In general we first need to round M to a rank 1 solution. Fortunately, a trivial rounding procedure is available: we simply use $\boldsymbol{\eta}$ (ignoring M) and re-solve for $\boldsymbol{\alpha}$ in (18). This is equivalent to replacing M with the rank 1 matrix $\tilde{N} = \boldsymbol{\eta}\|\boldsymbol{\eta}\|_1^{-1}\boldsymbol{\eta}^T \in \mathcal{N}_{\boldsymbol{\eta}}$, which restores feasibility in the original problem. Of course, such a rounding step will generally increase the objective value.

Reoptimization: Finally, the rounded solution can be locally improved by alternating between $\boldsymbol{\eta}$ and $\boldsymbol{\alpha}$ updates in (18) (or using any other local optimization method), yielding the final estimate $\tilde{\boldsymbol{\alpha}}$.

5 Properties

Although a tight *a priori* bound on the size of the optimality gap is difficult to achieve, a rigorous bound on the optimality gap can be recovered *post hoc* once the re-optimized estimator is computed. Let R_0 denote the minimum value of (18) (not efficiently computable); let R_1 denote the minimum value of (23) (the relaxed solution); let R_2 denote the value of (18) achieved by freezing $\boldsymbol{\eta}$ from the relaxed solution but re-optimizing $\boldsymbol{\alpha}$ (the rounded solution); and finally let R_3 denote the value of (18) achieved by re-optimizing $\boldsymbol{\eta}$ and $\boldsymbol{\alpha}$ from the rounded solution (the re-optimized solution). Clearly we have the relationships $R_1 \leq R_0 \leq R_3 \leq R_2$. An upper bound on the relative optimality gap of the final solution (R_3) can be determined by $(R_3 - R_0)/R_3 \leq (R_3 - R_1)/R_3$, since R_1 and R_3 are both known quantities.

Tractability: Under mild assumptions on ℓ and ψ, computation of the approximate estimator (solving the relaxed problem, rounding, then re-optimizing) admits a polynomial-time solution; see Appendix E in the supplement. (Appendix E also provides details for an efficient implementation for solving (23).) Once $\boldsymbol{\eta}$ is recovered from the relaxed solution, the subsequent optimizations of (18) can be solved efficiently under weak assumptions about ℓ and ψ; namely that they both satisfy the self-concordance and polynomial-time computation properties discussed in Section 2.

Robustness: Despite the approximation, the relaxation remains sufficiently tight to preserve some of the robustness properties of bounded loss minimization. To establish the robustness (and consistency) properties, we will need to make use of a specific technical definition of *outliers* and *inliers*.

Definition 2 (Outliers and Inliers)**.** *For an L-Lipschitz loss ℓ, an outlier is a point (x_i, y_i) that satisfies $\ell(y_i) > L^2K_{ii}/(2\lambda) - \psi'(0)$, while an inlier satisfies $\ell(y_i) + L^2K_{ii}/(2\lambda) < -\psi'(1)$.*

Theorem 3. *Assume the loss ρ is bounded and has a variational representation (9) such that ℓ is Lipschitz-continuous and ψ' is bounded. Also assume there is at least one (unperturbed) inlier, and consider the perturbation of a single data point (y_1, x_1). Under the following conditions, the rounded (re-optimized) estimator maintains bounded response:*
(i) If either y_1 remains bounded, or $\kappa(x_1, x_1)$ remains bounded.
(ii) If $|y_1| \to \infty$, $\kappa(x_1, x_1) \to \infty$ and $\ell(y_1)/\kappa(x_1, x_1) \to \infty$.
(Proof given in Appendix B of the supplement.)

Methods	Outlier Probability					
	$p = 0.4$		$p = 0.2$		$p = 0.0$	
L2	43.5	(13)	57.6	(21.21)	0.52	(0.01)
L1	4.89	(2.81)	3.6	(2.04)	0.52	(0.01)
Huber	4.89	(2.81)	3.62	(2.02)	0.52	(0.01)
LTS	6.72	(7.37)	8.65	(14.11)	0.52	(0.01)
GemMc	0.53	(0.03)	0.52	(0.02)	0.52	(0.01)
[9]	0.52	(0.01)	0.52	(0.01)	0.52	(0.01)
AltBndL2	0.52	(0.01)	0.52	(0.01)	0.52	(0.02)
AltBndL1	0.73	(0.12)	0.74	(0.16)	0.52	(0.01)
CvxBndL2	0.52	(0.01)	0.52	(0.01)	0.52	(0.01)
CvxBndL1	0.53	(0.02)	0.55	(0.05)	0.52	(0.01)

Table 1: RMSE on clean test data for an artificial data set with 5 features and 100 training points, with outlier probability p, and 10000 test data points. Results are averaged over 10 repetitions. Standard deviations are given in parentheses.

Note that the latter condition causes *any* convex loss ℓ to demonstrate unbounded response (see proof of Theorem 5 in Appendix B). Therefore, the approximate estimator is strictly more robust (in terms of bounded response) than regularized empirical risk minimization with a convex loss ℓ.

Consistency: Finally, we can establish consistency of the approximate estimator in a limited albeit non-trivial setting, although we have yet to establish it generally.

Theorem 4. *Assume ℓ is Lipschitz-continuous and $\psi(\eta) = 1 - \eta$. Assume that the data is generated from a mixture of* inliers *and* outliers*, where $P(\text{inlier}) > P(\text{outlier})$. Then the estimate $\hat{\boldsymbol{\theta}}$ produced by the rounded (re-optimized) method is loss consistent.*(Proof given in Appendix C.2.)

6 Experimental Evaluation

We conducted a set of experiments to evaluate the effectiveness of the proposed method compared to standard methods from the literature. Our experimental evaluation was conducted in two parts: first a synthetic experiment where we could control data generation, then an experiment on real data.

The first synthetic experiment was conducted as follows. A target weight vector $\boldsymbol{\theta}$ was drawn from $N(\mathbf{0}, I)$, with $X_{i:}$ sampled uniformly from $[0, 1]^m$, $m = 5$, and outputs y_i computed as $y_i = X_{i:}\boldsymbol{\theta} + \epsilon_i$, $\epsilon_i \sim N(0, \frac{1}{2})$. We then seeded the data set with outliers by randomly re-sampling each y_i and $X_{i:}$ from $N(0, 10^8)$ and $N(0, 10^4)$ respectively, governed by an outlier probability p. Then we randomly sampled 100 points as the training set and another 10000 samples are used for testing.

We implemented the proposed method with two different base losses, L_2 and L_1, respectively; referring to these as CvxBndL2 and CvxBndL1. We compared to standard L_2 and L_1 loss minimization, as well as minimizing the Huber minimax loss (Huber) [4]. We also considered standard methods from the robust statistics literature, including the least trimmed square method (LTS) [7, 24], and bounded loss minimization based on the Geman-McClure loss (GemMc) [8]. Finally we also compared to the alternating minimization strategies outlined at the end of Section 3 (AltBndL2 and AltBndL1 for L_2 and L_1 losses respectively), and implemented the strategy described in [9]. We added the Tikhonov regularization to each method and the regularization parameter λ was selected (optimally for each method) on a separate validation set. Note that LTS has an extra parameter n', which is the number of inliers. The ideal setting $n' = (1 - p)n$ was granted to LTS. We also tried 30 random restarts for LTS and picked the best result.

All experiments are repeated 10 times and the average root mean square errors (RMSE) (with standard deviations) on the clean test data are reported in Table 1. For $p = 0$ (*i.e.* no outliers), all methods perform well; their RMSEs are close to optimal (1/2, the standard deviation of ϵ_i). However, when outliers start to appear, the result of least squares is significantly skewed, while the results of classic robust statistics methods, Huber, L1 and LTS, indeed turn out to be more robust than the least squares, but nevertheless are still affected significantly. Both implementations of the new method performs comparably to the the non-convex Geman-McClure loss while substantially improving the alternating strategy under the L1 loss. Note that the latter improvement clearly demonstrates that

Methods	Datasets							
	cal-housing		abalone		pumadyn		bank-8fh	
L2	1185	(124.59)	7.93	(0.67)	1.24	(0.42)	18.21	(6.57)
L1	1303	(244.85)	7.30	(0.40)	1.29	(0.42)	6.54	(3.09)
Huber	1221	(119.18)	7.73	(0.49)	1.24	(0.42)	7.37	(3.18)
LTS	533	(398.92)	755.1	(126)	0.32	(0.41)	10.96	(6.67)
GemMc	28	(88.45)	2.30	(0.01)	0.12	(0.12)	0.93	(0.80)
[9]	967	(522.40)	8.39	(0.54)	0.81	(0.77)	3.91	(6.18)
AltBndL2	967	(522.40)	8.39	(0.54)	0.81	(0.77)	7.74	(9.40)
AltBndL1	1005	(603.00)	7.30	(0.40)	1.29	(0.42)	1.61	(2.51)
CvxBndL2	9	(0.64)	7.60	(0.86)	0.07	(0.07)	0.20	(0.05)
CvxBndL1	8	(0.28)	2.98	(0.08)	0.08	(0.07)	0.10	(0.07)
Gap(Cvx2)	2e-12	(3e-12)	3e-9	(4e-9)	0.025	(0.052)	0.001	(0.003)
Gap(Cvx1)	0.005	(0.01)	0.001	(0.001)	0.267	(0.269)	0.011	(0.028)

Table 2: RMSE on clean test data for 108 training data points and 1000 test data points, with 10 repeats. Standard deviations shown parentheses. The mean gap values of CvxBndL2 and CvxBndL1, Gap(Cvx2) and Gap(Cvx1) respectively, are given in the last two rows.

alternating can be trapped in poor local minima. The proposal from [9] was not effective in this setting (which differed from the one investigated there).

Next, we conducted an experiment on four real datasets taken from the StatLib repository[9] and DELVE.[10] For each data set, we randomly selected 108 points as the training set, and another random 1000 points as the test set. Here the regularization constant is tuned by 10-fold cross validation. To seed outliers, 5% of the training set are randomly chosen and their X and y values are multiplied by 100 and 10000, respectively. All of these data sets have 8 features, except pumadyn which has 32 features. We also estimated the scale factor on the training set by the mean absolute deviation method, a common method in robust statistics [3]. Again, the ideal parameter $n' = (1 - 5\%)n$ is granted to LTS and 30 random restarts are performed.

The RMSE on test set for all methods are reported in Table 2. It is clear that all methods based on convex losses (L2, L1, Huber) suffer significantly from the added outliers. The method proposed in this paper consistently outperform all other methods with a noticeable margin, except on the abalone data set where GemMc performs slightly better.[11] Again, we observe evidence that the alternating strategy can be trapped in poor local minima, while the method from [9] was less effective. We also measured the relative optimality gaps for the approximate CvxBnd procedures. The gaps were quite small in most cases (the gaps were very close to zero in the synthetic case, and so are not shown), demonstrating the tightness of the proposed approximation scheme.

7 Conclusion

We have developed a new robust regression method that can guarantee a form of robustness (bounded response) while ensuring tractability (polynomial run-time). The estimator has been proved consistent under some restrictive but non-trivial conditions, although we have not established general consistency. Nevertheless, an empirical evaluation reveals that the method meets or surpasses the generalization ability of state-of-the-art robust regression methods in experimental studies. Although the method is more computationally involved than standard approaches, it achieves reasonable scalability in real problems. We are investigating whether the proposed estimator achieves stronger robustness properties, such as high breakdown or bounded influence. It would be interesting to extend the approach to also estimate scale in a robust and tractable manner. Finally, we continue to investigate whether other techniques from the robust statistics and machine learning literatures can be incorporated in the general framework while preserving desired properties.

Acknowledgements

Research supported by AICML and NSERC.

[9] http://lib.stat.cmu.edu/datasets/

[10] http://www.cs.utoronto.ca/ delve/data/summaryTable.html

[11] Note that we obtain different results than [9] arising from a very different outlier process.

References

[1] T. Bernholt. Robust estimators are hard to compute. Technical Report 52/2005, SFB475, U. Dortmund, 2005.

[2] R. Nunkesser and O. Morell. An evolutionary algorithm for robust regression. *Computational Statistics and Data Analysis*, 54:3242–3248, 2010.

[3] R. Maronna, R. Martin, and V. Yohai. *Robust Statistics: Theory and Methods*. Wiley, 2006.

[4] P. Huber and E. Ronchetti. *Robust Statistics*. Wiley, 2nd edition, 2009.

[5] A. Christmann and I. Steinwart. Consistency and robustness of kernel-based regression in convex risk minimization. *Bernoulli*, 13(3):799–819, 2007.

[6] A. Christmann, A. Van Messem, and I. Steinwart. On consistency and robustness properties of support vector machines for heavy-tailed distributions. *Statistics and Its Interface*, 2:311–327, 2009.

[7] P. Rousseeuw and A. Leroy. *Robust Regression and Outlier Detection*. Wiley, 1987.

[8] M. Black and A. Rangarajan. On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *International Journal of Computer Vision*, 19(1): 57–91, 1996.

[9] Y. Yu, M. Yang, L. Xu, M. White, and D. Schuurmans. Relaxed clipping: A global training method for robust regression and classification. In *Advances in Neural Information Processings Systems (NIPS)*, 2010.

[10] A. Bental, L. El Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press, October 2009.

[11] H. Xu, C. Caramanis, and S. Mannor. Robust regression and Lasso. In *Advances in Neural Information Processing Systems (NIPS)*, volume 21, pages 1801–1808, 2008.

[12] H. Xu, C. Caramanis, and S. Mannor. Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, 10:1485–1510, 2009.

[13] S. Mukherjee, P. Niyogi, T. Poggio, and R. Rifkin. Learning theory: Stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1-3):161–193, 2006.

[14] G. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics*, 41(2):495–502, 1970.

[15] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

[16] Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

[17] F. Hampel, E. Ronchetti, P. Rousseeuw, and W. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, 1986.

[18] D. Donoho and P. Huber. The notion of breakdown point. In *A Festschrift for Erich L. Lehmann*, pages 157–184. Wadsworth, 1983.

[19] P. Davies and U. Gather. The breakdown point—examples and counterexamples. *REVSTAT Statistical Journal*, 5(1):1–17, 2007.

[20] Y. Nesterov and A. Nemiroviskii. *Interior-point Polynomial Methods in Convex Programming*. SIAM, 1994.

[21] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer, 2003.

[22] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge U. Press, 2004.

[23] J. Peng and Y. Wei. Approximating k-means-type clustering via semidefinite programming. *SIAM Journal on Optimization*, 18(1):186–205, 2007.

[24] P. Rousseeuw and K. Van Driessen. Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery*, 12(1):29–45, 2006.

[25] R. Horn and C. Johnson. *Matrix Analysis*. Cambridge, 1985.

Supplement to
A Polynomial-time Form of Robust Regression

Yaliang Yu, Özlem Aslan, Dale Schuurmans
Department of Computing Science, University of Alberta, Edmonton AB T6G 2E8, Canada
{yaoliang,ozlem,dale}@cs.ualberta.ca

This supplement contains proofs of the claims made in the paper as well as additional technical developments. The supplement is organized into separate appendices to allow convenient spot checking of particular details of interest.

Appendix A Derivation details for the approximate estimation strategy.

Appendix B Robustness and non-robustness proofs.

Appendix C Consistency properties.

Appendix D Variational representation of loss functions
(including details on when such representations exist and how they can be recovered).

Appendix E Computational issues
(including notes on efficient implementation, and proofs of NP-hardness).

A Deriving the Approximate Estimator

Lemma 1.
$$\min_{0\le\boldsymbol{\eta}\le 1}\min_{\boldsymbol{\alpha}} \boldsymbol{\eta}^T\boldsymbol{\ell}(\mathbf{y}-K\boldsymbol{\alpha}) + \mathbf{1}^T\boldsymbol{\psi}(\boldsymbol{\eta}) + \tfrac{\lambda}{2}\|\boldsymbol{\eta}\|_1\boldsymbol{\alpha}^T K\boldsymbol{\alpha} \tag{24}$$
$$= \min_{0\le\boldsymbol{\eta}\le 1}\sup_{\boldsymbol{\nu}} \mathbf{1}^T\boldsymbol{\psi}(\boldsymbol{\eta}) - \boldsymbol{\eta}^T(\boldsymbol{\ell}^*(\boldsymbol{\nu}) - \Delta(\mathbf{y})\boldsymbol{\nu}) - \tfrac{1}{2\lambda}\boldsymbol{\nu}^T\left(K\circ(\boldsymbol{\eta}\|\boldsymbol{\eta}\|_1^{-1}\boldsymbol{\eta}^T)\right)\boldsymbol{\nu}. \tag{25}$$

where the function evaluations are componentwise.

Proof. The lemma amounts to dualizing the interior convex minimization problem
$$\min_{\boldsymbol{\alpha}} \boldsymbol{\eta}^T\boldsymbol{\ell}(\mathbf{y}-K\boldsymbol{\alpha}) + \mathbf{1}^T\boldsymbol{\psi}(\boldsymbol{\eta}) + \tfrac{\lambda}{2}\|\boldsymbol{\eta}\|_1\boldsymbol{\alpha}^T K\boldsymbol{\alpha}. \tag{26}$$
Recall from the main body that the function ℓ is assumed to be closed and convex. Therefore, strong Fenchel duality holds and we can reexpress ℓ as $\ell(r) = \sup_\nu \nu r - \nu\ell^*(\nu)$ where ℓ^* is the Fenchel conjugate of ℓ [22, Ch.3]. Thus
$$\boldsymbol{\eta}^T\boldsymbol{\ell}(\mathbf{y}-K\boldsymbol{\alpha}) \;=\; \sup_{\boldsymbol{\nu}} \boldsymbol{\nu}^T\Delta(\boldsymbol{\eta})(\mathbf{y}-K\boldsymbol{\alpha}) - \boldsymbol{\nu}^T\boldsymbol{\ell}^*(\boldsymbol{\nu}) \tag{27}$$
and we can therefore re-write (26) as
$$(26) \;=\; \min_{\boldsymbol{\alpha}}\sup_{\boldsymbol{\nu}} \mathbf{1}^T\boldsymbol{\psi}(\boldsymbol{\eta}) - \boldsymbol{\eta}^T\boldsymbol{\ell}^*(\boldsymbol{\nu}) + \boldsymbol{\nu}^T\Delta(\mathbf{y})\boldsymbol{\eta} - \boldsymbol{\nu}^T\Delta(\boldsymbol{\eta})K\boldsymbol{\alpha} + \frac{\lambda}{2}\|\boldsymbol{\eta}\|_1\boldsymbol{\alpha}^T K\boldsymbol{\alpha} \tag{28}$$
$$= \sup_{\boldsymbol{\nu}}\min_{\boldsymbol{\alpha}} \mathbf{1}^T\boldsymbol{\psi}(\boldsymbol{\eta}) - \boldsymbol{\eta}^T\boldsymbol{\ell}^*(\boldsymbol{\nu}) + \boldsymbol{\nu}^T\Delta(\mathbf{y})\boldsymbol{\eta} - \boldsymbol{\nu}^T\Delta(\boldsymbol{\eta})K\boldsymbol{\alpha} + \frac{\lambda}{2}\|\boldsymbol{\eta}\|_1\boldsymbol{\alpha}^T K\boldsymbol{\alpha}, \tag{29}$$
where (29) follows by strong duality (since for all fixed $\boldsymbol{\nu}$ in (29) the sublevel sets in $\boldsymbol{\alpha}$ are bounded [22, Ch.3]).

Finally, $\boldsymbol{\alpha}$ can be eliminated from the inner problem by solving for a critical point (the inner objective is convex in $\boldsymbol{\alpha}$). Taking the gradient in $\boldsymbol{\alpha}$ and setting to zero gives the system of equations $\nabla_{\boldsymbol{\alpha}} = K\Delta(\boldsymbol{\eta})\boldsymbol{\nu} + \lambda\|\boldsymbol{\eta}\|_1 K\boldsymbol{\alpha} = 0$, which is satisfied by $\boldsymbol{\alpha} = \Delta(\boldsymbol{\eta})\boldsymbol{\nu}/(\lambda\|\boldsymbol{\eta}\|_1)$. Substituting this solution back into (29) yields
$$(29) \;=\; \sup_{\boldsymbol{\nu}} \mathbf{1}^T\boldsymbol{\psi}(\boldsymbol{\eta}) - \boldsymbol{\eta}^T(\boldsymbol{\ell}^*(\boldsymbol{\nu}) - \Delta(\mathbf{y})\boldsymbol{\nu}) - \frac{1}{2\lambda\|\boldsymbol{\eta}\|_1}\boldsymbol{\nu}^T\Delta(\boldsymbol{\eta})K\Delta(\boldsymbol{\eta})\boldsymbol{\nu} \tag{30}$$
$$= \sup_{\boldsymbol{\nu}} \mathbf{1}^T\boldsymbol{\psi}(\boldsymbol{\eta}) - \boldsymbol{\eta}^T(\boldsymbol{\ell}^*(\boldsymbol{\nu}) - \Delta(\mathbf{y})\boldsymbol{\nu}) - \frac{1}{2\lambda}\boldsymbol{\nu}^T\left(K\circ(\boldsymbol{\eta}\|\boldsymbol{\eta}\|_1^{-1}\boldsymbol{\eta}^T)\right)\boldsymbol{\nu}, \tag{31}$$
establishing the lemma. ■

To prove the next lemma from the main body, recall the definitions.
$$\mathcal{N}_{\boldsymbol{\eta}} \;=\; \{N : N\succcurlyeq 0, N\mathbf{1} = \boldsymbol{\eta}, \mathrm{rank}(N) = 1\} \tag{32}$$
$$\mathcal{M}_{\boldsymbol{\eta}} \;=\; \{M : M\succcurlyeq 0, M\mathbf{1} = \boldsymbol{\eta}, \mathrm{tr}(M) \le 1\}. \tag{33}$$

Lemma 2.
$$(25) \;=\; \min_{0\le\boldsymbol{\eta}\le 1}\min_{N\in\mathcal{N}_{\boldsymbol{\eta}}}\sup_{\boldsymbol{\nu}} \mathbf{1}^T\boldsymbol{\psi}(\boldsymbol{\eta}) - \boldsymbol{\eta}^T(\boldsymbol{\ell}^*(\boldsymbol{\nu}) - \Delta(\mathbf{y})\boldsymbol{\nu}) - \tfrac{1}{2\lambda}\boldsymbol{\nu}^T\left(K\circ N\right)\boldsymbol{\nu} \tag{34}$$
$$\ge \min_{0\le\boldsymbol{\eta}\le 1}\min_{M\in\mathcal{M}_{\boldsymbol{\eta}}}\sup_{\boldsymbol{\nu}} \mathbf{1}^T\boldsymbol{\psi}(\boldsymbol{\eta}) - \boldsymbol{\eta}^T(\boldsymbol{\ell}^*(\boldsymbol{\nu}) - \Delta(\mathbf{y})\boldsymbol{\nu}) - \tfrac{1}{2\lambda}\boldsymbol{\nu}^T\left(K\circ M\right)\boldsymbol{\nu}. \tag{35}$$

using the fact that $\mathcal{N}_{\boldsymbol{\eta}} \subseteq \mathcal{M}_{\boldsymbol{\eta}}$.

Proof. First, to prove the equality (34) consider any fixed $\boldsymbol{\eta} \in [0,1]^n$. We will show that $N = \boldsymbol{\eta}\|\boldsymbol{\eta}\|_1^{-1}\boldsymbol{\eta}^T \Leftrightarrow N \in \mathcal{N}_{\boldsymbol{\eta}}$, which will immediately yield the equality. The direction $\Rightarrow$ can be verified by a simple check. To prove $\Leftarrow$, assume $N \in \mathcal{N}_{\boldsymbol{\eta}}$. Since $N \succcurlyeq 0$ and $\mathrm{rank}(N) = 1$ we know that $N = \mathbf{q}\lambda\mathbf{q}^T$ for some $\lambda > 0$ and $\mathbf{q}$ such that $\|\mathbf{q}\| = 1$. But now, since $N\mathbf{1} = \boldsymbol{\eta}$, it must follow that $\mathbf{q} = \boldsymbol{\eta}/(\lambda\boldsymbol{\eta}^T\mathbf{1})$, hence $\mathbf{q} = \boldsymbol{\eta}/\|\boldsymbol{\eta}\|$ and $\lambda\mathbf{q}^T\mathbf{1} = \|\boldsymbol{\eta}\|$. Therefore, $\lambda = \|\boldsymbol{\eta}\|/\mathbf{q}^T\mathbf{1} = \|\boldsymbol{\eta}\|^2/(\boldsymbol{\eta}^T\mathbf{1})$. That is, N must have the form $N = \mathbf{q}\lambda\mathbf{q}^T = \boldsymbol{\eta}\|\boldsymbol{\eta}\|_1^{-1}\boldsymbol{\eta}^T$.

The inequality (35) then follows from the above argument, and the fact that $\|\boldsymbol{\eta}\|^2/(\boldsymbol{\eta}^T\mathbf{1}) \le 1$, which implies $\mathrm{tr}(N) \le 1$ for any $N \in \mathcal{N}_{\boldsymbol{\eta}}$. Therefore $\mathcal{N}_{\boldsymbol{\eta}} \subseteq \mathcal{M}_{\boldsymbol{\eta}}$. ■

B Robustness

B.1 Non-robustness of (Non-constant) Convex Losses

Theorem 5. *Empirical risk minimization based on a (non-constant) convex loss cannot have bounded response if the domain (or kernel) is unbounded, even under Tikhonov regularization.*

We will separately prove the two cases; first, assuming an explicit feature representation expressed in a data matrix X, and second, in the case of an RKHS.

B.1.1 Explicit Feature Case

Proof. Consider the L_2-norm regularized M-estimator:

$$\hat{\boldsymbol{\theta}} \quad = \quad \arg\min_{\boldsymbol{\theta}\in\Theta} \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2 + \sum_{i=1}^{n} \rho(y_i - X_{i:}\boldsymbol{\theta}), \tag{36}$$

where ρ is some (non-constant) convex function. For simplicity, we assume ρ is differentiable (otherwise one can consider subdifferentials to arrive at the same conclusion).

Suppose the theorem is false, then $\hat{\boldsymbol{\theta}}$ remains in a bounded interior subset. From the first order optimality condition (see, for instance, [22]), we know that

$$[\rho'(y_1 - X_{1:}\hat{\boldsymbol{\theta}}) + \lambda]X_{1:} + \sum_{i=2}^{n}[\rho'(y_i - X_{i:}\hat{\boldsymbol{\theta}}) + \lambda]X_{i:} \quad = \quad 0. \tag{37}$$

Now we perturb $(X_{1:}, y_1)$ to cause a contradiction. Since ρ is a univariate convex and non-constant function, we apparently have $\lim_{d\to\infty}\rho'(d) > 0$ or $\lim_{d\to-\infty}\rho'(d) < 0$ or both. Assume $\lim_{d\to\infty}\rho'(d) > 0$ (the other case can be proved similarly).

Let y_1 and $\|X_{1:}\|$ tend to infinity in a way that $\frac{y_1}{\|X_{1:}\|}$ also tends to infinity. Then

$$y_1 - X_{1:}\hat{\boldsymbol{\theta}} \quad \geq \quad y_1 - \|X_{1:}\| \cdot \|\hat{\boldsymbol{\theta}}\| \tag{38}$$

$$\quad = \quad \|X_{1:}\| \cdot \left(\frac{y_1}{\|X_{1:}\|} - \|\hat{\boldsymbol{\theta}}\|\right) \tag{39}$$

tends to infinity since $\hat{\boldsymbol{\theta}}$ is bounded. Therefore $\rho'(y_1 - X_{1:}\hat{\boldsymbol{\theta}})$ tends to a positive number. But then the first term in (37) is unbounded in norm while the second term is bounded in norm (for we did not perturb $\{(X_{i:}, y_i)\}_{i=2}^n$), contradiction. ■

Intuitively the meaning of this theorem is clear: If the loss function is unbounded, then any sample that is far enough can drag the estimate $\hat{\boldsymbol{\theta}}$ outside of any bounded interior subset. Tikhonov regularization is not able to overcome this effect. Note that we need to perturb both $X_{1:}$ and y_1, in order to derive a contradiction (for instance, if one only perturbs y_1, then convex functions with bounded derivatives will survive the proof).

B.1.2 RKHS Case

Proof. Let us consider the standard regularized M-estimator:

$$\hat{f}_{MR} \quad \in \quad \arg\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \rho(y_i - \langle\phi(x_i), f\rangle) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2, \tag{40}$$

where $\{(x_i, y_i)\}_{i=1}^n$ are samples from the domain $\mathcal{X}\times\mathbb{R}$; and $\mathcal{H}$ is the RKHS induced by some *unbounded* kernel κ, with its canonical feature map $\phi : \mathcal{X} \mapsto \mathcal{H}$. As above, we assume ρ is some non-constant convex loss function. For simplicity, we assume that the minimum in (40) is attained.

Suppose the theorem is false, then $\hat{f}_{MR}$ remains in a bounded set. The first order necessary condition for the optimality of $\hat{f}_{MR}$ yields

$$\rho'\left(y_1 - \langle\phi(x_1), \hat{f}_{MR}\rangle\right)\phi(x_1) + \sum_{i=2}^{n}\rho'\left(y_i - \langle\phi(x_i), \hat{f}_{MR}\rangle\right)\phi(x_i) + n\lambda\hat{f}_{MR} \quad = \quad 0, \tag{41}$$

where ρ' denotes the subdifferential of ρ (whose existence is guaranteed by convexity, moreover since ρ is finite-valued on $\mathbb{R}$, ρ' is also finite-valued on $\mathbb{R}$). The second term and third term above are bounded since $\hat{f}_{MR}$ is assumed to be bounded. We will perturb (x_1, y_1) such that the first term is not bounded in norm, hence creating a contradiction.

Since ρ is convex and non-constant, we must have either $\lim_{y\to\infty} \rho'(y) > 0$ or $\lim_{y\to-\infty} \rho'(y) > 0$ (or both). Let us assume $\lim_{y\to\infty} \rho'(y) > 0$ (the other case can be proved similarly).

Let both y_1 and $\kappa(x_1, x_1)$ tend to infinity in a way such that $\frac{y_1}{\sqrt{\kappa(x_1,x_1)}}$ also tends to infinity. Then

$$y_1 - \langle \phi(x_1), \hat{f}_{MR} \rangle \geq y_1 - \sqrt{\kappa(x_1, x_1)} \|\hat{f}_{MR}\|_{\mathcal{H}} \quad = \sqrt{\kappa(x_1, x_1)} \left(\frac{y_1}{\sqrt{\kappa(x_1, x_1)}} - \|\hat{f}_{MR}\|_{\mathcal{H}} \right), \tag{42}$$

hence

$$\left\| \rho' \left(y_1 - \langle \phi(x_1), \hat{f}_{MR} \rangle \right) \phi(x_1) \right\|_{\mathcal{H}} \quad = \quad \left| \rho' \left(y_1 - \langle \phi(x_1), \hat{f}_{MR} \rangle \right) \right| \sqrt{\kappa(x_1, x_1)} \ \to \ \infty, \tag{43}$$

due to our assumptions. ■

Note that again we need to perturb both $\kappa(x_1, x_1)$ and y_1 to reach a contradiction (for instance, if one only perturbs y_1, then convex functions with bounded derivatives will survive our proof).

It should be clear in both these proofs that one may replace the L_2-norm with other regularizers without affecting Theorem 5.

B.2 Robustness of the Approximate Regression Estimator

Theorem 3. *Assume the loss ρ is bounded and has a variational representation (9) such that ℓ is Lipschitz-continuous and ψ' is bounded. Also assume there is at least one (unperturbed) inlier, and consider the perturbation of a single data point (y_1, x_1). Under the following conditions, the rounded (re-optimized) estimator maintains bounded response:*
(i) If either y_1 remains bounded, or $\kappa(x_1, x_1)$ remains bounded.
(ii) If $|y_1| \to \infty$, $\kappa(x_1, x_1) \to \infty$ and $\ell(y_1)/\kappa(x_1, x_1) \to \infty$.

To prove this theorem, we first need some key definitions. First recall the definition of outlier and inlier from the main body.

Definition 2 (Outliers and Inliers)**.** *For an L-Lipschitz loss ℓ, an outlier is a point (x_i, y_i) that satisfies $\ell(y_i) > L^2 K_{ii}/(2\lambda) - \psi'(0)$, while an inlier satisfies $\ell(y_i) + L^2 K_{ii}/(2\lambda) < -\psi'(1)$.*

Also recall

$$\mathcal{N}_{\boldsymbol{\eta}} = \{N : N \succcurlyeq 0, N\mathbf{1} = \boldsymbol{\eta}, \operatorname{rank}(N) = 1\} \tag{44}$$
$$\mathcal{M}_{\boldsymbol{\eta}} = \{M : M \succcurlyeq 0, M\mathbf{1} = \boldsymbol{\eta}, \operatorname{tr}(M) \leq 1\}. \tag{45}$$

Let

$$R_0 = \min_{0 \leq \boldsymbol{\eta} \leq 1} \min_{\boldsymbol{\alpha}} \boldsymbol{\eta}^T \boldsymbol{\ell}(\mathbf{y} - K\boldsymbol{\alpha}) + \mathbf{1}^T \boldsymbol{\psi}(\boldsymbol{\eta}) + \frac{\lambda}{2} \|\boldsymbol{\eta}\|_1 \boldsymbol{\alpha}^T K \boldsymbol{\alpha} \tag{46}$$

$$R_1 = \min_{0 \leq \boldsymbol{\eta} \leq 1} \min_{M \in \mathcal{M}_{\boldsymbol{\eta}}} \sup_{\boldsymbol{\nu}} \mathbf{1}^T \boldsymbol{\psi}(\boldsymbol{\eta}) + \boldsymbol{\eta}^T (\Delta(\mathbf{y})\boldsymbol{\nu} - \boldsymbol{\ell}^*(\boldsymbol{\nu})) - \frac{1}{2\lambda} \boldsymbol{\nu}^T (K \circ M) \boldsymbol{\nu} \tag{47}$$

$$R_2 = \sup_{\boldsymbol{\nu}} \mathbf{1}^T \boldsymbol{\psi}(\boldsymbol{\eta}) + \boldsymbol{\eta}^T (\Delta(\mathbf{y})\boldsymbol{\nu} - \boldsymbol{\ell}^*(\boldsymbol{\nu})) - \frac{1}{2\lambda} \boldsymbol{\nu}^T \left(K \circ (\boldsymbol{\eta}\|\boldsymbol{\eta}\|_1^{-1} \boldsymbol{\eta}^T)\right) \boldsymbol{\nu} \tag{48}$$

$$= \min_{\boldsymbol{\alpha}} \boldsymbol{\eta}^T \boldsymbol{\ell}(\mathbf{y} - K\boldsymbol{\alpha}) + \mathbf{1}^T \boldsymbol{\psi}(\boldsymbol{\eta}) + \frac{\lambda}{2} \|\boldsymbol{\eta}\|_1 \boldsymbol{\alpha}^T K \boldsymbol{\alpha}, \tag{49}$$

where $\boldsymbol{\eta}$ in (48) is fixed at the optimal assignment determined by (47), and (49) follows by Lemma 1. Here, R_0 denotes the objective value obtained by the (intractable) oracle minimizer, R_1 is the objective value obtained by the relaxed solution, and R_2 denotes the objective value obtained by the rounded solution.

It is immediate that R_0 and R_1 must be bounded, since

$$R_1 \leq R_0 \leq n\psi(0) < \infty \tag{50}$$

where the first inequality follows from Lemma 2, and the second inequality is achieved by choosing $\boldsymbol{\eta} = \boldsymbol{\alpha} = 0$ in (47). The key question is whether R_2, the rounded objective value, remains bounded. Once this is established for each of the cases, we finally show that this will imply that $\|\hat{f}_2\|_{\mathcal{H}}$ remains bounded, and the theorem will be proved.

B.2.1 Proof for Case (i), Bounded y

Proof. Assume that $\mathbf{y}$ remains bounded. We will need to make use of the fact that, since ℓ is closed and convex, we have

$$\max_{\nu} \eta(y\nu - \ell^*(\nu)) = \eta \max_{\nu} y\nu - \ell^*(\nu) = \eta\ell(y). \tag{51}$$

Therefore, from (48) and (51) it follows that

$$R_2 \leq \sup_{\boldsymbol{\nu}} \mathbf{1}^T \boldsymbol{\psi}(\boldsymbol{\eta}) - \boldsymbol{\eta}^T (\boldsymbol{\ell}^*(\boldsymbol{\nu}) - \Delta(\mathbf{y})\boldsymbol{\nu}) \tag{52}$$
$$= \mathbf{1}^T \boldsymbol{\psi}(\boldsymbol{\eta}) + \boldsymbol{\eta}^T \boldsymbol{\ell}(\mathbf{y}) \tag{53}$$
$$\leq n\gamma \tag{54}$$

such that $\gamma = \max_{0 \leq \eta \leq 1} \psi(\eta) + \max_{|y| \leq B} \ell(y) < \infty$. ∎

B.2.2 Proof for Case (i), Bounded K

Proof. Assume K remains bounded; in particular that $|K_{ij}| \leq B$ for some $B < \infty$. We will need to make use of the fact that ℓ is Lipschitz-continuous. In particular, let ℓ have a Lipschitz-constant of L. Then it follows from Lemma 1 and the definition of Fenchel conjugate [26] that

$$
\begin{aligned}
R_2 &= \sup_{-L\leq\boldsymbol{\nu}\leq L} \mathbf{1}^T\boldsymbol{\psi}(\boldsymbol{\eta}) - \boldsymbol{\eta}^T(\boldsymbol{\ell}^*(\boldsymbol{\nu}) - \Delta(\mathbf{y})\boldsymbol{\nu}) - \frac{1}{2\lambda}\boldsymbol{\nu}^T\left(K\circ(\boldsymbol{\eta}\|\boldsymbol{\eta}\|_1^{-1}\boldsymbol{\eta}^T)\right)\boldsymbol{\nu} && (55)\\
&\leq \sup_{-L\leq\boldsymbol{\nu}\leq L} \mathbf{1}^T\boldsymbol{\psi}(\boldsymbol{\eta}) - \boldsymbol{\eta}^T(\boldsymbol{\ell}^*(\boldsymbol{\nu}) - \Delta(\mathbf{y})\boldsymbol{\nu}) && (56)\\
&= \sup_{-L\leq\boldsymbol{\nu}\leq L} \mathbf{1}^T\boldsymbol{\psi}(\boldsymbol{\eta}) - \boldsymbol{\eta}^T(\boldsymbol{\ell}^*(\boldsymbol{\nu}) - \Delta(\mathbf{y})\boldsymbol{\nu}) - \frac{1}{2\lambda}\boldsymbol{\nu}^T(K\circ M)\boldsymbol{\nu} + \frac{1}{2\lambda}\boldsymbol{\nu}^T(K\circ M)\boldsymbol{\nu} && (57)\\
&\leq \sup_{-L\leq\boldsymbol{\nu}\leq L}\left\{\mathbf{1}^T\boldsymbol{\psi}(\boldsymbol{\eta}) - \boldsymbol{\eta}^T(\boldsymbol{\ell}^*(\boldsymbol{\nu}) - \Delta(\mathbf{y})\boldsymbol{\nu}) - \frac{1}{2\lambda}\boldsymbol{\nu}^T(K\circ M)\boldsymbol{\nu}\right\}\\
&\qquad + \sup_{-L\leq\boldsymbol{\nu}\leq L}\frac{1}{2\lambda}\boldsymbol{\nu}^T(K\circ M)\boldsymbol{\nu} && (58)\\
&= R_1 + \sup_{-L\leq\boldsymbol{\nu}\leq L}\frac{1}{2\lambda}\boldsymbol{\nu}^T(K\circ M)\boldsymbol{\nu} && (59)\\
&\leq n\psi(0) + \sup_{-L\leq\boldsymbol{\nu}\leq L}\frac{1}{2\lambda}\boldsymbol{\nu}^T(K\circ M)\boldsymbol{\nu}. && (60)
\end{aligned}
$$

Therefore it suffices to bound

$$
\begin{aligned}
&\sup_{-L\leq\boldsymbol{\nu}\leq L}\boldsymbol{\nu}^T(K\circ M)\boldsymbol{\nu}\\
&\leq L^2\,\mathrm{tr}(KM) && (61)\\
&\leq L^2Bn, && (62)
\end{aligned}
$$

since $M \succcurlyeq 0$, $\mathrm{tr}(M) \leq 1$, and $\lambda_{max}(K) \leq Bn$ [25, p.292,Thm.5.6.9]. ∎

B.2.3 Proof for Case (ii)

Proof. The proof in this case proceeds differently than the previous cases. Here we assume we are given a fixed data set $(x_1, y_1), ..., (x_n, y_n)$, where only the first data point (x_1, y_1) is perturbed (without loss of generality), so that $|y_1| \to \infty$, $\kappa(x_1, x_1) \to \infty$ and $\ell(y_1)/\kappa(x_1, x_1) \to \infty$. Note that such a point must eventually become an outlier. We will show that this forces the corresponding weight η_1 to eventually satisfy $\eta_1 = 0$ in the relaxed solution (47), which will automatically imply that the rounded value R_2 stays at the same finite value thereafter (since no other data point is perturbed).

Consider the inner objective in (47):

$$
\begin{aligned}
&\mathbf{1}^T\boldsymbol{\psi}(\boldsymbol{\eta}) + \boldsymbol{\eta}^T(\Delta(\mathbf{y})\boldsymbol{\nu} - \boldsymbol{\ell}^*(\boldsymbol{\nu})) - \frac{1}{2\lambda}\boldsymbol{\nu}^T(K\circ M)\boldsymbol{\nu} && (63)\\
&= \mathbf{1}^T\boldsymbol{\psi}(M\mathbf{1}) + \boldsymbol{\eta}^T(\Delta(\mathbf{y})\boldsymbol{\nu} - \boldsymbol{\ell}^*(\boldsymbol{\nu})) - \frac{1}{2\lambda}\boldsymbol{\nu}^T(K\circ M)\boldsymbol{\nu}. && (64)
\end{aligned}
$$

The gradients with respect to the parameters for this objective are given by

$$
\begin{aligned}
\nabla_{\boldsymbol{\nu}} &= \Delta(\boldsymbol{\eta})(\mathbf{y} - \boldsymbol{\ell}^*(\boldsymbol{\nu})') - \frac{1}{\lambda}(K\circ M)\boldsymbol{\nu} && (65)\\
\nabla_M &= \boldsymbol{\psi}'(M\mathbf{1})\mathbf{1}^T + (\Delta(\mathbf{y})\boldsymbol{\nu} - \boldsymbol{\ell}^*(\boldsymbol{\nu}))\mathbf{1}^T - \frac{1}{2\lambda}(K\circ\boldsymbol{\nu}\boldsymbol{\nu}^T) && (66)\\
\frac{d}{d\eta_1} &= \psi'(\eta_1) + y_1\nu_1 - \ell^*(\nu_1) - \frac{1}{2\lambda n}K_{1:}\boldsymbol{\nu}\nu_1. && (67)
\end{aligned}
$$

Since ℓ is Lipschitz-continuous with Lipschitz constant L, we know that $|\nu| \leq L$ [26], hence

$$
\left|\frac{1}{2\lambda n}K_{1:}\boldsymbol{\nu}\nu_1\right| \leq \frac{\max|K_{1:}|}{2\lambda n}L^2n = \frac{L^2}{2\lambda}K_{11}. \qquad (68)
$$

Now consider the tentative assignment $M_{1:} = \mathbf{0}^T$, $\eta_1 = 0$. At this assignment, all of the quadratic terms in ν_1 have been nullified in (63) and (64), leaving the optimization over ν_1 as

$$\max_{\nu_1} y_1\nu_1 - \ell^*(\nu_1) \quad = \quad \ell(y_1) \tag{69}$$

by (51). Now note that at this solution for ν_1 we have

$$\frac{d}{d\eta_1} \quad = \quad \psi'(\eta_1) + \ell(y_1) - \frac{1}{2\lambda n} K_{1:}\boldsymbol{\nu}\nu_1 \tag{70}$$

$$\frac{d}{d\eta_1}\bigg|_{\eta_1=0} \quad = \quad \psi'(0) + \ell(y_1) - \frac{1}{2\lambda n} K_{1:}\boldsymbol{\nu}\nu_1. \tag{71}$$

Therefore, if $\ell(y_1) > \frac{L^2}{2\lambda}K_{11} - \psi'(0)$ (the outlier condition) then $\frac{d}{d\eta_1}\Big|_{\eta_1=0} > 0$, which implies that η_1 stays at 0. We conclude that once the outlier condition is achieved (guaranteed by the assumptions), R_2 retains the same finite value after all subsequent perturbations of (x_1, y_1), independent of $\ell(y_1)$ and $\kappa(x_1, x_1)$. ■

B.2.4 Final Step: Bounding $\|\hat{f}\|_{\mathcal{H}}$

Consider the rounded estimate $\hat{f}_2$, corresponding to the solution to (49). It remains to bound $\|\hat{f}_2\|_{\mathcal{H}}$ in all the three cases discussed above.

Proof. Observe that

$$\|\hat{f}_2\|_{\mathcal{H}} \quad = \quad \boldsymbol{\alpha}^T K \boldsymbol{\alpha} \;\leq\; \frac{2R_2}{\lambda\|\boldsymbol{\eta}\|_1}. \tag{72}$$

Since R_2 has been proved bounded under the stated assumptions, we only need to consider the behavior of $\|\boldsymbol{\eta}\|_1$, which we do in two cases: $\boldsymbol{\eta} = 0$ and $\boldsymbol{\eta} \neq 0$.

First, assume $\boldsymbol{\eta} = 0$. Then $\boldsymbol{\alpha} = 0$ is an optimal solution of (49), implying that $\|\hat{f}_2\|_{\mathcal{H}} = 0$, which is obviously bounded.

So it only remains to consider $\boldsymbol{\eta} \neq 0$. In this case it suffices to bound $\|\boldsymbol{\eta}\|_1$ away from zero. We achieve this by appealing to the assumption that there is at least one *inlier*; that is, an unperturbed point (x_i, y_i), $i \neq 1$, such that $\ell(y_i) + L^2K_{ii}/(2\lambda) < -\psi'(1)$. For any such point, we can establish that $\eta_i = 1$, and the result will follow.

To show that any inlier gets weight η_i in the relaxed solution (47), tentatively consider the assignment $\eta_i = 1$ then recall from (67) that

$$\frac{d}{d\eta_i} \quad = \quad \psi'(\eta_i) + y_i\nu_i - \ell^*(\nu_i) - \frac{1}{2\lambda n} K_{i:}\boldsymbol{\nu}\nu_i \tag{73}$$

$$\frac{d}{d\eta_i}\bigg|_{\eta_i=1} \quad = \quad \psi'(1) + y_i\nu_i - \ell^*(\nu_i) - \frac{1}{2\lambda n} K_{i:}\boldsymbol{\nu}\nu_i. \tag{74}$$

in the relaxed problem. By (69) above we know that $y_i\nu_i - \ell^*(\nu_i) \leq \ell(y_i)$. Furthermore, by (68) we know that $|K_{i:}\boldsymbol{\nu}\nu_i/(2\lambda n)| \leq L^2K_{ii}/(2\lambda)$. Therefore, if the condition $\ell(y_i) + L^2K_{ii}/(2\lambda) < -\psi'(1)$ is satisfied then $\frac{d}{d\eta_i}\Big|_{\eta_i=1} < 0$, which implies that η_i stays at 1. ■

C Consistency

C.1 Consistency of Standard Regression Estimators

Let us start by establishing some notation. Recall that $\mathcal{H}$ is the RKHS induced by some kernel κ, $\{(x_i, y_i)\}_{i=1}^n$ are *i.i.d.*samples from the underlying training distribution $\mathbb{P}(x, y)$. For any function f (possibly random), define

$$\mathcal{R}(f) := \int_{\mathcal{X}\times\mathbb{R}} \rho(y - f(x))\mathrm{d}\mathbb{P}(x, y).$$

Then the following functions all aim at minimizing $\mathcal{R}(f)$ in one way or another:

$$f^* \in \operatorname*{argmin}_f \mathcal{R}(f) \tag{75}$$

$$f_{\mathcal{H}} \in \operatorname*{argmin}_{f\in\mathcal{H}} \mathcal{R}(f) \tag{76}$$

$$f_{\mathcal{H},\lambda} \in \operatorname*{argmin}_{f\in\mathcal{H}} \mathcal{R}(f) + \lambda\|f\|_{\mathcal{H}}^2 \tag{77}$$

$$\hat{f}_{\mathcal{H},\lambda} \in \operatorname*{argmin}_{f\in\mathcal{H}} \hat{\mathcal{R}}(f) + \lambda\|f\|_{\mathcal{H}}^2, \tag{78}$$

where $\hat{\mathcal{R}}(f) := \frac{1}{n}\sum_{i=1}^n \rho(y_i - f(x_i))$. For simplicity, we have tacitly assumed the existence of all minimizers in the above. We will also ignore measurability issues for the time being. Note that only the last function $\hat{f}_{\mathcal{H},\lambda}$ depends on the data.

The regularized M-estimator $\hat{f}_{\mathcal{H},\lambda}$ is said to be (loss) consistent if

$$\mathcal{R}(\hat{f}_{\mathcal{H},\lambda}) - \mathcal{R}(f^*) \to 0 \tag{79}$$

as the sample size n increases to infinity and the regularization constant λ decreases to zero. To investigate when consistency can be assured, the following decomposition is standard and helpful [15]:

$$\begin{aligned}\mathcal{R}(\hat{f}_{\mathcal{H},\lambda}) - \mathcal{R}(f^*) = \quad & \mathcal{R}(\hat{f}_{\mathcal{H},\lambda}) - \mathcal{R}(f_{\mathcal{H},\lambda}) - \lambda\|f_{\mathcal{H},\lambda}\|_{\mathcal{H}}^2 && (80)\\ & + \mathcal{R}(f_{\mathcal{H},\lambda}) + \lambda\|f_{\mathcal{H},\lambda}\|_{\mathcal{H}}^2 - \mathcal{R}(f_{\mathcal{H}}) && (81)\\ & + \mathcal{R}(f_{\mathcal{H}}) - \mathcal{R}(f^*), && (82)\end{aligned}$$

where the last term is usually called the approximation error, which measures how well can functions in $\mathcal{H}$ approximate f^* under the ρ loss; the second term can be thought of as some stability error (where the regularization constant λ plays the role of perturbation); and the first term is related to the sampling error. Note that the last two terms depend only on the interaction between the RKHS $\mathcal{H}$ (hence the kernel κ), the training distribution $\mathbb{P}(x, y)$, and the loss ρ. It is however independent of any estimation procedure (except perhaps providing some insights on how to practically tune the regularization constant). Very general bounds for the last two terms exist in the learning theory literature, see, for instance [15] and [27]. To make our presentation less complicated we will simply assume the sum of the last two terms, as a function of λ, goes to 0 when λ itself decreases to 0. The interested reader can consult the books [15] and [27] for precise technical conditions under which this is indeed so.

The right-hand side in (80) is apparently upper bounded by

$$\begin{aligned}\mathcal{R}(\hat{f}_{\mathcal{H},\lambda}) + \lambda\|\hat{f}_{\mathcal{H},\lambda}\|_{\mathcal{H}}^2 - \mathcal{R}(f_{\mathcal{H},\lambda}) - \lambda\|f_{\mathcal{H},\lambda}\|_{\mathcal{H}}^2 &= \mathcal{R}(\hat{f}_{\mathcal{H},\lambda}) - \hat{\mathcal{R}}(\hat{f}_{\mathcal{H},\lambda})\\ &\quad + \hat{\mathcal{R}}(\hat{f}_{\mathcal{H},\lambda}) + \lambda\|\hat{f}_{\mathcal{H},\lambda}\|_{\mathcal{H}}^2 - \hat{\mathcal{R}}(f_{\mathcal{H},\lambda}) - \lambda\|f_{\mathcal{H},\lambda}\|_{\mathcal{H}}^2\\ &\quad + \mathcal{R}(f_{\mathcal{H},\lambda}) - \hat{\mathcal{R}}(f_{\mathcal{H},\lambda})\\ &\le \mathcal{R}(\hat{f}_{\mathcal{H},\lambda}) - \hat{\mathcal{R}}(\hat{f}_{\mathcal{H},\lambda}) + \mathcal{R}(f_{\mathcal{H},\lambda}) - \hat{\mathcal{R}}(f_{\mathcal{H},\lambda})\\ &\le \sup_{f:\|f\|_{\mathcal{H}}\le\frac{1}{\sqrt{\lambda}}} |\mathcal{R}(f) - \hat{\mathcal{R}}(f)|, \qquad (83)\end{aligned}$$

where the first inequality is due to the optimality of $\hat{f}_{\mathcal{H},\lambda}$, and the second inequality follows under the assumption $0 \le \rho \le 1$. Therefore we have related the right-hand side in (80) to the (uniform) sampling error. Applying standard uniform convergence bounds, for instance, the Rademacher complexity bound in [28, Theorem 8; Theorem 12; Lemma 22], leads to the following consistency result:

Proposition 1. *Assuming* (81) *and* (82) *approach 0 when* $\lambda \to 0$*, the loss* ρ *is Lipschitz and bounded between 0 and 1,* $\sup_{x \in \mathcal{X}} \kappa(x, x) \leq 1$*, then the regularized* M*-estimator defined in* (78) *is (*ρ*-loss) consistent.*

It is also possible to derive consistency results that depends on the capacity of the RKHS [15].

C.2 Consistency of Approximate Regression Estimators

Theorem 4. *Assume* ℓ *is Lipschitz-continuous and* $\psi(\eta) = 1 - \eta$*. Assume that the data is generated from a mixture of* inliers *and* outliers*, where* $P(inlier) > P(outlier)$*. Then the estimate* $\hat{\boldsymbol{\theta}}$ *produced by the rounded (re-optimized) method is loss consistent.*

Proof. From the proof of Theorem 3 above, we know that the relaxed solution will set $\eta_i = 0$ for all outliers, and $\eta_i = 1$ for all inliers. Since only outliers and inliers are present in the data, the solution $\boldsymbol{\eta}$ will be discrete $\{0, 1\}^n$, and all outliers will be ignored. Thus the reoptimized estimator $\hat{f}_2$ is based only on the inliers, and on such data the estimator is solved by a standard regularized empirical risk minimization. The consistency then follows from the standard results on regularized empirical risk minimization above. ■

D Variational Factorization of Loss Functions

Our objective in this section is to study some variational representation of loss functions. Our treatment is inspired by but different from that presented in [8].

Given any loss function $\rho(r)$ (not necessarily convex), we want to find (closed and proper) convex functions $\ell(r)$ and $\psi(\eta)$ such that

$$\rho(r) \quad = \quad \min_{\eta\in[0,1]} \eta\ell(r) + \psi(\eta). \tag{84}$$

Our main motivation to find such a variational representation for ρ is from optimization: the right hand side is separately (although not jointly) convex in both η and r, hence we can optimize ρ by dealing alternatively with ℓ and ψ, if the latter are in much simpler forms. This variational representation also allows us to develop a promising convex relaxation, as we will see soon.

Note that due to the closedness assumption on ψ, the minimum in (84) is always attained hence justifies our notation of $\min$ instead of $\inf$. Also due to the convexity we pose on ℓ, the effective domain of ρ is necessarily convex. We also mention that the convexity of ψ simplifies our presentation but is not essential, for we can always replace ψ by its (closed) convex hull over the interval $[0, 1]$.

It is obvious that all closed convex functions ρ admit a variational representation (84): Just set $\ell(r) = \rho(r), \psi(\eta) = \delta_{\{1\}}(\eta)$.[12] In fact, much more can be said. But let us prove a technical lemma first.

For any closed convex function $\psi(\eta)$, define its *restricted* Fenchel conjugate as:

$$\hat{\psi}(\hat{\eta}) \quad := \quad \max_{\eta\in[0,1]} \eta\hat{\eta} - \psi(\eta). \tag{85}$$

Note that $\hat{\psi}(\hat{\eta}) \le \max_{\eta\in[0,1]} \eta\hat{\eta} - \min_{\eta\in[0,1]} \psi(\eta) = (\hat{\eta})_+ - \min_{\eta\in[0,1]} \psi(\eta) < \infty$, due to the closedness assumption on ψ. Hence $\hat{\psi}$ is defined on the whole real line.

Denote F as the set of all closed (proper) convex functions on $\bar{\mathbb{R}}$, $\hat{F}$ the set of all closed (proper) convex functions $\hat{\ell}$ with effective domain $\mathbb{R}$ and satisfying[13]

$$0 \le \limsup_{\hat{\eta}\to-\infty} \frac{\hat{\psi}(\hat{\eta})}{\hat{\eta}} \le \liminf_{\hat{\eta}\to\infty} \frac{\hat{\psi}(\hat{\eta})}{\hat{\eta}} \le 1. \tag{86}$$

Define the mapping $\mathcal{F}$ which maps $\psi \in F$ into its restricted Fenchel conjugate $\hat{\psi}$.

Lemma 3. *$\mathcal{F}$ maps F onto $\hat{F}$.*

Proof. We first prove the range of the mapping $\mathcal{F}$ is contained in $\hat{F}$. By definition,

$$\liminf_{\hat{\eta}\to\infty} \frac{\hat{\psi}(\hat{\eta})}{\hat{\eta}} \quad = \quad \liminf_{\hat{\eta}\to\infty} \max_{\eta\in[0,1]} \eta - \frac{\psi(\eta)}{\hat{\eta}} \tag{87}$$

$$\le \quad \liminf_{\hat{\eta}\to\infty} 1 - \frac{\min_{\eta\in[0,1]} \psi(\eta)}{\hat{\eta}} \tag{88}$$

$$= \quad 1. \tag{89}$$

The last equality is because of the closedness of ψ. The other inequality in (86) can be similarly proved.

Let $\hat{\varphi} \in \hat{F}$, consider its (ordinary) Fenchel conjugate:

$$\hat{\varphi}^*(\eta) \quad = \quad \sup_{\hat{\eta}} \eta\hat{\eta} - \hat{\varphi}(\hat{\eta}) \quad = \quad \sup_{\hat{\eta}} \hat{\eta}\Big(\eta - \frac{\hat{\varphi}(\hat{\eta})}{\hat{\eta}}\Big). \tag{90}$$

[12] As is conventional in convex analysis, we use $\delta_C(\eta)$ to denote the indicator function for the point set C; i.e., $\delta_C(\eta) = 0$ if $\eta \in C$, otherwise $\delta_C(\eta) = \infty$.

[13] It does not matter if we change $\limsup$ to $\liminf$ or vice versa.

By the conditions (86) on $\hat{F}$, there exists a sequence $\hat{\eta}_i \to \infty$ as $i \to \infty$ such that $\frac{\hat{\psi}(\hat{\eta}_i)}{\hat{\eta}_i} \le 1+\epsilon$ for arbitrarily small $\epsilon > 0$ and all sufficiently large i. Now if $\eta > 1$ (hence $\eta \ge 1 + 2\epsilon$ for suitably small ϵ), then it is clear that $\hat{\varphi}^*(\eta) = \infty$. Similar arguments apply when $\eta < 0$, hence the effective domain of $\hat{\varphi}^*$ is contained in $[0,1]$. Let $\psi = \hat{\varphi}^*$, then $\hat{\psi} = \hat{\varphi}^{**} = \hat{\varphi}$. This proves the mapping $\mathcal{F}$ is onto. ■

Remark 1. *If we restrict the effective domain of the functions in F to be in $[0,1]$, then an easy argument of duality establishes the one-one property of the mapping $\mathcal{F}$.*

Remark 2. *It should be clear that the interval $[0,1]$ plays no special role here. Lemma 3 remains true (with obvious modifications) if we replace $[0,1]$ with any other (non-degenerate) compact interval $[a,b]$. However, for our later convex relaxation, it is desirable to have $a \ge 0$.*

Theorem 5. *The function ρ admits a variational representation (84) if and only if $\rho = -\hat{\psi} \circ (-\ell)$ for some $\hat{\psi} \in \hat{F}$, $\ell \in F$. This shall be called the variational factorization of ρ.*

Proof. This is an immediate consequence of Lemma 3:

$$-\rho(r) := -\min_{\eta \in [0,1]} \ \eta \ell(r) + \psi(\eta) \tag{91}$$

$$= \max_{\eta \in [0,1]} \ \eta(-\ell(r)) - \psi(\eta) \tag{92}$$

$$= \hat{\psi}(-\ell(r)). \tag{93}$$

■

Let us now first illustrate a few examples to see how one can apply Theorem 5.

Example 1 (Variational representation of the Geman-McClure function $\frac{r^2}{1+r^2}$)**.** *This example was also discussed in [8], through some integral arguments. Our derivation here, based on Theorem 5, is different. Rewrite $-\rho(r) = -1 + \frac{1}{1+r^2}$, which suggests us choose $\ell(r) = r^2$. Then $\hat{\psi}(\eta) = -1 + \frac{1}{1-\eta}, \forall \eta \le 0$. Note that the decomposition $-\rho = \hat{\psi} \circ (-\ell)$ only determines $\hat{\psi}$ on the range of $-\ell$, hence giving us some flexibility to "complete" $\hat{\psi}$ on the entire real line such that (86) are satisfied. The easiest (nontrivial) way to extend $\hat{\psi}$ is to glue with it an affine function whose slope dominates $\hat{\psi}$. That is, define $\hat{\psi}(\eta) = a\eta + b, \forall \eta \ge 0$, where $a \ge \sup_{\eta \le 0} \hat{\psi}'(\eta) = 1$. Picking $a = 1$ then satisfies both conditions in (86). To make $\hat{\psi}$ closed, we must set $b = \lim_{\eta \to 0} \hat{\psi}(\eta) = 0$. To summarize, $\rho(r) = \frac{r^2}{1+r^2}$ admits the variational representation (84) with $\ell(r) = r^2$, $\psi(\eta) = \hat{\psi}^*(\eta) = \sup_{\eta^*} \eta\eta^* - \hat{\psi}(\eta^*) = (\sqrt{\eta} - 1)^2$.*

Example 2 (Variational representation of the sigmoid function $\frac{\exp(r)}{1+\exp(r)}$)**.** *This example is new. It was not discussed in [8], probably because there ℓ was restricted to the squared loss, which turns out to be prohibitive in this case. We simply repeat the procedure in Example 1. The similarity between the sigmoid function and the Geman-McClure function suggests us choose $\ell(r) = \exp(r)$, hence $\hat{\psi}(\eta)$ is accordingly determined on $(-\infty, 0)$ as $-1 + \frac{1}{1-\eta}$. Arguing exactly as in Example 1 we get the variational representation of the sigmoid function: $\ell(r) = \exp(r), \psi(\eta) = (\sqrt{\eta} - 1)^2$.*

Example 3 (Variational representation of the Geman-Reynolds function $\frac{-1}{1+|r|}$)**.** *This example was discussed in [8] where $\ell(r) = r^2$, but the final result turned out to be very complicated. We now demonstrate that by choosing a better ℓ, the representation can be significantly simplified. Note that $\rho(\eta) + 1 = \frac{|\eta|}{1+|\eta|}$. By analogy to the previous examples, it is easy to see that $\ell(r) = |r|$, $\psi(\eta) = (\sqrt{\eta} - 1)^2 - 1$ is a correct, and perhaps better, representation.*

The above three examples can be summarized as: $\frac{\ell(r)}{1+\ell(r)}$ with $\ell(r)$ positive and convex has variational representation $\ell(r), \psi(\eta) = (\sqrt{\eta} - 1)^2$.

Example 4 (Some other examples)**.** *These examples (with $\ell(r) = r^2$) were all discussed in [8]. We include them mainly for the purpose of completeness.*

Lorentzian function $\rho(r) = \log(1 + \ell(r))$ *with* ℓ *positive and convex has variational representation:* $\ell(r), \psi(\eta) = \eta - 1 - \log \eta$.

Tukey's biweight function $\rho(r) = \min\{\ell^2(r) - \ell^4(r) + \ell^6(r)/3, 1/3\}$ *with* ℓ *convex has variational representation:* $\ell(r), \psi(\eta) = \frac{1}{3} - \eta + \frac{2}{3}\eta^{3/2}$.

Leclerc's function $\rho(r) = 1 - \exp(-\ell^2(r))$ *with* ℓ *convex has variational representation:* $\ell(r), \psi(\eta) = \eta \log \eta - \eta + 1$.

The mean-field function $\rho(r) = -\log(1 + \exp(-\ell^2(r)))$ *with variational representation:* $\ell(r), \psi(\eta) = \eta \log \eta + (1 - \eta) \log(1 - \eta)$.

The nontrivial part in all examples is to identify the correct form of ℓ; once that is done, the rest is fairly routine. Fortunately, for many useful loss functions, a few trial-and-error usually suffice. While we will develop a more principled way to yield a factorization, we nevertheless find the naive guessing method convenient and that is why we present it first.

We now turn to two immediate questions about Theorem 5:

- Is the variational factorization unique?
- Do all functions admit a variational factorization?

The first question is easily seen to be false since one can scale (or more generally, affinely transform) $\hat{\psi}$ and ℓ accordingly. The answer to the second question is also negative, for as we noted before, the effective domain of ρ needs to be convex. Moreover, by Alexandrov's theorem, $-\rho$ must have second derivative almost everywhere, in order to be a candidate of compositions of convex-concave functions. But it is well-known that there exist plenty of continuous and nowhere differentiable functions. Nevertheless, as noted before, all convex functions do admit a variational factorization (another trivial factorization would be $\hat{\psi}(\eta) = \eta$, $\ell = \rho$, a concrete example of the non-uniquesness). All concave functions whose negations are in $\hat{F}$ also admit a variational factorization (again, a trivial factorization would be $\hat{\psi} = -\rho$, $\ell(r) = -r$).

To have a more satisfying (and involved) answer for the above two questions, we need to put some conditions on the loss function and its variational factors.[14] We shall call a variational factorization $(\hat{\psi}, \ell)$ minimal if $\hat{\psi}$ is strictly increasing, and for any other factorization $(\hat{\varphi}, \xi)$, the function $\hat{\psi}^{-1} \circ \hat{\varphi}$ is convex. Now we can have a positive answer for the uniqueness question.

Lemma 4. *Minimal factorizations, if exist, are unique up to some (appropriate) affine transforms.*

Proof. Let $(\hat{\psi}, \ell), (\hat{\varphi}, \xi)$ both be minimal, then from the definition, we know both $\hat{\psi}^{-1} \circ \hat{\varphi}$ and $\hat{\varphi}^{-1} \circ \hat{\psi}$ are convex. Since both $\hat{\psi}$ and $\hat{\varphi}$ are assumed to be strictly increasing, it suffices to prove that only certain affine function can possibly be strictly increasing, convex, and having a convex inverse. But this is immediate, for we have $\forall \lambda \in [0, 1]$,

$$\begin{aligned} \lambda x + (1 - \lambda)y &= f^{-1} \circ f(\lambda x + (1 - \lambda)y) && (94)\\ &\leq f^{-1}(\lambda f(x) + (1 - \lambda) f(y)) && (95)\\ &\leq \lambda f^{-1} \circ f(x) + (1 - \lambda) f^{-1} \circ f(y) && (96)\\ &= \lambda x + (1 - \lambda)y. && (97) \end{aligned}$$

The equality for all pairs of (x, y) forces f to be affine. ■

The next result requires some tools from real analysis, in particular, the Riesz representation theorem for bounded linear functionals (for a review, refer to [30]).

Lemma 5. *Let* $-\rho$ *be continuous and strictly decreasing on the open interval* $\mathbb{I} \subseteq \mathbb{R}$ *and admits a variational factorization* $(\hat{\psi}, \ell)$ *which are convex, continuous and strictly increasing on* $\{-\ell(r) : r \in \mathbb{I}\}$ *and* $\mathbb{I}$*, respectively, then*

1. *the right derivative,* ρ_r*, of* ρ *exists on* $\mathbb{I}$*;*

[14] Our treatment here is mainly inspired by [29].

2. $\log \rho_r$ *is locally of bounded variation, hence corresponds (uniquely) to a (regular) Radon measure on* $\mathbb{I}$*, which we denote as* $\mathrm{d}(\log \rho_r)$*;*

3. *a minimal factorization of* $-\rho$ *is given as* $(\hat{\varphi}, \xi)$*, with*

$$\xi(x) := \int_a^x \exp(\int_b^y \mathrm{d}\nu_+)\mathrm{d}y, \quad \hat{\varphi}(x) = -\rho(\xi^{-1}(-x)), \tag{98}$$

provided that (98) is finite valued. Here $a = \inf \mathbb{I}, b \in \mathbb{I}$ *and* $\mathrm{d}\nu_+$ *is the positive part of the Radon measure* $\mathrm{d}(\log \rho_r)$*.*

Proof. 1. The proof is obvious hence omitted.

2. Due to the chain rule for derivatives:

$$\begin{aligned} (-\rho)_r &= (\hat{\psi} \circ (-\ell))_r && (99)\\ &= (\hat{\psi}_r \circ (-\ell)) \cdot (-\ell_r), \quad \text{hence} && (100)\\ \log \rho_r &= \log(\hat{\psi}_r \circ (-\ell)) + \log \ell_r. && (101) \end{aligned}$$

Note that all quantities inside the *log* are positive due to the monotonicity assumption. Moreover, the two terms on the right hand side are decreasing and increasing, respectively, due to the convexity assumption. Therefore $\log \rho_r$ is locally of bounded variation. Through the usual Riemann-Stieltjes integral, $\log \rho_r$ induces a bounded linear functional on $C_c(\mathbb{I})$ (continuous functions on $\mathbb{I}$ with compact support), hence by the Riesz representation theorem, corresponds to a unique (regular) Radon measure on $\mathbb{I}$.

3. Suppose $(\hat{\varphi}, \xi)$ is the minimal factorization and $(\hat{\psi}, \ell)$ is any variational factorization of $-\rho$, then from $-\rho = \hat{\varphi} \circ (-\xi) = \hat{\psi} \circ (-\ell)$ and the minimality of $\hat{\varphi}$ we know $f := \hat{\varphi}^{-1} \circ \hat{\psi} = (-\xi) \circ (-\ell)^{-1}$ is convex, hence $-\xi = f \circ (-\ell)$. It follows that $\log \xi_r = \log(f_r \circ (-\ell)) + \log \ell_r$. Since f is convex, $\log(f_r \circ (-\ell))$ is decreasing, therefore $\mathrm{d}(\log \xi_r) \leq \mathrm{d}(\log \ell_r)$. This proves some minimal property of the induced Radon measure of the minimal factorization.

Recall (101) which decomposes $\mathrm{d}(\log \rho_r)$ into a positive measure and a negative one. The result above suggests us choose $\mathrm{d}(\log \xi_r) = \mathrm{d}\nu_+$, for the latter, being the positive part, is minimal by the Hahn-Jordan decomposition. This choice also forces $\hat{\varphi}_r$ to be increasing. Assuming $(\hat{\varphi}, \xi)$ as defined in (98) is finite valued, it is easy to verify the required convexity and monotonicity. ■

Remark 3. *A careful inspection of the proof reveals that the minimal factorization in (98) does not hinge on the fact that* $-\rho$ *indeed admits a variational factorization, as long as* $\log \rho_r$ *is locally of bounded variation. Therefore finding the minimal factorization does not involve more work than finding merely a factorization.*

Again, we digest the previous lemma through an example.

Example 5 (The sigmoid function revisited)**.** $\mathbb{I} = (-\infty, \infty)$ *in this example, hence* $a = -\infty$ *and we pick* $b = 0$ *(any other value would lead to the same result up to some constants). Simple calculation confirms* $\mathrm{d}(\log \rho_r) = \frac{1-\exp(-x)}{1+\exp(x)}\mathrm{d}x$*, where* $\mathrm{d}x$*, as usual, is the Lebesgue measure on* $\mathbb{I}$*. Therefore* $\mathrm{d}\nu_+ = \mathbf{1}\{x \leq 0\} \cdot \frac{1-\exp(-x)}{1+\exp(x)}\mathrm{d}x$*. Direct integration in (98) gives* $\xi(x) = \mathbf{1}\{x \leq 0\} \cdot \frac{4\exp(x)}{1+\exp(x)} + \mathbf{1}\{x > 0\} \cdot (2 + x)$ *and* $\hat{\varphi}(x) = \mathbf{1}\{x \geq -2\} \cdot \frac{x}{4} - \mathbf{1}\{x < -2\} \cdot \frac{1}{1+\exp(x+2)}$*. One easily verifies that the conditions in (86) are indeed satisfied. By Theorem 5 we know there exists a variational representation* (φ, ξ) *for the sigmoid function, although its explicit form is too cumbersome to write down. This variational factorization is very different from the one in Example 2, and perhaps is harder to come up with through guessing. From the point of view of optimization, the current factorization is inferior, despite of its theoretical value.*

Lemma 5 cannot be applied directly to the Geman-McClure function due to the lack of monotonicity on the entire real line. However, such functions have the symmetry property that we can exploit. The idea is to restrict $\rho(x)$ to $\mathbb{R}_+$ where monotonicity is guaranteed, then apply Lemma 5 we get ξ on $\mathbb{R}_+$ (and $\hat{\varphi}$). Finally extend ξ to $\mathbb{R}_-$ by symmetry.

Example 6 (The Geman-McClure function revisited)**.** *Following the above recipe,* $\mathbb{I} = (0, \infty)$*,* $a = 0$ *and we leave* b *unspecified (as it only affects constants). Simple calculation confirms* $\mathrm{d}(\log \rho_r) = \frac{1-3x^2}{x(1+x^2)}\mathrm{d}x$*. Therefore* $\mathrm{d}\nu_+ = \mathbf{1}\{0 \le x \le 1/\sqrt{3}\} \cdot \frac{1-3x^2}{x(1+x^2)}\mathrm{d}x$*. Direct integration in (98) gives* $\xi(x) = \mathbf{1}\{0 \le x \le 1/\sqrt{3}\} \cdot \frac{x^2}{1+x^2} + \mathbf{1}\{x > 1/\sqrt{3}\} \cdot (x + 1/4 - 1/\sqrt{3})$ *and* $\hat{\varphi}(x) = \mathbf{1}\{x \ge -1/4\} \cdot x - \mathbf{1}\{x < -1/4\} \cdot \frac{(x+1/4-1/\sqrt{3})^2}{1+(x+1/4-1/\sqrt{3})^2}$*. For* $x < 0$*, we set* $\xi(x) = \xi(-x)$ *while leave* $\hat{\varphi}$ *untouched. One again verifies that the conditions in (86) are satisfied. By Theorem 5 we know there exists a variational representation* (φ, ξ) *for the sigmoid function, although its explicit form is again too cumbersome to write down. This variational factorization is very different from the one in Example 1, and is inferior from the point of view of optimization.*

The above two examples might lead one to conclude that the minimal factorization given by Lemma 5 is of little use from the viewpoint of optimization, which is our main motivation of developing the variational representation. However, as we point out below, the true value of Lemma 5 is to ensure one when a particular loss function does admit a variational factorization. Note that once the existence issue is solved, finding a meaningful factorization can usually be done through a few trial-and-error. We also notice that without the help of Lemma 5, it seems hard to envision the following theorem merely from the conclusion we drew in Theorem 5.

Theorem 6. *All bounded strictly increasing* C^2 *functions admit a variational factorization that satisfies (86), provided* $\xi(x)$ *defined below in (102) is finite valued and* $\frac{\rho'(x)}{\xi'(x)}$ *is bounded from above.*

Proof. Let ρ be bounded and strictly increasing. (98) now simplifies to

$$\xi(x) \quad = \quad \int_a^x \exp\left(\int_b^y \frac{\max\{\rho''(z), 0\}}{\rho'(z)}\mathrm{d}z\right)\,\mathrm{d}y. \tag{102}$$

We need only verify (86). Notice that $\hat{\varphi}$ in (98) is only determined in $\{-\xi(x) : x \in \mathrm{dom}\rho\}$, but we can always complete it using the trick presented in Example 1 and 2.

Suppose $\hat{\varphi}(x)$ is defined for all sufficiently small x before we do the "completion", then

$$\limsup_{x\to-\infty} \frac{\hat{\varphi}(x)}{x} \quad = \quad \limsup_{x\to-\infty} \frac{-\rho(\xi^{-1}(-x))}{x} \quad = \quad 0, \tag{103}$$

since ρ is bounded. The other condition in (86) can be similarly argued. On the other hand, if $\hat{\varphi}(x)$ (for sufficiently small x) is only defined after we do the "completion", then we need to verify that $[-\rho(\xi^{-1}(-x))]' = \frac{\rho'(\xi^{-1}(-x))}{\xi'(\xi^{-1}(-x))} = \frac{\rho'(y)}{\xi'(y)}$ is between 0 and some positive number c (scaling $\hat{\varphi}$ and ξ appropriately will make $c = 1$). This holds because both ρ and ξ are strictly increasing, ρ' is bounded from above (for ρ is bounded), plus our assumption. ■

While it is possible to pin down a sufficient condition to remove the assumption in Theorem 6, we prefer not to do so, simply because this assumption itself is easily checkable. Apparently, Theorem 6 can be slightly modified to accommodate symmetric (and strictly increasing on, say, $\mathbb{R}_+$) loss functions. Refer to Examples 5 and 6.

Remark 4. *It is clear that we may replace the strictly* decreasing *assumption on the loss* ρ *with strictly* increasing *in both Lemma 5 and Theorem 6 (under obvious modifications), although such increasing functions seem to be uninteresting for loss minimization.*

In summary, we wonder what kind of functions can be factorized into the composition of a concave function and a convex function, where the concave part satisfies some condition like (86)? Of course, we would also like a principled method to find such factorizations, hopefully useful ones in the sense of optimization. Our results in Lemma 5 and Theorem 6 provide partial solutions to the above request, but seem to rely heavily on the monotonicity assumption. It would be nice to replace this assumption with something lighter.

E Computation

E.1 Efficient Implementation of the Approximate Estimator

Recall that the approximate estimator is computed in three steps: First, the relaxed lower bound (23) is computed, recovering $\boldsymbol{\eta}$ (see Lemma 2 in Appendix A). Second, the parameters $\boldsymbol{\alpha}$ are recovered by fixing $\boldsymbol{\eta}$ and re-solving for $\boldsymbol{\alpha}$ in (18) (see Lemma 1 in Appendix A). Third, $\boldsymbol{\eta}$ and $\boldsymbol{\alpha}$ are locally reoptimized in (18) using the previous $(\boldsymbol{\eta}, \boldsymbol{\alpha})$ as the initial point. We discuss an efficient implementation strategy, and the conditions under which polynomial run-time can be ensured.

For clarity we consider the case $\psi(\eta) = 1 - \eta$. (An efficient algorithm is possible for general ψ, but the details unnecessarily complicate the presentation.)

Relaxation: The first step of the approximate estimator is to compute the lower bound (23) given in Lemma 2, and recover the optimal $\boldsymbol{\eta}$ and M. This optimization can be solved efficiently as follows. Recall the definition $M_{\boldsymbol{\eta}} = \{M \succcurlyeq 0, M\mathbf{1} = \boldsymbol{\eta}, \mathrm{tr}(M) \leq 1\}$, apply the definition $\psi(\eta) = 1 - \eta$, and observe

$$
\begin{aligned}
(23) &= \max_{\boldsymbol{\nu}} \min_{0 \leq \boldsymbol{\eta} \leq 1} \min_{M \in \mathcal{M}_{\boldsymbol{\eta}}} n - \mathbf{1}^T\boldsymbol{\eta} - \boldsymbol{\eta}^T(\boldsymbol{\ell}^*(\boldsymbol{\nu}) - \Delta(\mathbf{y})\boldsymbol{\nu}) - \frac{1}{2\lambda}\boldsymbol{\nu}^T(K \circ M)\boldsymbol{\nu} && (104)\\
&= \max_{\boldsymbol{\nu}} \max_{\mathbf{a} \geq 0, \mathbf{b} \geq 0} \min_{\boldsymbol{\eta}} \min_{M \in \mathcal{M}_{\boldsymbol{\eta}}} n - \mathbf{1}^T\boldsymbol{\eta} - \boldsymbol{\eta}^T(\boldsymbol{\ell}^*(\boldsymbol{\nu}) - \Delta(\mathbf{y})\boldsymbol{\nu}) \\
&\qquad - \frac{1}{2\lambda}\boldsymbol{\nu}^T(K \circ M)\boldsymbol{\nu} - \mathbf{a}^T M\mathbf{1} + \mathbf{b}^T(M\mathbf{1} - \mathbf{1}) && (105)\\
&= \max_{\boldsymbol{\nu}} \max_{\mathbf{a} \geq 0, \mathbf{b} \geq 0} \min_{M \succcurlyeq 0, \mathrm{tr}(M) \leq 1} n - \mathbf{1}^T M\mathbf{1} - \mathbf{1}^T M(\boldsymbol{\ell}^*(\boldsymbol{\nu}) - \Delta(\mathbf{y})\boldsymbol{\nu}) \\
&\qquad - \frac{1}{2\lambda}\boldsymbol{\nu}^T(K \circ M)\boldsymbol{\nu} - \mathbf{a}^T M\mathbf{1} + \mathbf{b}^T(M\mathbf{1} - \mathbf{1}) && (106)\\
&= \max_{\boldsymbol{\nu}, \mathbf{a} \geq 0, \mathbf{b} \geq 0} n - \mathbf{b}^T\mathbf{1} - \max_{M \succcurlyeq 0, \mathrm{tr}(M) \leq 1} \mathbf{1}^T M\mathbf{1} + \mathbf{1}^T M(\boldsymbol{\ell}^*(\boldsymbol{\nu}) - \Delta(\mathbf{y})\boldsymbol{\nu}) \\
&\qquad + \frac{1}{2\lambda}\boldsymbol{\nu}^T(K \circ M)\boldsymbol{\nu} + \mathbf{a}^T M\mathbf{1} - \mathbf{b}^T M\mathbf{1}, && (107)
\end{aligned}
$$

where (104) follows by Sion's minimax theorem [26, Cor.37.3.2]; (105) follows by Lagrange duality; (106) follows by substituting $M\mathbf{1} = \boldsymbol{\eta}$ to replace the constraint and eliminate $\boldsymbol{\eta}$; and (107) is simple regrouping. Therefore (107) can be solved as a nonsmooth maximization:

$$
\max_{\boldsymbol{\nu}, \mathbf{a} \geq 0, \mathbf{b} \geq 0} n - \mathbf{b}^T\mathbf{1} - f(\boldsymbol{\nu}, \mathbf{a}, \mathbf{b}), \tag{108}
$$

where

$$
\begin{aligned}
f(\boldsymbol{\nu}, \mathbf{a}, \mathbf{b}) &= \max_{M \succcurlyeq 0, \mathrm{tr}(M) \leq 1} \mathbf{1}^T M\mathbf{1} + \mathbf{1}^T M\mathbf{c} + \frac{1}{2\lambda}\boldsymbol{\nu}^T(K \circ M)\boldsymbol{\nu} + \mathbf{a}^T M\mathbf{1} - \mathbf{b}^T M\mathbf{1} && (109)\\
&= \frac{1}{2}\max_{M \succcurlyeq 0, \mathrm{tr}(M) \leq 1} \mathrm{tr}(MC(\boldsymbol{\nu}, \mathbf{a}, \mathbf{b})), && (110)
\end{aligned}
$$

such that

$$
\begin{aligned}
C(\boldsymbol{\nu}, \mathbf{a}, \mathbf{b}) &= 2\mathbf{1}\mathbf{1}^T + \mathbf{1}\mathbf{c}(\boldsymbol{\nu})^T + \mathbf{c}(\boldsymbol{\nu})\mathbf{1}^T + \frac{1}{\lambda}\left(K \circ \boldsymbol{\nu}\boldsymbol{\nu}^T\right) + \mathbf{a}\mathbf{1}^T + \mathbf{1}\mathbf{a}^T - \mathbf{b}\mathbf{1}^T - \mathbf{1}\mathbf{b}^T && (111)\\
\mathbf{c}(\boldsymbol{\nu}) &= \boldsymbol{\ell}^*(\boldsymbol{\nu}) - \Delta(\mathbf{y})\boldsymbol{\nu}. && (112)
\end{aligned}
$$

Note that in the maximization problem (108), $f(\boldsymbol{\nu}, \mathbf{a}, \mathbf{b})$ must be convex, since it is a pointwise maximum of linear functions [22, Ch.3]; hence (108) is a concave maximization problem. Each evaluation of $f(\boldsymbol{\nu}, \mathbf{a}, \mathbf{b})$ requires no more than $O(n^3)$ time, since (110) can be solved by computing the maximum eigenvector of $C(\boldsymbol{\nu}, \mathbf{a}, \mathbf{b})$ [31]. Moreover, a subgradient in $(\boldsymbol{\nu}, \mathbf{a}, \mathbf{b})$ can easily be recovered from the maximizer M, based on the fact that

$$
\partial f \quad \ni \quad \begin{bmatrix} \Delta(M\mathbf{1})(\boldsymbol{\ell}'(\boldsymbol{\nu}) - \mathbf{y}) + \frac{1}{\lambda}(K \circ M)\boldsymbol{\nu} \\ M\mathbf{1} \\ -M\mathbf{1} \end{bmatrix} \tag{113}
$$

by Danskin's theorem [32, Ch.6]. (Note that the maximizer M might not be unique, so f is non-smooth at points where this occurs.) Finally, at a solution $(\boldsymbol{\nu}, \mathbf{a}, \mathbf{b}; M)$ the weight vector $\boldsymbol{\eta}$ is recovered via $\boldsymbol{\eta} = M\mathbf{1}$.

Therefore, computing the relaxed solution requires solving a nonsmooth concave maximization over $3n$ variables, where each function evaluation (and subgradient) can be computed in $O(n^3)$ time. An ellipsoid algorithm can therefore be used to solve (108) in polynomial-time [21] [33].

Rounding: The rounding procedure involves a simple, smooth convex minimization of $\boldsymbol{\alpha}$ in (18), where $\boldsymbol{\eta}$ is fixed from the relaxation step. This problem can be solved in polynomial-time provided only that the base loss ℓ (assumed convex) is also self-concordant [20].

Reoptimization: Finally, in the reopimization step, both $\boldsymbol{\eta}$ and $\boldsymbol{\alpha}$ are jointly (and locally) optimized in (18), starting from the solution above. A local optimum can, once again, be recovered in polynomial-time adding only the assumption that ψ (in addition to being convex) is also self-concordant.

Therefore, the estimation procedure requires only polynomial-time under the stated assumptions.

E.2 Proofs of NP-hardness

Our goal is to prove that minimizing a bounded loss function is NP-hard in general. After establishing the preliminary definitions required to be sufficiently precise about the results, we first prove that a special case—*clipped-loss* minimization—is strongly NP-hard in Section E.2.1. Then the NP-hardness of bounded loss minimization can be easily established by a reduction from clipped-loss minimization in Section E.2.2.

E.2.1 Hardness of Clipped Loss Minimization

Definition 3 (Loss function)**.** *A function $\ell : \mathbb{R} \to \mathbb{R}$ is a* loss function *if (i) $\ell(r) \geq 0$ for all r; (ii) $\ell(0) = 0$; and (iii) $\ell(r)$ is nondecreasing in r away from $r = 0$; that is, $r_1 \leq r_2 \leq 0$ implies $\ell(r_1) \geq \ell(r_2)$, and $0 \leq r_1 \leq r_2$ implies $\ell(r_1) \leq \ell(r_2)$.*

Definition 4 (τ-minimal two-sided loss)**.** *A loss function ℓ is a* τ-minimal two-sided loss *if there exists a finite $B_\tau > 0$ such that (i) $r \leq B_\tau$ implies $\ell(r) \geq \tau$; and (ii) $r \geq B_\tau$ implies $\ell(r) \geq \tau$.*

Definition 5 (β-bounded loss)**.** *A loss function ρ is a* β-bounded loss *if $\rho(y - \hat{y}) \leq \beta$ for all y and $\hat{y}$.*

Definition 6 (weakly τ-minimal loss)**.** *A loss function ℓ is a* weakly τ-minimal loss *if it is a $(\tau - \epsilon)$-minimal loss for all $\epsilon > 0$.*

Definition 7 (Clipped loss minimization)**.**
Instance: *$t \times n$ matrix X of training data, $t \times 1$ vector $\mathbf{y}$ of training labels, 1-minimal two-sided loss function ℓ, nonnegative threshold number c.*
Question: *Is there an $n \times 1$ vector $\boldsymbol{\theta}$ such that $\sum_{i=1}^{t} \min(1, \ell(y_i - X_{i:}\boldsymbol{\theta})) \leq c$?*

Definition 8 (Bounded loss minimization)**.**
Instance: *$t \times n$ matrix X of training data, $t \times 1$ vector $\mathbf{y}$ of training labels, 1-bounded and weakly 1-minimal two-sided loss function ℓ, nonnegative threshold number b.*
Question: *Is there an $n \times 1$ vector $\boldsymbol{\theta}$ such that $\sum_{i=1}^{t} \ell(y_i - X_{i:}\boldsymbol{\theta}) \leq b$?*

Definition 9 (Maximum 2-satisfiability)**.**
Instance: *Set U of variables, collection C of clauses over U such that each clause $c \in C$ has $|c| = 2$, positive integer $K \leq |C|$.*
Question: *Is there a truth assignment for U that simultaneously satisfies at least K of the clauses in C?*

Note that MAX2SAT is known to be NP-complete in general [34]. It is solvable in polynomial-time if $K = |C|$ [34], but NP-hard to approximate within a multiplicative constant better than 21/22 = 0.95454 [35].

Theorem 7. *Clipped loss minimization is strongly NP-hard.*

Proof. We transform MAX2SAT to clipped-loss minimization. Let (U, C, K) constitute an instance of MAX2SAT.

Let $\mathbf{1}_i$ denote a boolean vector with a single 1 in position i. We will also use a scale factor s that will be specified below (we will be able to choose a value for s that is polynomial in $|U|$).

Widget construction: For each variable in $u_j \in U$ associate a feature $X_{:j}$ with corresponding weight θ_j. For each clause $c \in C$ construct three training examples in the form $(\mathbf{x}^\top, y)$ as follows:

- For $u_i \vee u_j$ clauses, construct three examples $(s\mathbf{1}_i^\top, s)$, $(s\mathbf{1}_j^\top, s)$ and $(s(\mathbf{1}_i + \mathbf{1}_j)^\top, s)$.
- For $u_i \vee \neg u_j$ clauses, construct three examples $(s\mathbf{1}_i^\top, s)$, $(s\mathbf{1}_j^\top, 0)$ and $(s(\mathbf{1}_i - \mathbf{1}_j)^\top, 0)$.
- For $\neg u_i \vee u_j$ clauses, construct three examples $(s\mathbf{1}_i^\top, 0)$, $(s\mathbf{1}_j^\top, s)$ and $(s(\mathbf{1}_i - \mathbf{1}_j)^\top, 0)$.
- For $\neg u_i \vee \neg u_j$ clauses, construct three examples $(s\mathbf{1}_i^\top, 0)$, $(s\mathbf{1}_j^\top, 0)$ and $(s(\mathbf{1}_i + \mathbf{1}_j)^\top, s)$.

To illustrate how this construction will work, consider a training example $(s\mathbf{1}_i^\top, s)$. For a weight vector $\boldsymbol{\theta}$ one will obtain a prediction $\hat{y} = s\mathbf{1}_i^\top \boldsymbol{\theta} = s\theta_i$, which is compared to the target value $y = s$.

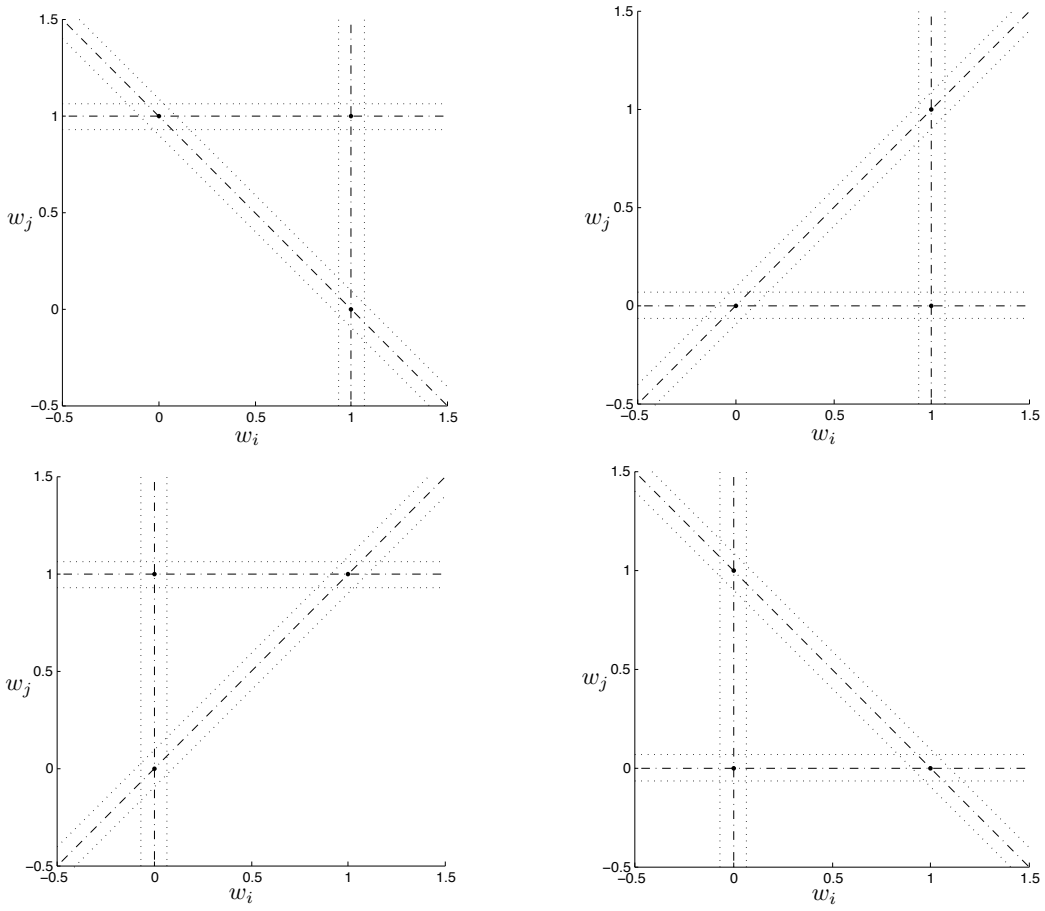

Figure 1: Depiction of the error surface in (θ_i, θ_j) for the three training examples constructed for each type of clause (when $i \neq j$): a $u_i \vee u_j$ clause, a $u_i \vee \neg u_j$ clause, a $\neg u_i \vee u_j$ clause, and a $\neg u_i \vee \neg u_j$ clause. (The trough widths are controlled by the scale factor s.) Note that the minimum total loss of 1 can only be achieved by setting (θ_i, θ_j) to a satisfying assignment in each of the four cases.

Note that if $\theta_i = 1$, then the loss must be zero on this example, since $\hat{y} = s = y$. However, as θ_i moves away from 1 the loss must increase until it "saturates". Let

$$\rho(y - \hat{y}) \quad = \quad \min(1, \ell(y - \hat{y})) \tag{114}$$

denote the "clipped" loss. Since ℓ is a 1-minimal two-sided loss there must exist a finite $B_1 > 0$ such that for any y, $\hat{y} \leq y - B_1$ implies $\rho(y - \hat{y}) = 1$ and $\hat{y} \geq y + B_1$ implies $\rho(y - \hat{y}) = 1$. Conversely, if $\rho(y - \hat{y}) < 1$ then it must follow that $y - B_1 < \hat{y} < y + B_1$. The role of the scale factor s, therefore, is to control the width of the "trough" where the losses remain strictly less than 1: Note that for any scale factor $s > 0$, if $\rho(sy - s\hat{y}) < 1$ then $sy - B_1 < s\hat{y} < sy + B_1$, which holds if and only if $y - B_1/s < \hat{y} < y + B_1/s$. Thus, the larger the choice of s, the narrower the trough where losses less than 1 can be achieved for a given training example.

Figure E.2.1 depicts the error surfaces created by the set of three training examples for each of the four clause types (for sufficiently large s) when the variables are distinct, $i \neq j$. Note that the minimum loss any weight vector can achieve on a clause (i.e., on its associated set of three training examples) is always 1, and this can only be achieved by assigning boolean weights that "satisfy" the clause.

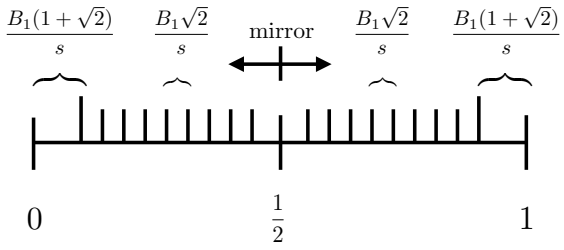

Figure 2: Illustrating bins of width $\frac{B_1\sqrt{2}}{s}$ between $\frac{B_1(1+\sqrt{2})}{s}$ and $\frac{1}{2}$, and their mirror images between $\frac{1}{2}$ and $1 - \frac{B_1(1+\sqrt{2})}{s}$.

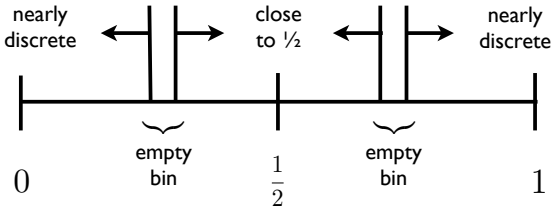

Figure 3: Illustrating how an empty (bin, mirror bin) pair can be used to define which component weight values are "nearly discrete" versus "close to 1/2".

Choosing the scale factor s: The scale factor s needs to be fixed to a sufficiently large value so that, for any weight vector $\boldsymbol{\theta}$, there exists a "gap" between 0 and $\frac{1}{2}$ (and a corresponding mirror gap between $\frac{1}{2}$ and 1) that contains no individual weight component θ_i. In particular, consider a partition of the interval $[\frac{B_1(1+\sqrt{2})}{s}, \frac{1}{2}]$ into disjoint bins of size $\frac{B_1\sqrt{2}}{s}$. There are at least $\lfloor\left(\frac{1}{2} - \frac{B_1(1+\sqrt{2})}{s}\right)/\frac{B_1\sqrt{2}}{s}\rfloor = \lfloor\frac{s}{B_1 2\sqrt{2}} - \frac{1}{\sqrt{2}} - 1\rfloor$ such bins that fit entirely within the interval.[15] So, by setting

$$s \;=\; B_1 2\sqrt{2}(|U|+3) \tag{115}$$

it follows that there must be at least $|U|+1$ disjoint bins of width $\frac{B_1\sqrt{2}}{s}$ within the interval $[\frac{B_1(1+\sqrt{2})}{s}, \frac{1}{2}]$, plus a mirror set of $|U|+1$ disjoint bins within the interval $[\frac{1}{2}, 1 - \frac{B_1(1+\sqrt{2})}{s}]$; see Figure 2. Given that the weight vector only has $|U|$ components, by construction, we know that for any particular $\boldsymbol{\theta}$ there must be at least one (bin, mirror bin) pair such that neither bin contains any θ_i value; see Figure 3. We will use these empty bins to define which θ_i values are considered to be "nearly discrete" versus "close to 1/2", as shown in Figure 3.

Main argument: We now proceed to show that there is always a boolean weight vector that is a global minimizer of the total clipped loss. Let

$$\rho(\boldsymbol{\theta}) \;=\; \sum_{i=1}^{t} \rho(y_i - X_{i:}\boldsymbol{\theta}) \tag{116}$$

denote the total clipped loss obtained by a weight vector $\boldsymbol{\theta}$ on the constructed set of training examples. Let $\boldsymbol{\theta}^{(0)}$ be a global minimizer of (116). Then we know that there must be some (bin, mirror bin) pair that contains no component of $\boldsymbol{\theta}^{(0)}$, which we can use to define "nearly discrete" versus "close to 1/2" weight values (Figure 3). Then by Lemma 6 below we know there must be some

[15] The reason for considering the interval $[\frac{B_1(1+\sqrt{2})}{s}, \frac{1}{2}]$ instead of $[0, \frac{1}{2}]$ is that we also need to ensure the empty bin does not intersect the parallelogram at the interesction between an axis-parallel and diagonal trough. For a given s this will be true for any bin that is at least $\frac{B_1(1+\sqrt{2})}{s}$ away from 0 or 1.

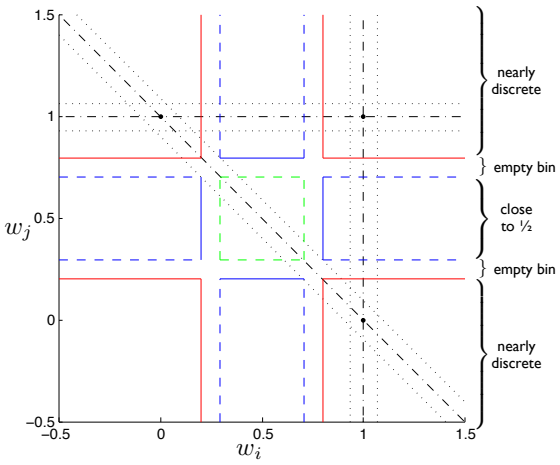

Figure 4: Illustrating the rounding scheme of Lemma 6 for a $u_i \vee u_j$ clause. (The scheme for the other three clause types depicted in Figure E.2.1 is isomorphic. Weight pairs (θ_i, θ_j) in the four *red* corner quadrants are rounded to the boolean corners. Weight pairs (θ_i, θ_j) in the four *blue* semi-rectangles on the edges are rounded to the line where the "nearly discrete" value becomes boolean. Weight pairs (θ_i, θ_j) in the *green* central square are not rounded. No weight pair occurs in the cross-hatch formed by the empty bins.

weight vector $\boldsymbol{\theta}^{(1)}$ which contains only boolean or "close to $1/2$" values such that $\rho(\boldsymbol{\theta}^{(1)}) \leq \rho(\boldsymbol{\theta}^{(0)})$. Then by Lemma 7 below we know there must exist a pure boolean weight vector $\boldsymbol{\theta}^{(2)}$ such that $\rho(\boldsymbol{\theta}^{(2)}) \leq \rho(\boldsymbol{\theta}^{(1)})$. Finally given a boolean assignment $\boldsymbol{\theta}^{(2)}$, a corresponding truth assignment $\mathbf{u}^{(2)}$ can be directly recovered via the transformation

$$\begin{aligned} u_i = \text{true} \quad &\Leftrightarrow \quad \theta_i = 1 \\ u_i = \text{false} \quad &\Leftrightarrow \quad \theta_i = 0. \end{aligned} \tag{117}$$

Furthermore, for a boolean weight vector $\boldsymbol{\theta}^{(2)}$ we have that

$$\begin{aligned} \rho(\boldsymbol{\theta}^{(2)}) &= 3 \times |\{\text{clauses falsified by } \mathbf{u}^{(2)}\}| \\ &\quad + 1 \times |\{\text{clauses satisfied by } \mathbf{u}^{(2)}\}| \qquad (118) \\ &= 3|C| - 2 \times |\{\text{clauses satisfied by } \mathbf{u}^{(2)}\}|. \qquad (119) \end{aligned}$$

It then follows that $|\{\text{clauses satisfied by } \mathbf{u}^{(2)}\}| \geq K$ if and only if $\rho(\boldsymbol{\theta}^{(2)}) \leq 3|C| - 2K$. That is, the minimum clipped loss achieved can be used to decide whether K clauses can be satisfied in the original instance. ■

Lemma 6. *For any weight vector $\boldsymbol{\theta}$ there is always a weight vector $\tilde{\boldsymbol{\theta}}$ with only boolean components or components that are "close to $1/2$" such that $\rho(\tilde{\boldsymbol{\theta}}) \leq \rho(\boldsymbol{\theta})$.*

Proof. Fix $\boldsymbol{\theta}$ and use one of its empty bins to define which component values are "nearly discrete" versus "close to $1/2$". Let $\tilde{\boldsymbol{\theta}}$ be the vector obtained by rounding all "nearly discrete" values of $\boldsymbol{\theta}$ to their nearest boolean value. We show that the total clipped loss obtained by $\tilde{\boldsymbol{\theta}}$ cannot be greater than that of $\boldsymbol{\theta}$.

Consider Figure 4, which depicts the effect of rounding for a clause of the $u_i \vee u_j$ type (the other three cases are isomorphic). For any clause, there are two situations to consider: when both weights are rounded and when only one weight is rounded.

First, consider the case where both weights (θ_i, θ_j) are "nearly discrete" and hence both are rounded to their nearest boolean value. This corresponds to being in one of the *red* corner quadrants shown in Figure 4. By inspection, one can see that the boolean assignment within each quadrant achieves the minimum loss attainable within the quadrant. In particular, in the three "satisfying" quadrants

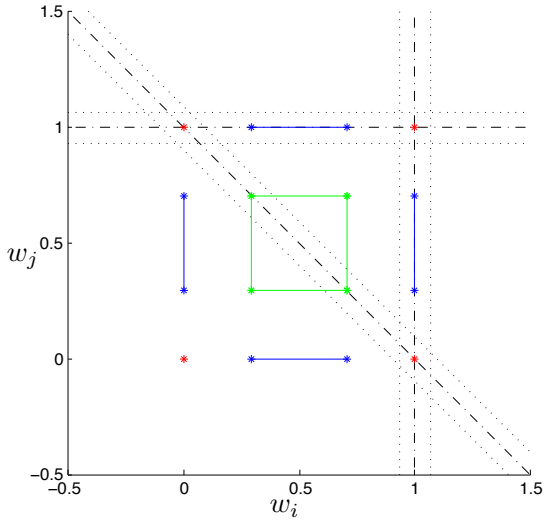

Figure 5: Possible weight pairs that can occur after rounding.

the boolean assignment is the strict minimizer of the loss, whereas in the "falsifying" quadrant all assignments achieve the same loss of 3. Therefore, the total loss cannot have increased on such a clause due to rounding.

Second, consider the case where one of the weights is "nearly discrete" and hence rounded, while the other is "close to $1/2$" and hence not rounded. This corresponds to being in one of the *blue* semi-rectangles on the edges shown in Figure 4. Once again, by inspection one can see that rounding the "nearly discrete" weight to its nearest boolean value cannot increase the loss obtained. In particular, rounding the "nearly discrete" variable yields the minimum loss assignment in the top and right semi-rectangles, while every assignment in the bottom and left semi-rectangles obtains the same loss of 3. This latter property is precisely what is achieved by the bin definitions: the empty bin width of $\frac{B_1\sqrt{2}}{s}$ is sufficient to ensure that the *blue* semi-rectangles do not intersect the diagonal trough. Therefore, the total loss cannot have increased on such a clause due to rounding.

In clauses where both weights are "close to $1/2$" (the *green* square in the middle) neither weight is rounded, hence the loss does not change. The result follows. ■

Lemma 7. *For any weight vector $\boldsymbol{\theta}$ with only boolean components or components that are "close to $1/2$" there is always a boolean weight vector $\hat{\boldsymbol{\theta}}$ such that $\rho(\hat{\boldsymbol{\theta}}) \leq \rho(\boldsymbol{\theta})$.*

Proof. Consider Figure 5, which depicts the possible weight pairs (θ_i, θ_j) participating in a clause of type $u_i \vee u_j$. (The situation for any of the clause types is isomorphic.) The isolated *red* points are the only purely boolean values, the *blue* line segments indicate pairs where one value is boolean, and the *green* square shows the region where both weights are "close to $1/2$". Observe that any weight pair with one or more "close to $1/2$" values (i.e., the union of the *blue* lines and the *green* square) has loss at least 2.

Given the weight vector $\boldsymbol{\theta}$, patition the clauses into three subsets: the set of clauses where both weight values are boolean ("closed clauses"), the set of clauses where one weight is boolean and the other is not ("mixed clauses"), and the set of clauses where both weights are "close to $1/2$" ("open clauses"). Let the set of "non-closed clauses" consist of the union of the "mixed clauses" and the "open clauses". From above, we know that $\boldsymbol{\theta}$ achieves a loss of at least 2 on each "non-closed clause". Let c denote the loss per clause achieved by $\boldsymbol{\theta}$ on the "closed clauses". There are two cases to consider.

Case 1: ($c \geq 2$) In this case the overall loss per clause achieved by $\boldsymbol{\theta}$ must be at least 2, which implies the total loss is $\rho(\boldsymbol{\theta}) \geq 2|C|$. Therefore, by Lemma 8 below we know that there must exist

some assignment $\hat{\mathbf{u}}$ to the variables in U that satisfies at least $\frac{|C|}{2}$ of the clauses in C. Let $\hat{\boldsymbol{\theta}}$ denote the corresponding boolean weight vector, recovered via the translation (117). By (119) we know that

$$\begin{aligned} \rho(\hat{\boldsymbol{\theta}}) &= 3|C| - 2 \times |\{\text{clauses satisfied by } \mathbf{u}^{(2)}\}| && (120) \\ &\leq 2|C| && (121) \\ &\leq \rho(\boldsymbol{\theta}). && (122) \end{aligned}$$

Case 2: ($c < 2$) Let B denote the indices where $\boldsymbol{\theta}$ is boolean, and let N denote the indices where $\boldsymbol{\theta}$ is "close to $1/2$". (Since $c < 2$ there must be at least one component θ_i that is boolean.) Fix the boolean components of $\boldsymbol{\theta}_B$, which in turn fixes the outcomes on the "closed clauses". For any "mixed clause" let θ_i denote the weight that is boolean. We would like to preserve the assignment of θ_i. Therefore, temporarily replace such a mixed clause with a new singleton clause defined by substituting u_i with its partner u_j. Consider the new set of altered "mixed clauses", in union with the set of "open clauses". Denote this new set $\tilde{C}$. Note that $\tilde{C}$ is defined only on the subset of variables U_N corresponding to weights in $\boldsymbol{\theta}$ that are "close to $1/2$". By Lemma 8 below, there must be some assignment $\tilde{\mathbf{u}}_N$ that satisfies at least half of the clauses in $\tilde{C}$, hence there exists a corresponding boolean weight vector $\tilde{\boldsymbol{\theta}}_N$ that achieves a loss of at most $2|\tilde{C}|$ on $\tilde{C}$.

Now consider the boolean weight vector $\hat{\boldsymbol{\theta}}$ formed by conjoining $\boldsymbol{\theta}_B$ with $\tilde{\boldsymbol{\theta}}_N$. Note that for any "closed clause" $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ behave identically and hence achieve the same loss. However, recall $\boldsymbol{\theta}$ achieves a loss of at least 2 on each "non-closed clause", while by the above construction, $\hat{\boldsymbol{\theta}}$ achieves a loss of at most 2 per clause over the "non-closed clauses". (In particular, even though $\hat{\mathbf{u}}_N$ was constructed by satisfying temporarily altered "mixed clauses" above, if it satisfies such a clause, then it must also satisfy the original "mixed clause".) Hence $\rho(\hat{\boldsymbol{\theta}}) \leq \rho(\boldsymbol{\theta})$.

Since in each of the two cases we were able to identify a boolean weight vector $\hat{\boldsymbol{\theta}}$ that achieves a loss no worse than $\boldsymbol{\theta}$, the result must follow. ■

Lemma 8. *For any MAX2SAT instance (U, C, K) and any assignment $\mathbf{u}$ to its variables U, either $\mathbf{u}$ or its negation $\neg\mathbf{u}$ must satisfy at least $\frac{|C|}{2}$ of the clauses in C.*

Proof. Consider any assignment $\mathbf{u}$. Note that $\neg\mathbf{u}$ must satisfy each clause that $\mathbf{u}$ falsifies. Therefore, if $\mathbf{u}$ satisfies fewer than $\frac{|C|}{2}$ clauses, it must falsify at least $\frac{|C|}{2}$ of the clauses, hence $\neg\mathbf{u}$ would have to satisfy at least $\frac{|C|}{2}$ of the clauses. Otherwise, $\mathbf{u}$ satisfies at least $\frac{|C|}{2}$ of the clauses. ■

E.2.2 Hardness of Bounded Loss Minimization

Finally, we are in a position to prove Theorem 2 from the main body. Recall the definition of bounded loss minimization (Definition 8).

Theorem 2. *Bounded (non-constant) loss minimization is NP-hard.*

Proof. We transform MAX2SAT to bounded-loss minimization. Let (U, C, K) constitute an instance of MAX2SAT. Use the same widget construction as in the proof of Theorem 7. Let

$$t = 3|C| \tag{123}$$

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{t} \ell(y_i - X_{i:}\boldsymbol{\theta}) \tag{124}$$

$$\ell^* = \min_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) \tag{125}$$

on the constructed training set.

To prove that minimizing the bounded loss ℓ can be used to decide whether K clauses can be satisfied, first consider an intermediate clipped-loss version of the problem. In particular, let

$$\hat{\ell}(y - \hat{y}) = \frac{1}{1 - \frac{1}{2t}} \ell(y - \hat{y}) \tag{126}$$

$$\hat{\rho}(y - \hat{y}) = \min(1, \hat{\ell}(y - \hat{y})) \tag{127}$$

$$\hat{\rho}(\boldsymbol{\theta}) = \sum_{i=1}^{t} \hat{\rho}(y_i - X_{i:}\boldsymbol{\theta}) \tag{128}$$

$$\hat{\rho}^* = \min_{\boldsymbol{\theta}} \hat{\rho}(\boldsymbol{\theta}) \tag{129}$$

on the constructed training set. Note that $\hat{\ell}$ is a 1-minimal loss with $\hat{B}_1 = B_{1-\frac{1}{2t}}$ (see Definition 4). Therefore, by Theorem 7, $\hat{\rho}^*$ can be used to decide whether K clauses can be satisfied in the original MAX2SAT instance. Recall also that $\hat{\rho}^*$ must be integer valued in this case.

Lemma 9 below shows that $\hat{\rho}^* - \frac{1}{2} \leq \ell^* \leq \hat{\rho}^*$. It then follows that for any integer n, $\hat{\rho}^* \leq n$ if and only if $\ell^* < n + \frac{1}{2}$. To see why this must hold, note: ($\Leftarrow$) if $\ell^* < n + \frac{1}{2}$ then $\hat{\rho}^* \leq \ell^* + \frac{1}{2} < n + 1$, hence $\hat{\rho}^* \leq n$ since $\hat{\rho}^*$ is integer valued; and also ($\Rightarrow$) if $\ell^* \geq n + \frac{1}{2}$ then $\hat{\rho}^* \geq \ell^* \geq n + \frac{1}{2}$, hence $\hat{\rho}^* > n$ since $\hat{\rho}^*$ is integer valued. Therefore, ℓ^* can be used to decide what integer value $\hat{\rho}^*$ achieves, which in turn can be used to decide whether K clauses can be satisfied by Theorem 7. ■

Note that bounded-loss minimization is strongly NP-hard if $B_{1-\frac{1}{2t}}$ is polynomial in $t = 3|C|$.

Lemma 9. $\hat{\rho}^* - \frac{1}{2} \leq \ell^* \leq \hat{\rho}^*$.

Proof. First, note that for any $(y, \hat{y})$

$$\hat{\rho}(y - \hat{y}) - \tfrac{1}{2t} \leq (1 - \tfrac{1}{2t})\hat{\rho}(y - \hat{y}) \tag{130}$$

$$= (1 - \tfrac{1}{2t}) \min(1, \hat{\ell}(y - \hat{y})) \tag{131}$$

$$= \min(1 - \tfrac{1}{2t}, \ell(y - \hat{y})) \tag{132}$$

$$\leq \ell(y - \hat{y}) \tag{133}$$

$$\leq \hat{\rho}(y - \hat{y}). \tag{134}$$

Therefore, for any $\boldsymbol{\theta}$

$$\hat{\rho}(\boldsymbol{\theta}) - \tfrac{1}{2} \leq \ell(\boldsymbol{\theta}) \leq \hat{\rho}(\boldsymbol{\theta}). \tag{135}$$

If we let $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$ and $\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \hat{\rho}(\boldsymbol{\theta})$, it is then immedate that

$$\hat{\rho}(\hat{\boldsymbol{\theta}}) - \tfrac{1}{2} \leq \hat{\rho}(\boldsymbol{\theta}^*) - \tfrac{1}{2} \tag{136}$$

$$\leq \ell(\boldsymbol{\theta}^*) \tag{137}$$

$$\leq \ell(\hat{\boldsymbol{\theta}}) \tag{138}$$

$$\leq \hat{\rho}(\hat{\boldsymbol{\theta}}). \tag{139}$$

■

Auxiliary References

[26] R. Rockafellar. *Convex Analysis*. Princeton U. Press, 1970.

[27] F. Cucker and D. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge, 2007.

[28] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2003.

[29] K. Sturm. Generalized Orlicz spaces and Wasserstein distances for convex-concave scale functions. *Bulletin des Sciences Mathématiques*, 135:795–802, 2011.

[30] W. Rudin. *Real and Complex Analysis*. McGraw-Hill, 3rd edition, 1987.

[31] M. Overton and R. Womersley. Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices. *Mathematical Programming*, 62(2):321–357, 1993.

[32] D. Bertsekas. *Nonlinear Programming*. Athena, 1995.

[33] M. Primak and B. Kheyfets. A modification of the inscribed ellipsoid method. *Mathematical and Computer Modelling*, 21(11):69–76, 1995.

[34] M. Garey and D. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, 1979.

[35] J. Håstad. Some optimal inapproximability results. *J. ACM*, 48(4):798–859, 2001.