Journal of Machine Learning Research 24 (2023) 1-79 Submitted 9/21; Revised 5/23; Published 5/23

On Tilted Losses in Machine Learning: Theory and Applications

Tian Li* TIANLI@CMU.EDU
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Ahmad Beirami*† BEIRAMI@GOOGLE.COM
Google Research
New York, NY 10011, USA

Maziar Sanjabi MAZIARS@FB.COM
Meta AI
Menlo Park, CA 94025, USA

Virginia Smith SMITHV@CMU.EDU
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Editor: Zaid Harchaoui

Abstract

Exponential tilting is a technique commonly used in fields such as statistics, probability, information theory, and optimization to create parametric distribution shifts. Despite its prevalence in related fields, tilting has not seen widespread use in machine learning. In this work, we aim to bridge this gap by exploring the use of tilting in risk minimization. We study a simple extension to ERM—tilted empirical risk minimization (TERM)—which uses exponential tilting to flexibly tune the impact of individual losses. The resulting framework has several useful properties: We show that TERM can increase or decrease the influence of outliers, respectively, to enable fairness or robustness; has variance-reduction properties that can benefit generalization; and can be viewed as a smooth approximation to the tail probability of losses. Our work makes connections between TERM and related objectives, such as Value-at-Risk, Conditional Value-at-Risk, and distributionally robust optimization (DRO). We develop batch and stochastic first-order optimization methods for solving TERM, provide convergence guarantees for the solvers, and show that the framework can be efficiently solved relative to common alternatives. Finally, we demonstrate that TERM can be used for a multitude of applications in machine learning, such as enforcing fairness between subgroups, mitigating the effect of outliers, and handling class imbalance. Despite the straightforward modification TERM makes to traditional ERM objectives, we find that the framework can consistently outperform ERM and deliver competitive performance with state-of-the-art, problem-specific approaches.

Keywords: Exponential tilting, empirical risk minimization, Value-at-Risk, superquantile optimization, fairness, robustness.

*Equal contribution.
†Work done at Meta AI.

1. Introduction

Many statistical estimation procedures rely on the concept of empirical risk minimization (ERM), in which the parameter of interest, $\theta \in \Theta \subseteq \mathbb{R}^d$, is estimated by minimizing an average loss over the data $\{x_1,...,x_N\}$:

$$\overline{R}(\theta) := \frac{1}{N} \sum_{i \in [N]} f(x_i; \theta)\,. \tag{1}$$

Although ERM is widely used in machine learning, it is known to perform poorly in situations where average performance is not an appropriate surrogate for the problem of interest. Significant research has thus been devoted to developing alternatives to traditional ERM for diverse applications, such as learning in the presence of noisy/corrupted data (Khetan et al., 2018; Jiang et al., 2018), performing classification with imbalanced data (Lin et al., 2017; Malisiewicz et al., 2011), ensuring that subgroups within a population are treated fairly (Hashimoto et al., 2018; Samadi et al., 2018), or developing solutions with favorable out-of-sample performance (Duchi and Namkoong, 2019).

In this paper, we suggest that deficiencies in ERM can be flexibly addressed via a unified framework, *tilted empirical risk minimization (TERM)*. TERM encompasses a family of objectives, parameterized by a real-valued hyperparameter, t. For $t \in \mathbb{R}^{\setminus 0}$, the t-tilted loss (TERM objective) is given by:

$$\widetilde{R}(t; \theta) := \frac{1}{t} \log \left(\frac{1}{N} \sum_{i \in [N]} e^{t f(x_i; \theta)} \right). \tag{2}$$

TERM generalizes ERM as the 0-tilted loss recovers the average loss, i.e., $\widetilde{R}(0, \theta) = \overline{R}(\theta)$.[1] It also recovers other popular alternatives such as the max-loss ($t \to +\infty$) and min-loss ($t \to -\infty$) (Lemma 4). As we discuss below, although tilted risk minimization is not widely used in machine learning, variants of tilting have been extensively studied in related fields including statistics, applied probability, optimization, and information theory.

1.1 Perspectives on Exponential Tilting

We begin by defining *exponential tilting* and discussing uses of tilting in various fields. Let $\mathcal{P} := \{p_\theta\}$ be a set of parametric distributions. For any $x \in \mathcal{X}$, we let $f(x; \theta)$ be the information of x under θ, which is defined as (Cover and Thomas, 1991):

$$f(x; \theta) := -\log p_\theta(x). \tag{3}$$

Further assume that X is a random variable drawn from distribution $p(\cdot)$, which is not necessarily matched to $\mathcal{P}$, i.e., the model family may be misspecified. The cumulant generating function of the information random variable, $f(X; \theta)$, can be stated as (Dembo and Zeitouni, 2009, Section 2.2):

$$\Lambda_X(t; \theta) := \log \left(\mathbb{E} \left[e^{t f(X; \theta)} \right] \right) = \log \sum_x p(x) p_\theta(x)^{-t}, \tag{4}$$

1. $\widetilde{R}(0; \theta)$ is defined in (20) via the continuous extension of $R(t; \theta)$.

where in this paper $\mathbb{E}[\cdot]$ denotes expectation with respect to the true distribution p unless otherwise stated. This expectation is commonly referred to as an *exponential tilt* of the information density, and can induce parametric distribution shifts that have varied applications in probability, statistics, and information theory. In particular, it is noteworthy that if $\mathcal{P}$ is an exponential family of distributions parameterized by θ, then the tilted distribution $p_\theta(x)^t$ (when normalized by $\int_{\mathcal{X}} p_\theta(x)^t dx$) also belongs to the same exponential family. Further, given samples $\{x_i\}_{i\in[N]}$, the empirical cumulant generating function is defined as:

$$\widetilde{\Lambda}(t;\theta) := \log\left(\frac{1}{N}\sum_{i\in[N]}\left\{e^{tf(x_i;\theta)}\right\}\right). \tag{5}$$

It is thus evident that TERM (2) can be viewed as an appropriately scaled variant of the empirical cumulant generating function in (5). Although tilting of this form has been used in a number of related disciplines, uses of exponential tilting in machine learning are relatively unexplored. We provide several perspectives on exponential tilting from other fields below.

Statistics. Exponential tilting is well-known as a distribution shifting technique in statistics, where the main idea is to draw samples from an exponentially tilted version of the original distribution to improve the convergence properties of statistical estimation, especially when the distribution of interest belongs to an exponential family, such as Gaussian or multinomial. Common use cases include rejection sampling, rare-event simulation, saddle-point approximation (Butler, 2007, p. 156), and importance sampling (Siegmund, 1976).

Applied probability. In large deviations theory, exponential tilting lies at the heart of deriving concentration bounds. For example, Chernoff bounds apply Markov's inequality to e^{tX}, which results in a parametric set of bounds by using exponential tilts of various orders. The bound may then be further optimized on the real tilt value to derive the tightest possible bound (Dembo and Zeitouni, 2009).

Information theory. While source coding limits and channel capacity are characterized by Shannon entropy and Shannon mutual information (which are simple averages over the information (3)) (Cover and Thomas, 1991), there are other elements of information theory that are not characterized by the average, such as error exponents in channel decoding (Gallager, 1968), probability of error in list decoding (Merhav, 2014), and computational cost in sequential decoding (Massey, 1994; Arıkan, 1996). These fundamental elements of information theory are asymptotically determined by a non-zero tilted cumulant generating function of the information random variable (3) (see (Beirami et al., 2018) for further discussion).

Optimization. Exponential tilting has also appeared as a minimax smoothing approach in optimization (Kort and Bertsekas, 1972; Pee and Royset, 2011; Liu and Theodorou, 2019). Such smooth approximations to the max often appear through LogSumExp functions, with applications in geometric programming (Calafiore and El Ghaoui, 2014, Sec. 9.7), and boosting (Mason et al., 1999; Shen and Li, 2010). We discuss min-max objectives and the connections with TERM in several subsequent sections of the paper.

Machine learning. Despite the rich history of tilted objectives in related fields, they have not seen widespread use in ML beyond limited applications such as robust regression (Wang

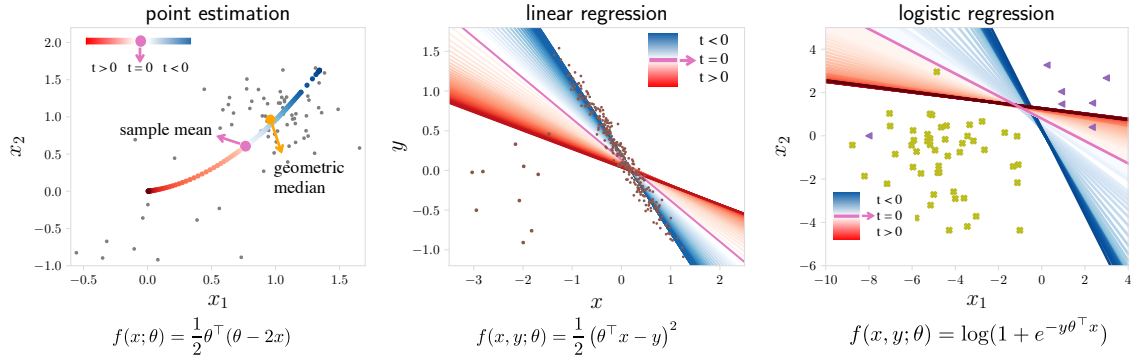


Figure 1: Toy examples illustrating TERM as a function of t: (a) finding a point estimate from a set of 2D samples, (b) linear regression with outliers, and (c) logistic regression with imbalanced classes. While positive values of t magnify outliers, negative values suppress them. Setting t=0 recovers the original ERM objective (1).

et al., 2013) and sequential decision making (Howard and Matheson, 1972; Borkar, 2002). In this work, we argue that tilting is a critical yet undervalued tool in machine learning. We demonstrate the effectiveness of tilting by (i) rigorously studying properties of the TERM objective, and (ii) exploring its utility for a wide range of ML applications. Surprisingly, we find that this simple extension to ERM can match or exceed state-of-the-art performance from highly tuned, bespoke solutions to common ML problems, from learning with noisy data to ensuring fair performance between subgroups. We highlight several motivating applications of TERM below and provide an outline of the remainder of the paper in Section 1.3.

1.2 Motivating Examples

To motivate how the TERM objective (2) may be used in machine learning, we provide several running examples below, which are illustrated in Figure 1.

(a) Point estimation: As a first example, consider determining a point estimate from a set of samples that contain some outliers. We plot an example 2D dataset in Figure 1a, with data centered at (1,1). Using traditional ERM (i.e., TERM with $t = 0$) recovers the *sample mean*, which can be biased towards outlier data. By setting $t < 0$, TERM can suppress outliers by reducing the relative impact of the largest losses (i.e., points that are far from the estimate) in (2). A specific value of $t < 0$ can in fact approximately recover the geometric median, as the objective in (2) can be viewed as approximately optimizing specific loss quantiles (a connection which we make explicit in Section 2). In contrast, if these 'outlier' points are important to estimate, setting $t > 0$ will push the solution towards a point that aims to minimize variance, as we prove in Section 2, Theorem 3.

(b) Linear regression: A similar interpretation holds for the case of linear regression (Figure 2b). As $t \to -\infty$, TERM finds a line of best while ignoring outliers. However, this solution may not be preferred if we have reason to believe that these 'outliers' should not be ignored. As $t \to +\infty$, TERM recovers the min-max solution, which aims to minimize the worst loss, thus ensuring the model is a reasonable fit for *all* samples (at the expense of possibly being a worse fit for many). Similar criteria have been used, e.g., in defining

notions of fairness (Hashimoto et al., 2018; Samadi et al., 2018). We explore several use-cases involving robust regression and fairness in more detail in Section 7.

(c) Logistic regression: Finally, we consider a binary classification problem using logistic regression (Figure 2c). For $t \in \mathbb{R}$, the TERM solution varies from the nearest cluster center ($t \to -\infty$), to the logistic regression classifier ($t=0$), towards a classifier that magnifies the misclassified data ($t \to +\infty$). We note that it is common to modify logistic regression classifiers by adjusting the decision threshold from 0.5, which is equivalent to moving the intercept of the decision boundary. This is fundamentally different than what is offered by TERM (where the slope is changing). As we show in Section 7, this added flexibility affords TERM with competitive performance on a number of classification problems, such as those involving noisy data, class imbalance, or a combination of the two.

1.3 Contributions

In this work, we explore the use of tilting in machine learning through TERM, a simple, unified framework that can flexibly address various challenges with empirical risk minimization. We first analyze the objective and its solutions, showcasing the behavior of TERM with varying tilt parameters t (Section 2). We also establish connections between TERM and related approaches such as distributionally robust optimization in Section 3.

We rigorously analyze the relations between TERM and other risks (e.g, Value-at-Risk (VaR), Conditional Value-at-Risk (CVaR), and Entropic Value-at-Risk (EVaR)) in Section 4. In particular, we introduce a new risk measure based on TERM, called Tilted Value-at-Risk (TiVaR), to approximate VaR. We show that TiVaR can provide a better approximation of VaR than CVaR in certain regimes, and improves upon EVaR in all regimes.

We develop efficient first-order batch and stochastic methods for solving TERM, both for hierarchical and non-hierarchical cases (Section 5 and 6). We provide convergence rates scaling with the hyperparameter t on both convex and non-convex problems for both batch and stochastic algorithms. Our solvers run within 2–3$\times$ wall-clock time compared with that of ERM in all explored case studies.

Finally, we show via numerous case studies that TERM is competitive with existing, problem-specific state-of-the-art solutions (Section 7). We also extend TERM to handle compound issues, such as the simultaneous existence of noisy samples and imbalanced classes (Section 6). Our results demonstrate the effectiveness and versatility of tilted objectives in machine learning.

We note that the material in this paper was presented in part at ICLR 2021 (Li and Beirami et al., 2021). Compared to this earlier work, the current manuscript provides additional historical background of tilting (Section 1), establishes stronger and novel relationships between tilted losses and other risk-averse objectives in the literature (Section 3 and Section 4), provides convergence guarantees for our stochastic solver of TERM (Section 5), offers comprehensive details on applications of the framework in practice, and considers new applications of TERM to meta-learning and heteroskedastic deep learning (Section 7).

Outline. This paper is organized as follows. We discuss general properties and interpretations of TERM in Section 2. We connect TERM with other prior risk measures in Section 3 and propose a new risk motivated by TERM in Section 4. In Section 5, we develop both batch and stochastic algorithms for optimizing TERM and provide convergence guarantees for them.

We extend TERM to hierarchical multi-objective tilting in Section 6 and demonstrate the flexibility and competitive performance of the TERM framework via real-world applications in Section 7. We discuss related work in Section 8 and conclude the paper with Section 9.

2. TERM: Properties and Interpretations

To better understand the performance of the t-tilted losses in (2), in this section we provide several interpretations of the TERM solutions, leaving the full proofs to the appendix. We make no distributional assumptions on the data, and study properties of TERM under the assumption that the loss function forms a generalized linear model, e.g., L_2 loss and logistic loss. However, we also obtain favorable empirical results using TERM with other objectives such as PCA and deep neural networks in Section 7, motivating the extension of this part of our theory beyond GLMs in future work.

2.1 Assumptions

We first provide notation and assumptions that are used throughout our theoretical analyses. The results in this paper are derived under one of the following three nested assumptions (the assumptions become progressively more restrictive, i.e., $3 \to 2 \to 1$):

Assumption 1 (Continuous differentiability)**.** *For $i \in [N]$, the loss function $f(x_i;\theta)$ belongs to the differentiability class C^1 (i.e., continuously differentiable) with respect to $\theta \in \Theta \subseteq \mathbb{R}^d$.*

Assumption 2 (Smoothness and strong convexity condition)**.** *Assume that Assumption 1 is satisfied. In addition, for any $i \in [N]$, $f(x_i;\theta)$ belongs to differentiability class C^2 (i.e., twice differentiable with continuous Hessian) with respect to θ. We further assume that there exist $\beta_{\min}, \beta_{\max} \in \mathbb{R}^{>0}$ such that for $i \in [N]$ and any $\theta \in \Theta \subseteq \mathbb{R}^d$,*

$$\beta_{\min}\mathbf{I} \preceq \nabla^2_{\theta\theta^\top} f(x_i;\theta) \preceq \beta_{\max}\mathbf{I}, \tag{6}$$

where $\mathbf{I}$ is the identity matrix of appropriate size (in this case $d \times d$), and there does **not** *exist any $\theta \in \Theta$, such that $\nabla_\theta f(x_i;\theta) = 0$ for all $i \in [N]$.*

Assumption 3 (Generalized linear model condition (Wainwright and Jordan, 2008))**.** *Assume that Assumption 2 is satisfied. Further, assume that the loss function $f(x;\theta)$ is given by*

$$f(x;\theta) = A(\theta) - \theta^\top T(x), \tag{7}$$

where $A(\cdot)$ is a convex function such that there exists $\beta_{\max}$ where for any $\theta \in \Theta \subseteq \mathbb{R}^d$,

$$\beta_{\min}\mathbf{I} \preceq \nabla^2_{\theta\theta^\top} A(\theta) \preceq \beta_{\max}\mathbf{I}\,, \tag{8}$$

and

$$\sum_{i\in[N]} T(x_i)T(x_i)^\top \succ 0. \tag{9}$$

This set of assumptions become the most restrictive with Assumption 3, which essentially requires that the loss be the negative log-likelihood of an exponential family. While the assumption is stated using the natural parameter of an exponential family for ease

of presentation, the results hold for any bijective and smooth reparameterization of the exponential family. For example, Assumption 3 is satisfied by the commonly used L_2 loss for regression and logistic loss for classification (see toy examples (b) and (c) in Figure 1). $\sum_{i\in[N]} T(x_i)T(x_i)^\top \succ 0$ assumes a reasonable regularity on the dataset $\{x_i\}_{i\in[N]}$. For instance, in the case of linear regression ($T(x_i) = x_i \in \mathbb{R}^d$), it reduces to the standard regularity assumption $XX^T \succ 0$ (where $X := [x_1, \cdots, x_N] \in \mathbb{R}^{d\times N}$). While Assumption 3 is not satisfied when we use neural network function approximators in Section 7, we observe favorable numerical results motivating the extension of these results beyond the cases that are theoretically studied in this paper.

In the sequel, many of the results are concerned with characterizing the t-tilted solutions defined as the parametric set of solutions of t-tiled losses by sweeping $t \in \mathbb{R}$,

$$\breve{\theta}(t) \in \arg\min_{\theta\in\Theta} \widetilde{R}(t;\theta), \tag{10}$$

where $\Theta \subseteq \mathbb{R}^d$ is an open subset of $\mathbb{R}^d$. Further, let the optimal tilted objective be defined as

$$\widetilde{F}(t) := \widetilde{R}(t;\breve{\theta}(t)). \tag{11}$$

We state a final assumption, on $\breve{\theta}(t)$, below.

Assumption 4 (Strict saddle property (Definition 4 in Ge et al. (2015))). *We assume that the set* $\arg\min_{\theta\in\Theta} \widetilde{R}(t;\theta)$ *is non-empty for all* $t \in \mathbb{R}$. *Further, we assume that for all* $t \in \mathbb{R}$, $\widetilde{R}(t;\theta)$ *is a "strict saddle" as a function of* θ, *i.e., for all local minima,* $\nabla^2_{\theta\theta^\top}\widetilde{R}(t;\theta)\succ 0$, *and for all other stationary solutions,* $\lambda_{\min}(\nabla^2_{\theta\theta^\top}\widetilde{R}(t;\theta)) < 0$, *where* $\lambda_{\min}(\cdot)$ *is the minimum eigenvalue of the matrix.*

We use the strict saddle property in order to reason about the properties of the t-tilted solutions. In particular, since we are solely interested in the local minima of $\widetilde{R}(t;\theta)$, the strict saddle property implies that for every $\breve{\theta}(t) \in \arg\min_{\theta\in\Theta} \widetilde{R}(t;\theta)$, for a sufficiently small r, for all $\theta \in \mathcal{B}(\breve{\theta}(t), r)$,

$$\nabla^2_{\theta\theta^\top}\widetilde{R}(t;\theta) \succ 0, \tag{12}$$

where $\mathcal{B}(\breve{\theta}(t), r)$ denotes a d-ball of radius r around $\breve{\theta}(t)$. We will show later in Section 2.2 that the strict saddle property is readily verified for $t \in \mathbb{R}^{>0}$ under Assumption 2, and we need Assumption 4 to be able to reason about $t \in \mathbb{R}^{<0}$.

2.2 General Properties of TERM

We begin by noting several general properties of the TERM objective (2). In particular: (i) $\widetilde{R}(t;\theta)$ is L-Lipschitz continuous in θ if $f(x;\theta)$ is L-Lipschitz (Lemma 1); (ii) If $f(x;\theta)$ is strongly convex, the t-tilted loss is strongly convex for $t > 0$ (Lemma 2); and (iii) Given a smooth $f(x;\theta)$, the t-tilted loss is smooth for all finite t (Lemma 3). We state these properties more formally below.

Lemma 1 (Lipschitzness of $\widetilde{R}(t;\theta)$). *For any* $t \in \mathbb{R}$ *and* $\theta \in \Theta$, *if for* $i \in [N]$, $f(x_i;\theta)$ *is* L*-Lipschitz conditnuous in* θ, *then* $\widetilde{R}(t;\theta)$ *is* L*-Lipschitz in* θ.

Lemma 2 (Tilted Hessian and strong convexity for $t \in \mathbb{R}^{>0}$)**.** *Under Assumption 2, for any* $t \in \mathbb{R}$,

$$\nabla^2_{\theta\theta^\top}\widetilde{R}(t;\theta) = \frac{t}{N}\sum_{i\in[N]}(\nabla_\theta f(x_i;\theta) - \nabla_\theta \widetilde{R}(t;\theta))(\nabla_\theta f(x_i;\theta) - \nabla_\theta \widetilde{R}(t;\theta))^\top e^{t(f(x_i;\theta)-\widetilde{R}(t;\theta))} \tag{13}$$

$$+ \frac{1}{N}\sum_{i\in[N]}\nabla^2_{\theta\theta^\top} f(x_i;\theta) e^{t(f(x_i;\theta)-\widetilde{R}(t;\theta))}. \tag{14}$$

In particular, for all $\theta \in \Theta$ *and all* $t \in \mathbb{R}^{>0}$*, the* t*-tilted objective is strongly convex. That is*

$$\nabla^2_{\theta\theta^\top}\widetilde{R}(t;\theta) \succ \beta_{\min}\mathbf{I}. \tag{15}$$

Lemma 1 and 2 are proved in Appendix A. Lemma 2 also implies that under Assumption 2, the strict saddle assumption (Assumption 4) is readily verified.

Lemma 3 (Smoothness of $\widetilde{R}(t;\theta)$)**.** *For any* $t \in \mathbb{R}$*, let* $\beta(t)$ *be the smoothness parameter of twice differentiable* $\widetilde{R}(t;\theta)$*:*

$$\beta(t) := \lambda_{\max}\left(\nabla^2_{\theta\theta^\top}\widetilde{R}(t;\theta)\right), \tag{16}$$

where $\nabla^2_{\theta\theta^\top}\widetilde{R}(t;\theta)$ *is the Hessian of* $\widetilde{R}(t;\theta)$ *at* θ *and* $\lambda_{\max}(\cdot)$ *denotes the largest eigenvalue. Under Assumption 2, for any* $t \in \mathbb{R}$*,* $\widetilde{R}(t;\theta)$ *is a* $\beta(t)$*-smooth function of* θ*. Further, for* $t \in \mathbb{R}^{\leqslant 0}$*,*[2]

$$\beta(t) < \beta_{\max}, \tag{17}$$

where $\beta_{\max}$ *is defined in Assumption 2. For* $t \in \mathbb{R}^{>0}$*,*

$$0 < \lim_{t\to+\infty}\frac{\beta(t)}{t} < +\infty. \tag{18}$$

Lemma 3 (proved in Appendix A.1) indicates that t-tilted losses are $\beta(t)$-smooth for all t. $\beta(t)$ is bounded for all negative t and moderately positive t, whereas it scales linearly with t as $t \to +\infty$, which has been previously studied in the context of exponential smoothing of the max (Kort and Bertsekas, 1972; Pee and Royset, 2011). This can also be observed visually via the toy example in Figure 2.

As discussed in Section 1, TERM can recover traditional ERM (t=0), the max-loss ($t\to+\infty$), and the min-loss ($t\to-\infty$). We formally state this in Lemma 4 below.

Lemma 4. *Under Assumption 1,*

$$\widetilde{R}(-\infty;\theta) := \lim_{t\to-\infty}\widetilde{R}(t;\theta) = \check{R}(\theta), \tag{19}$$

$$\widetilde{R}(0;\theta) := \lim_{t\to 0}\widetilde{R}(t;\theta) = \overline{R}(\theta), \tag{20}$$

$$\widetilde{R}(+\infty;\theta) := \lim_{t\to+\infty}\widetilde{R}(t;\theta) = \widehat{R}(\theta), \tag{21}$$

2. $\mathbb{R}^{\leqslant 0}$ denotes the set of non-positive real numbers.

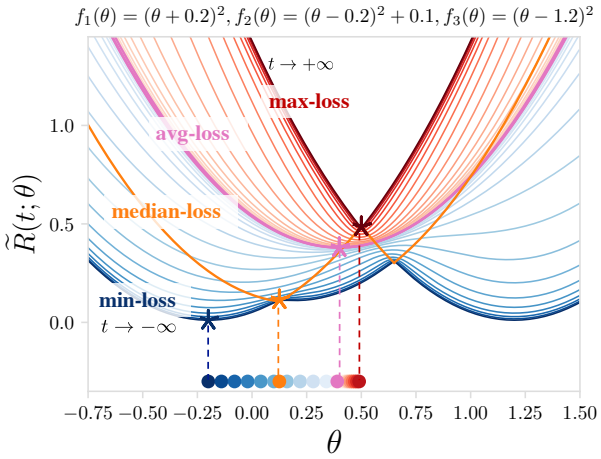


Figure 2: TERM objectives for a squared loss with three samples (N=3). As t moves from $-\infty$ to $+\infty$, t-tilted losses recover min-loss, avg-loss, and max-loss. TERM is smooth for all finite t and convex for positive t.

where $\widehat{R}(\theta)$ *is the max-loss and* $\check{R}(\theta)$ *is the min-loss*[3]*:*

$$\widehat{R}(\theta) := \max_{i \in [N]} f(x_i; \theta), \qquad \check{R}(\theta) := \min_{i \in [N]} f(x_i; \theta). \tag{22}$$

Note that Lemma 4 has been studied or observed before in the entropic risk literature (e.g., Ahmadi-Javid, 2012), as well as other contexts (Cohen and Shashua, 2014). This lemma also implies that $\check{\theta}(0)$ is the ERM solution, $\check{\theta}(+\infty)$ is the min-max solution, and $\check{\theta}(-\infty)$ is the min-min solution. In other words, a benefit of TERM is that it offers a continuum of solutions between the min and max losses.

Providing a smooth trade-off between these specific losses can be beneficial for a number of practical use-cases—both in terms of the resulting solution and the difficulty of solving the problem itself. We empirically demonstrate the benefits of such a trade-off in Section 7. We also visualize the solutions to TERM for a toy problem in Figure 2, which allows us to illustrate several special cases of the general framework. Interestingly, we additionally show that the TERM solution can be viewed as a smooth approximation to the *tail probability of losses*, which effectively minimizes quantiles of losses such as the median loss (Section 4). In Figure 2, it is clear to see why this may be beneficial, as the median loss (orange) can be highly non-smooth in practice. In Theorem 1 and 2 below, we formally characterize how tilted objectives change as a function of values t (proofs provided in Appendix A).

Theorem 1 (Tilted objective is increasing with t)**.** *Under Assumption 3, for all* $t \in \mathbb{R}$*, and all* $\theta \in \Theta$*,*

$$\frac{\partial}{\partial t} \widetilde{R}(t; \theta) \geqslant 0. \tag{23}$$

3. When the argument of the max-loss or the min-loss is not unique, for the purpose of differentiating the loss function, we define $\widehat{R}(\theta)$ as the average of the individual losses that achieve the maximum, and $\check{R}(\theta)$ as the average of the individual losses that achieve the minimum.

Theorem 2 (Optimal tilted objective is increasing with t). *Under Assumption 3, for all $t \in \mathbb{R}$, and all $\theta \in \Theta$,*

$$\frac{\partial}{\partial t}\widetilde{F}(t) = \frac{\partial}{\partial t}\widetilde{R}(t;\breve{\theta}(t)) \geqslant 0. \tag{24}$$

Recall that TERM as $t \to -\infty$ and $t \to \infty$ corresponds to min-loss and max-loss, respectively. We discuss in Section 4.2 that solving TERM with any $t \in \mathbb{R}$ can indeed be viewed as approximately minimizing the k-th smallest loss ($k \in [N]$) among all N individual losses. As we increase k from 1 to N, the corresponding value of t sweeps in $(-\infty, \infty)$. Theorem 2 hence roughly states that the optimal k-th smallest loss is non-decreasing with k, which is intuitively expected.

We next provide two interesting interpretations of the TERM framework to further understand its behavior.

2.3 Interpretation 1: Re-Weighting Samples to Magnify/Suppress Outliers

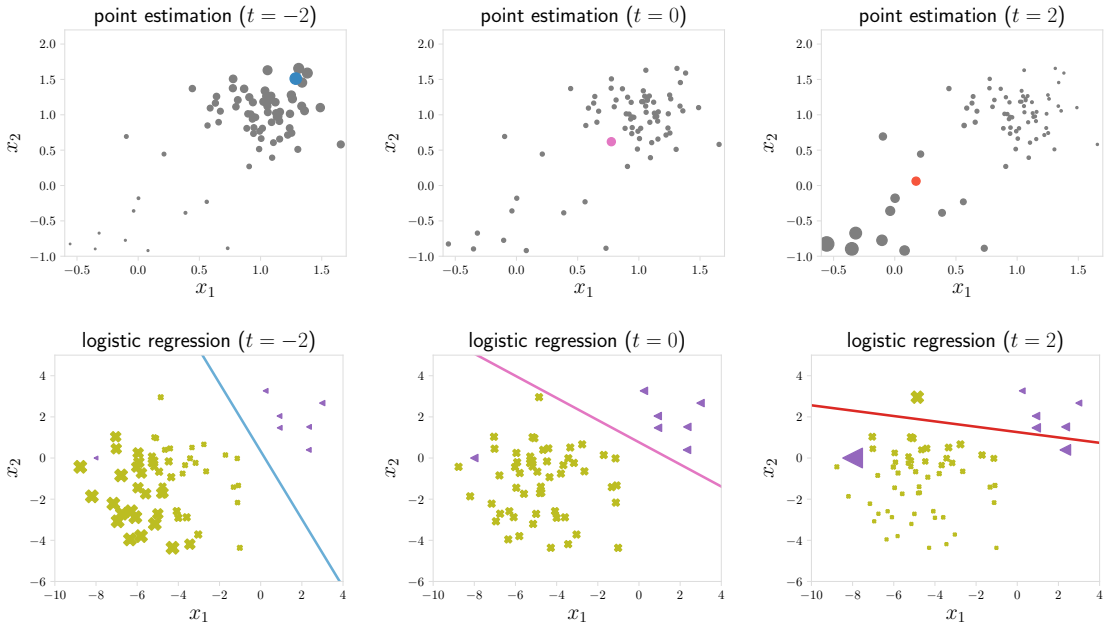


Figure 3: We visualize the size of the samples using their gradient weights. Negative t's ($t = -2$ on the left) focus on the inlier samples (suppressing outliers), while positive t's ($t = 2$ on the right) magnify the outlier samples.

As discussed via the toy examples in Section 1, TERM can be tuned (using t) to magnify or suppress the influence of outliers. We make this notion rigorous by exploring the *gradient* of the t-tilted loss in order to reason about the solutions to the objective defined in (2).

Lemma 5 (Tilted gradient). *For a smooth loss function $f(x;\theta)$,*

$$\nabla_\theta \widetilde{R}(t;\theta) = \sum_{i\in[N]} w_i(t;\theta)\nabla_\theta f(x_i;\theta), \tag{25}$$

where tilted weights are given by

$$w_i(t;\theta) := \frac{e^{tf(x_i;\theta)}}{\sum_{j\in[N]} e^{tf(x_j;\theta)}} = \frac{1}{N} e^{t(f(x_i;\theta)-\widetilde{R}(t;\theta))}. \tag{26}$$

Proof. Under Assumption 1, we have:

$$\nabla_\theta \widetilde{R}(t;\theta) = \nabla_\theta \left\{ \frac{1}{t} \log \left(\frac{1}{N} \sum_{i\in[N]} e^{tf(x_i;\theta)} \right) \right\} = \frac{\sum_{i\in[N]} \nabla_\theta f(x_i;\theta) e^{tf(x_i;\theta)}}{\sum_{i\in[N]} e^{tf(x_i;\theta)}}. \tag{27}$$

□

Lemma 5 provides the gradient of the tilted objective, which has been studied previously in the context of exponential smoothing (see Pee and Royset (2011, Proposition 2.1)). From this, we can observe that the tilted gradient is a weighted average of the gradients of the original individual losses, where each data point is weighted exponentially proportional to the value of its loss. Note that $t = 0$ recovers the uniform weighting associated with ERM, i.e., $w_i(t;\theta) = 1/N$. For positive t, this has the effect of *magnifying* the outliers—samples with large losses—by assigning more weight to them, and for negative t, it *suppresses* the outliers by assigning less weight to them (Figure 3).

Generalizing the notion of tilted gradients (weighted average of individual gradients), we define tilted empirical mean over any N-vector $\mathbf{u} \in \mathbf{R}^N$ below, which will be used throughout the paper.

Definition 1 (Tilted empirical mean and variance)**.** *For $\mathbf{u} \in \mathbf{R}^N$, let weighted empirical mean with weights $\mathbf{w} \in \Delta^N$ (where Δ^N stands for N dimensional simplex) be defined as*

$$\widehat{E}_{\mathbf{w}}(\mathbf{u}) := \sum_{i\in[N]} w_i u_i. \tag{28}$$

Tilted empirical mean is weighted empirical mean with tilted weights, i.e.,

$$\widehat{E}_{\mathbf{w}(t;\theta)}(\mathbf{u}) := \sum_{i\in[N]} w_i(t;\theta) u_i, \tag{29}$$

$$\widehat{E}_{\mathbf{w}(t;\breve{\theta}(t))}(\mathbf{u}) := \sum_{i\in[N]} w_i(t;\breve{\theta}(t)) u_i, \quad \widehat{E}_t := \widehat{E}_{\mathbf{w}(t;\breve{\theta}(t))}(\mathbf{u}), \tag{30}$$

where $w_i(t;\theta)$ is defined in Eq. (26), and $\breve{\theta}(t)$ is defined in Eq. (10). We also refer to $\widehat{E}_t$ as the "t-tilted empirical mean". Similarly, tilted empirical variance is defined as

$$\widehat{\mathrm{var}}_{\mathbf{w}(t;\theta)}(\mathbf{u}) := \widehat{E}_{\mathbf{w}(t;\theta(t))} \left(u_i - \widehat{E}_{\mathbf{w}(t;\theta(t))}(\mathbf{u}) \right)^2, \tag{31}$$

$$\widehat{\mathrm{var}}_{\mathbf{w}(t;\breve{\theta}(t))}(\mathbf{u}) := \widehat{E}_t (u_i - \widehat{E}_t(\mathbf{u}))^2, \quad \widehat{\mathrm{var}}_t := \widehat{\mathrm{var}}_{\mathbf{w}(t;\breve{\theta}(t))}(\mathbf{u}), \tag{32}$$

and we refer to $\widehat{\mathrm{var}}_t$ as the "t-tilted empirical variance".

As discussed before, the full gradient of TERM is tilted empirical mean of individual gradients $\{\nabla_\theta f(x_i;\theta)\}_{i\in[N]}$ with weights proportional to $e^{tf(x_i;\theta)}$. In the next section as well as Appendix A.3, we will prove other properties of TERM using tilted empirical mean and variance defined here.

2.4 Interpretation 2: Empirical Bias/Variance Trade-off

Another key property of the TERM solutions is that for any $t \in \mathbb{R}$, t-tilted empirical variance of the losses across all samples will decrease if we increase t by a small amount of value. We formally stated this in Theorem 3.

Theorem 3 (Variance reduction)**.** *Let* $\mathbf{f}(\theta) := (f(x_1;\theta)), \ldots, f(x_N;\theta))$*. Then, under Assumption 3 and Assumption 4, for any* $t \in \mathbb{R}$*,*

$$\frac{\partial}{\partial t}\left\{\widehat{\text{var}}_\tau(\mathbf{f}(\breve{\theta}(t)))\right\}\Big|_{t=\tau} < 0. \tag{33}$$

Note that $\widehat{\text{var}}_\tau$ is τ-tilted empirical variance defined in Eq. (32). Hence, for any t, the t-tilted empirical variance among N losses will decrease if we increase t by a small value. When $\tau = 0$, $\widehat{\text{var}}_\tau$ reduces to standard empirical variance. In particular, Theorem 3 states that the empirical variance of the loss vector decreases if t is chosen to be a small positive value. Therefore, it is possible to trade off between optimizing the average loss vs. reducing variance, allowing the solutions to potentially achieve a better bias-variance trade-off for generalization (Maurer and Pontil, 2009; Bennett, 1962; Hoeffding, 1994). At a high level, this property is consistent with and extends the approximation of TERM mentioned by Liu and Theodorou (2019, Section V.A), which approximates TERM as the empirical risk regularized with variance of the loss at $t = 0$. We rely on this property to achieve better generalization in classification in Section 7.

In addition to empirical variance across all losses, there are other related distribution uniformity measures. In Theorem 4 below, we also prove that entropy of the weight distribution at solution $\breve{\theta}(t)$ tilted by τ close to t is increasing with t, which indicates that larger t's encourages more uniform solutions measured via entropy.

Theorem 4 (Gradient weights become more uniform by increasing t)**.** *Under Assumption 3 and Assumption 4, for any* $t \in \mathbb{R}^{>0}$*,*

$$\frac{\partial}{\partial t} H(\mathbf{w}(\tau; \breve{\theta}(t)))\Big|_{\tau=t} > 0, \tag{34}$$

where $H(\cdot)$ *denotes the Shannon entropy function measured in nats,*

$$H\left(\mathbf{w}(t;\theta)\right) := -\sum_{i\in[N]} w_i(t;\theta)\log w_i(t;\theta). \tag{35}$$

Full proofs of the theorems presented in this section can be found in Appendix A.3. In the next section, we connect TERM to other objectives. Note that the results in all subsequent sections do not require the GLMs assumption, unless stated otherwise.

3. Connections to Other Risk Measures

In this section (and subsequently in Section 4) we explore TERM by comparing, contrasting, and drawing connections between TERM and other common risk measures. To do so, we

first introduce a distributional version of TERM, which is closely related to entropic risk (measure) in previous literature (Ahmadi-Javid, 2012; Föllmer and Schied, 2004). Entropic risk, denoted as $R_X(t;\theta)$, can be viewed as the scaled cumulant generating function of $f(X;\theta)$, i.e.,

$$R_X(t;\theta) := \frac{1}{t}\Lambda_X(t;\theta) = \frac{1}{t}\log\left(\mathbb{E}\left[e^{tf(X;\theta)}\right]\right) = \frac{1}{t}\log\sum_x p(x)p_\theta(x)^{-t}. \tag{36}$$

We note that entropic risk is usually defined over $t \in \mathbb{R}^{>0}$ in the literature (Föllmer and Schied, 2004). In Eq. (36) above, we naturally extend its definition to support $t \in \mathbb{R}$. The TERM objective $\widetilde{R}(t;\theta)$ is the empirical version of entropic risk $R_X(t;\theta)$ ($t \in \mathbb{R}$). One of the contributions of this work can be viewed as providing an operational meaning to the value of the (empirical) entropic risk and rigorously investigating its properties for $t \in \mathbb{R}^{<0}$. In the next sections (Section 3.1–Section 3.3), we characterize various relations between tilted risks (TERM or entropic risk) and other common risk measures, both in terms of the empirical variants (involving TERM) and distributional forms (involving entropic risk).

3.1 TERM and Rényi Cross Entropy

We begin by demonstrating that TERM can be viewed as form of Rényi cross entropy minimization, which helps to explain the uniformity properties of TERM discussed in Section 2.4. Consider the cross entropy between p and p_θ defined by

$$H(p\|p_\theta) := \mathbb{E}\left[f(X;\theta)\right] = \sum_x p(x)\log\left(\frac{1}{p_\theta(x)}\right). \tag{37}$$

Hence, minimizing $\mathbb{E}\left[f(X;\theta)\right]$ is equivalent to minimizing the cross entropy between the true distribution and the postulated distribution. The empirical variant of (37) would be empirical risk minimization (1).

For $\rho \in \mathbb{R}^{>0}$, let Rényi cross entropy of order ρ between p and q be defined as:[4]

$$H_\rho(p\|q) := \frac{1}{1-\rho}\log\left(\sum_x p(x)q(x)^{\rho-1}\right). \tag{38}$$

Rényi cross entropy can be viewed as a natural extension of cross entropy, and in fact it recovers cross entropy for $\rho = 1$, i.e., $H_1(p\|q) = H(p\|q)$. Rényi cross-entropy can also be viewed as a natural extension of Rényi entropy, which it recovers when $p = q$, i.e., $H_\rho(p\|p) = H_\rho(p)$, where Rényi entropy of order ρ is defined as

$$H_\rho(p) := \frac{1}{1-\rho}\log\left(\sum_x p(x)^\rho\right). \tag{39}$$

It is straightforward to see that the entropic risk can be expressed in terms of Rényi cross entropy:

$$R_X(t;\theta) = H_{1-t}(p\|p_\theta). \tag{40}$$

4. H_1 is defined via continuous extension.

Equivalently, in the empirical world, TERM can be expressed as:

$$\widetilde{R}(t;\theta) = H_{1-t}(\mathbf{u}\|\mathbf{w}(1;\theta)), \tag{41}$$

where $\mathbf{u}$ denotes the uniform N-vector and $\mathbf{w}(1;\theta) := (w_1(1;\theta),\ldots,w_n(1;\theta))$ with $w_i(1;\theta)$ defined in Eq. (26), and for any two N-vectors $\mathbf{p}$ and $\mathbf{q}$,

$$H_\rho(\mathbf{p}\|\mathbf{q}) := \frac{1}{1-\rho}\log\left(\sum_{i\in N} p_i q_i^{\rho-1}\right). \tag{42}$$

In other words, if we treat the loss $f(x_i;\theta)$ as log-likelihood of the sample x_i under p_θ, this implies that TERM is the Rényi entropy of order $(1-t)$ between the uniform vector and the normalized likelihood vector of all samples, $\mathbf{w}(1;\theta)$. Hence, minimizing over θ is encouraging the *uniformity* of $\mathbf{w}(1;\theta)$ in the sense of the Rényi cross entropy with the uniform vector.

3.2 TERM as a Regularizer to Empirical Risk

TERM can also be interpreted as a form of regularization in traditional ERM. We first note that by Taylor series expansion at $t = 0$, TERM can be approximately decomposed into empirical risk regularized by t times the empirical variance of the loss, for small t (Liu and Theodorou, 2019, Section V.A). Here, we provide an exact interpretation of TERM as regularized ERM for all t. We first look at the distributional case, i.e., relating $R_X(t;\theta)$ to cross entropy as follows.

Lemma 6. *The entropic risk of order t can be stated as:*

$$R_X(t;\theta) = H(p\|p_\theta) + \frac{1}{t}D(p\|T(p,p_\theta,-t)), \tag{43}$$

where D denotes KL divergence between two distributions and $T(p,p_\theta,-t)$ is a mismatched tilted distribution defined as (Salamatian et al., 2019, Definition 1)

$$T(p,p_\theta,-t)(x) := \frac{p(x)p_\theta(x)^{-t}}{\sum_u p(u)p_\theta(u)^{-t}}. \tag{44}$$

Proof. Consider the following equation:

$$\sum_x p(x)\log\left(\frac{p(x)}{T(p,p_\theta,-t)(x)}\right) = -t\sum_x p(x)\log\frac{1}{p_\theta(x)} + \log\left(\sum_x p(x)p_\theta(x)^{-t}\right), \tag{45}$$

which directly implies the desired identity. □

In other words, entropic risk of order t is equivalent to the cross entropy risk regularized via a tilted mismatched distribution. Let $\mathbf{w}(t;\theta) := (w_1(t;\theta),\ldots,w_n(t;\theta))$ denote the tilted weight vector of the n samples. Our next result is an empirical variant of Lemma 6.

Lemma 7. *TERM objective can be restated as follows:*

$$\widetilde{R}(t;\theta) = \bar{R}(\theta) + \frac{1}{t}D(\mathbf{u}\|\mathbf{w}(t;\theta)), \tag{46}$$

where $\bar{R}(\theta)$ *is the empirical risk* (1), $\mathbf{u}$ *denotes the uniform* N*-vector, i.e.,* $\mathbf{u} := \left(\frac{1}{N}, \ldots, \frac{1}{N}\right)$, *and where for* N*-vectors* $\mathbf{p}$ *and* $\mathbf{q}$,

$$D(\mathbf{p}\|\mathbf{q}) := \sum_{i\in[N]} p_i \log\left(\frac{p_i}{q_i}\right). \tag{47}$$

Proof. The proof is a consequence of the following identity:

$$\frac{1}{t}\frac{1}{N}\sum_{i\in[N]} \log\left(\frac{\frac{1}{N}}{w_i(t;\theta)}\right) + \frac{1}{N}\sum_{i\in[N]} f(x_i;\theta) = \frac{1}{t}\log\left(\frac{1}{N}\sum_{i\in[N]} e^{tf(x_i;\theta)}\right). \tag{48}$$

□

Hence, TERM aims to minimize an average loss regularized by the KL divergence between the weight vector (which exponentially tilts the individual losses) and the uniform vector.

3.3 TERM and Distributionally Robust Risks

Finally, we note that TERM is closely related to distributionally robust optimization (DRO) objectives (e.g., Namkoong and Duchi, 2017; Duchi and Namkoong, 2019; Chen and Paschalidis, 2020; Gürbüzbalaban et al., 2022; Duchi and Namkoong, 2018). In particular, TERM with $t > 0$ is equivalent to a form of DRO with a max-entropy regularizer, i.e., the constraint set is determined by a KL ball around uniform distribution (Föllmer and Knispel, 2011; Qi et al., 2020b; Shapiro et al., 2014):

$$\widetilde{R}(t;\theta) = \max_{q\in\Delta_N}\left\{\sum q_i f(x_i;\theta) - \frac{1}{t}\sum_{i\in[N]} q_i \log N q_i\right\} = \max_{q\in\Delta_N}\left\{H(\mathbf{q}\|\mathbf{w}(1;\theta)) - \frac{1}{t}D(\mathbf{q}\|\mathbf{u})\right\}, \tag{49}$$

and the corresponding relations in the distributional form is

$$R_X(t;\theta) = \max_q\left\{\mathbb{E}_q[f(X;\theta)] - \frac{1}{t}D(q\|p)\right\} = \max_q\left\{H(q\|p_\theta) - \frac{1}{t}D(q\|p)\right\}. \tag{50}$$

This relation is also a special case of Donsker-Varadhan Variational Formula (Dupuis and Ellis, 1997). We note that similar connections between DRO and TERM have also been explored in concurrent works by Qi et al. (2020a,b) specifically in the limited context of stochastic optimization methods for solving class imbalance with $t > 0$.

In the next section, we propose a new risk motivated by TERM, which may be of independent interest.

4. Tilted Value-at-Risk and Value-at-Risk

In this section we provide connections between TERM and risk measures such as Value-at-Risk (VAR) that specifically target loss quantiles. In particular, based on TERM, we propose a new risk—*Tilted Value-at-Risk (TiVaR)* and discuss its relations with existing risks (Section 4.2). We find that TiVaR is a computationally efficient alternative to VaR that provides tighter approximations to VaR than prior risks, which again helps to motivate the use of TERM.

4.1 Tail Probabilities of Losses and Value-at-Risk (VaR)

The tail probabilities of losses focus on quantiles of losses that exceed a certain threshold, as formally defined below.

Definition 2 (Tail probability of losses)**.** *For all $\gamma \in \mathbb{R}$, let $Q_X(\gamma;\theta)$ denote the probability of the losses $f(X;\theta)$ no smaller than γ, i.e.,*

$$Q_X(\gamma;\theta) := P\left[f(X;\theta) \geqslant \gamma\right]. \tag{51}$$

Equivalently, define the empirical variant $\widetilde{Q}(\gamma;\theta)$ over samples x_i for $i \in [N]$:

$$\widetilde{Q}(\gamma;\theta) := \frac{1}{N}\sum_{i\in[N]} \mathbb{I}\left\{f(x_i;\theta) \geqslant \gamma\right\} \tag{52}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function.

Notice that $\widetilde{Q}(\gamma;\theta) \in \left\{0, \frac{1}{N}, \ldots, 1\right\}$ quantifies the fraction of the data for which loss is at least γ. For example, optimizing for 90% of the individual losses (ignoring the worst-performing 10%) could be a more reasonable practical objective than the pessimistic min-max objective. Another common application of this is to use the median in contrast to the mean in the presence of noisy outliers.

Using tail distribution of losses, Value-at-Risk (VaR) (Jorion, 1996) with confidence α $(0 < \alpha < 1)$ is defined as

$$\mathrm{VaR}_X(1-\alpha;\theta) := \min_{\gamma}\left\{\gamma : Q_X(\gamma;\theta) \leqslant \alpha\right\}, \tag{53}$$

and the empirical variant for $\alpha \in \{\frac{k}{N}\}_{k\in[N]}$ is

$$\widetilde{\mathrm{VaR}}(1-\alpha;\theta) := \min_{\gamma}\left\{\gamma : \widetilde{Q}(\gamma;\theta) \leqslant \alpha\right\}. \tag{54}$$

Notice that when we view the loss as log-likelihood of a parametric probability distribution function, $Q_X(\gamma;\theta)$ (Definition 2) can be viewed as the complementary cumulative distribution function (CDF) of the information random variable $f(X;\theta)$. Given the definition of VaR, $Q_X(\gamma;\theta)$ can also be viewed as 'inverted' VaR, as we formalize and prove in Lemma 8 and 9 below. Let

$$Q_X^0(\gamma) := \min_{\theta} Q_X(\gamma;\theta), \quad \theta_X^0(\gamma) \in \arg\min_{\theta} Q_X(\gamma;\theta), \tag{55}$$

$$\widetilde{Q}^0(\gamma) := \min_{\theta} \widetilde{Q}(\gamma;\theta), \quad \theta^0(\gamma) \in \arg\min_{\theta} \widetilde{Q}(\gamma;\theta). \tag{56}$$

where Q_X and $\widetilde{Q}$ is defined in Definition 2. Optimizing $\widetilde{Q}(\gamma;\theta)$ is equivalent to optimizing VaR. Formally, we have the following lemmas.

Lemma 8. *Assume $\min_\theta Q_X(\gamma;\theta)$ is strictly decreasing with γ. We note*

$$\min_{\theta}\left\{\mathit{VaR}_X(1-Q_X^0(\gamma);\theta)\right\} = \gamma, \quad \arg\min_{\theta}\left\{\mathit{VaR}_X(1-Q_X^0(\gamma);\theta)\right\} \ni \theta_X^0(\gamma). \tag{57}$$

Note that $\min_\theta Q_X(\gamma;\theta)$ is non-increasing as γ increases by definition. The additional strict monotonic assumption on $\gamma \mapsto \min_\theta Q_X(\gamma;\theta)$ can be easily satisfied if $f(X;\theta)$ is a continuous random variable and γ is in the range of f. Lemma 8 is proved as follows.

Proof. First, we note for any θ and γ_0 such that $Q_X(\gamma_0;\theta) \leqslant Q_X^0(\gamma)$, we have $\gamma_0 \geqslant \gamma$. Otherwise, there exist $\theta', \gamma' < \gamma$ and $Q_X(\gamma';\theta') \leqslant Q_X^0(\gamma)$, which in turn implies that

$$\min_\theta P[f(X;\theta) \geqslant \gamma'] \leqslant P[f(X;\theta') \geqslant \gamma'] \leqslant Q_X^0(\gamma), \tag{58}$$

contradicting $Q_X^0(\gamma') > Q_X^0(\gamma)$. The proof completes combining with the fact that the function value of $\mathrm{VaR}_X(1 - Q_X^0(\gamma);\theta)$ can achieve γ at any $\theta_X^0(\gamma)$. □

Lemma 9 below describes the empirical variant, which does not require the strict monotonic assumption.

Lemma 9. *For any $\gamma \in (\widetilde{F}(-\infty), \widetilde{F}(+\infty))$ where $\widetilde{F}(t)$ is defined as the optimal tilted objective as in Eq. (11), let $\gamma^0 = \min\left\{\gamma' | \widetilde{Q}^0(\gamma') = \widetilde{Q}^0(\gamma)\right\}$. Then*

$$\min_\theta \left\{\widetilde{VaR}(1 - \widetilde{Q}^0(\gamma^0);\theta)\right\} = \gamma^0, \quad \arg\min_\theta \left\{\widetilde{VaR}(1 - \widetilde{Q}^0(\gamma^0);\theta)\right\} \ni \theta^0(\gamma^0). \tag{59}$$

Both tail distribution of losses and VaR are usually non-smooth and non-convex, and solving them to global optimality is very challenging. In the next section, we show that TiVaR (an objective based on TERM) provides a good upper bound on VaR, and is computationally more efficient, as VaR is not even continuous. In parallel, in Appendix B, we prove that TERM also provides a reasonable approximate solution to the minimizer of tail probability of losses (i.e., inverted VaR).

The proof of one of the main theorems of this section (Theorem 16) relies on a new variant of Chernoff bound for non-negative random variables, which may be of independent interest.

Theorem 5 (Chernoff bound for non-negative random variables)**.** *Let X be a non-negative random variable. Further assume that $E\left[e^{tX}\right] < \infty$ for all $t \in \mathbb{R}$. Then for $\gamma > 0$,*

$$P[X \geqslant \gamma] \leqslant \inf_{t \in \mathbb{R}} \left\{\frac{E\left[e^{tX}\right] - 1}{e^{t\gamma} - 1}\right\} \leqslant \inf_{t \in \mathbb{R}^+} \left\{\frac{E\left[e^{tX}\right]}{e^{t\gamma}}\right\}, \tag{60}$$

where the latter term is the generic Chernoff bound with $\gamma > 0$.

Proof. The theorem holds by applying Markov's inequality twice on $e^{tX} - 1$ ($t \geqslant 0$) and $1 - e^{tX}$ ($t < 0$), and noting that

$$P[X \geqslant \gamma] \leqslant \min\left\{\inf_{t \in \mathbb{R}^{\geqslant 0}} \left\{\frac{E\left[e^{tX}\right] - 1}{e^{t\gamma} - 1}\right\}, \inf_{t \in \mathbb{R}^-} \left\{\frac{1 - E\left[e^{tX}\right]}{1 - e^{t\gamma}}\right\}\right\} = \inf_{t \in \mathbb{R}} \left\{\frac{E\left[e^{tX}\right] - 1}{e^{t\gamma} - 1}\right\}. \tag{61}$$

□

Theorem 5 presents a tighter Chernoff bound for non-negative random variables. To the best of our knowledge, despite the fact that this bound is a simple extension of the generic Chernoff bound, and the existing variants of Chernoff bounds in prior works (Boucheron et al., 2013; Yang and Rosenthal, 2017), we have not seen the result we have here appear elsewhere in this form. In particular, notice that the search for an optimal value of t has been extended from non-negative values to all real numbers. This can result in significantly tighter bounds, especially in small deviations regime, as visualized empirically on two simple distributions in Figure 15, Appendix B. We will see how this leads to significantly better bounds in robustness applications.

4.2 TiVaR: Tilted Value-at-Risk

In this section, we introduce a new risk measure, called Tilted Value-at-Risk (TiVaR). To put TiVaR in perspective, we briefly state other existing risks first. Conditional Value-at-Risk (CVaR) minimizes the average risk of tail events where the risk is above some threshold (Rockafellar et al., 2000; Rockafellar and Uryasev, 2002). One form of CVaR is

$$\mathrm{CVaR}_X(1-\alpha;\theta) := \min_{\gamma} \left\{ \gamma + \frac{1}{\alpha}\mathbb{E}[f(X;\theta)-\gamma]_+ \right\}. \tag{62}$$

It is worth noting that $\mathrm{CVaR}_X(1-\alpha;\theta)$ is a dual formulation of DRO with an uncertainty set that perturbs arbitrary parts of the data by an amount up to $\frac{1}{\alpha}$ (Rockafellar et al., 2000; Curi et al., 2020). Formally, the dual of DRO $\max_{Q:\left\{\frac{dQ}{dP}\leqslant\frac{1}{\alpha}\right\}} \mathbb{E}_Q[f(X;\theta)]$ is $\mathrm{CVaR}_X(1-\alpha;\theta) = \min_\gamma \left\{\gamma + \frac{1}{\alpha}\mathbb{E}[f(X;\theta)-\gamma]_+\right\}$. Some previous works implicitly minimize CVaR by only training on samples with top-k losses (e.g., Fan et al., 2017). Entropic Value-at-Risk (EVaR) is proposed as an upper bound of CVaR and VaR that could be more computationally efficient (Ahmadi-Javid, 2012). EVaR with a confidence level α ($0<\alpha<1$) is defined as:

$$\mathrm{EVaR}_X(1-\alpha;\theta) := \min_{t\in\mathbb{R}^{>0}} \left\{ \frac{1}{t}\log\left(\frac{\mathbb{E}[e^{tf(X;\theta)}]}{\alpha}\right)\right\} = \min_{t\in\mathbb{R}^{>0}} \left\{ R_X(t;\theta) - \frac{1}{t}\log\alpha \right\}. \tag{63}$$

Similarly, for $\alpha \in \{\frac{k}{N}\}_{k\in[N]}$, the empirical variants of CVaR and EVaR are

$$\widetilde{\mathrm{CVaR}}(1-\alpha;\theta) := \min_{\gamma} \left\{ \gamma + \frac{1}{\alpha}\frac{1}{N}\sum_{i\in[N]}[f(x_i;\theta)-\gamma]_+ \right\},$$

$$\widetilde{\mathrm{EVaR}}(1-\alpha;\theta) := \min_{t\in\mathbb{R}^{>0}} \left\{ \frac{1}{t}\log\left(\frac{\frac{1}{N}\sum_{i\in[N]} e^{tf(x_i;\theta)}}{\alpha}\right)\right\} = \min_{t\in\mathbb{R}^{>0}} \left\{ \widetilde{R}(t;\theta) - \frac{1}{t}\log\alpha \right\}.$$

Notice that TERM objective appears as part of the objective in $\widetilde{\mathrm{EVaR}}$, and particularly optimizing $\widetilde{\mathrm{EVaR}}$ with respect to θ would be equivalent to solving TERM for some value of t implicitly defined through α (see Lemma 25 and Lemma 26 in the appendix).

It is known that $\mathrm{VaR}_X(1-\alpha;\theta) \leqslant \mathrm{CVaR}_X(1-\alpha;\theta) \leqslant \mathrm{EVaR}_X(1-\alpha;\theta)$ (Ahmadi-Javid, 2012) which directly yields $\widetilde{\mathrm{VaR}}(1-\alpha;\theta) \leqslant \widetilde{\mathrm{CVaR}}(1-\alpha;\theta) \leqslant \widetilde{\mathrm{EVaR}}(1-\alpha;\theta)$. Meanwhile, to the best of our knowledge, it is not clear from existing works how entropic risk (or TERM) is related to VaR or EVaR. Next, based on TERM, we propose a new risk-averse objective Tilted Value-at-Risk, showing that it upper bounds VaR and lower bounds EVaR.

Definition 3 (Tilted Value-at-Risk (TiVaR)). *Let TiVaR for $\alpha \in (0,1]$ be defined as*

$$TiVaR_X(1-\alpha;\theta) := \min_{t\in\mathbb{R}} \left\{ F_X(-\infty) + \frac{1}{t}\log\left[\frac{e^{(R_X(t;\theta)-F_X(-\infty))t} - (1-\alpha)}{\alpha}\right]_+ \right\}. \tag{64}$$

Similarly, empirical TiVaR is defined for $\alpha \in (0,1)$,

$$\widetilde{TiVaR}\,(1-\alpha;\theta) := \min_{t\in\mathbb{R}} \left\{ \widetilde{F}(-\infty) + \frac{1}{t}\log\left[\frac{e^{(\widetilde{R}(t;\theta)-\widetilde{F}(-\infty))t} - (1-\alpha)}{\alpha}\right]_+ \right\}. \tag{65}$$

We note that TiVaR is not a coherent risk measure (see the work of Artzner (1997); Artzner et al. (1999) for definition of coherent risks), despite that it can be tighter than CVaR in some cases, as discussed in detail later. We next present our main result on relations between TiVaR, VaR, and EVaR.

Theorem 6. *For $\alpha \in (0,1]$ and any θ,*

$$VaR_X(1-\alpha;\theta) \leqslant TiVaR_X(1-\alpha;\theta) \leqslant EVaR_X(1-\alpha;\theta). \tag{66}$$

Similarly, for $\alpha \in \{\frac{k}{N}\}_{k\in[N]}$ and any θ,

$$\widetilde{VaR}\,(1-\alpha;\theta) \leqslant \widetilde{TiVaR}\,(1-\alpha;\theta) \leqslant \widetilde{EVaR}\,(1-\alpha;\theta)\,. \tag{67}$$

We defer the proof to Appendix B, where the main steps include applying the new Chernoff bound variant (Theorem 5). Theorem 6 indicates that $\widetilde{\text{TiVaR}}\,(1-\alpha;\theta)$ is a tighter approximation to $\widetilde{\text{VaR}}\,(1-\alpha;\theta)$ than $\widetilde{\text{EVaR}}\,(1-\alpha;\theta)$.

Comparing TiVaR and CVaR. In general, TiVaR and CVaR are not directly comparable, as both of them can be viewed as approximations to VaR and neither dominates the other, i.e., one risk can be tighter than the other depending on the quantile value, 1-α. In the regimes where α is a large value between some intermediate constant and 1, TiVaR provides a tighter approximation to VaR than CVaR. For instance, in the extreme case when $\alpha \to 1$, $\widetilde{\text{VaR}}$ will be close to the min-loss ($\min_{i\in[N]} f(x_i;\theta)$), while the value of $\widetilde{\text{CVaR}}$ is the mean of the losses ($\frac{1}{N}\sum_{i\in[N]} f(x_i;\theta)$). $\widetilde{\text{TiVaR}}$ reduces to the min-loss in this case. In other words, both $\widetilde{\text{VaR}}$ and $\widetilde{\text{TiVaR}}$ sweep the values between the min-loss and max-loss; whereas $\widetilde{\text{CVaR}}$ sweeps the values between the avg-loss and max-loss. We compare TiVaR with CVaR and other risks in Figure 4 on mean estimation and linear regression problems, and demonstrate that TiVaR is tighter than CVaR especially when α is close 1 (corresponding to robustness applications).[5]

We also note that there exist other risk-averse or risk-seeking formulations that focus on the upper or lower tail of losses, such as the mean-semideviation framework (Kalogerias and Powell, 2018). Mean-semideviation recovers a set of risk measures including mean-upper-semideviations and entropic mean-semideviation. Nevertheless, these risks usually cannot handle both fairness and robustness in a single formulation, and can incur more per-iteration gradient evaluations or worse convergence rates compared to vanilla ERM (Kalogerias and Powell, 2018; Gürbüzbalaban et al., 2022; Zhu et al., 2023).

5. While CVaR focuses on upper quantiles, one may explore 'inverse' CVaR to better approximate the lower quantiles. However, inverse CVaR, ranging from avg-loss to min-loss, is not a valid upper bound of VaR. Despite this, we empirically explore this approximation to solving VaR, among others, in Appendix B.

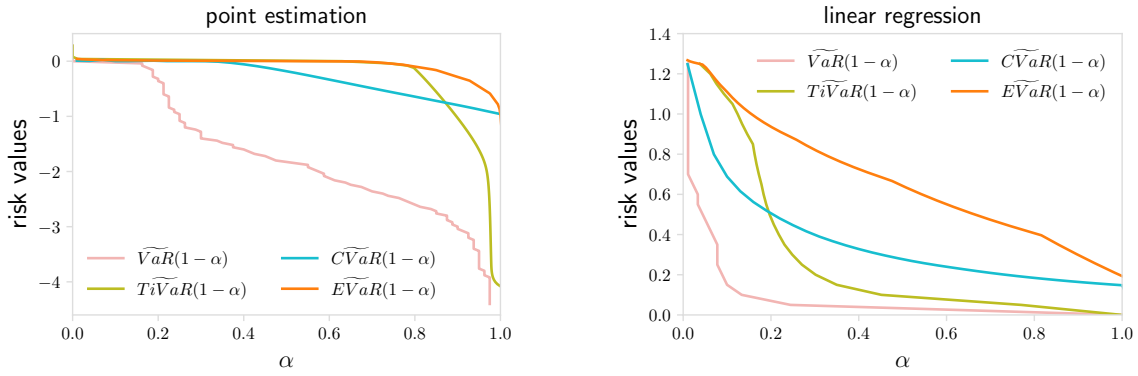


Figure 4: Comparing values of VaR, TiVaR, CVaR, and EVaR. $\widetilde{\mathrm{VaR}}(1-\alpha) := \min_\theta \widetilde{\mathrm{VaR}}(1-\alpha;\theta)$, and $\widetilde{\mathrm{TiVaR}}(1-\alpha)$, $\widetilde{\mathrm{CVaR}}(1-\alpha)$, and $\widetilde{\mathrm{EVaR}}(1-\alpha)$ are defined in a similar way. From Theorem 6, we know $\widetilde{\mathrm{VaR}}(1-\alpha;\theta) \leqslant \widetilde{\mathrm{TiVaR}}(1-\alpha;\theta) \leqslant \widetilde{\mathrm{EVaR}}(1-\alpha;\theta)$, which is also visualized here. Both CVaR and TiVaR values are between VaR and EVaR. TiVaR provides a tighter approximation to VaR than CVaR when α is closer to 1.

Finally, we draw connections between the above results and the k-loss, defined as the k-th smallest loss of N (i.e., 1-loss is the min-loss, N-loss is the max-loss, $(N-1)/2$-loss is the median-loss). Formally, let $R_{(k)}(\theta)$ be the k-th order statistic of the loss vector. Hence, $R_{(k)}$ is the k-th smallest loss, and particularly

$$R_{(1)}(\theta) = \check{R}(\theta), \quad R_{(N)}(\theta) = \hat{R}(\theta). \tag{68}$$

Thus, for any $k \in [N]$, we define

$$R^*_{(k)} := \min_\theta R_{(k)}(\theta), \quad \theta^*(k) := \arg\min_\theta R_{(k)}(\theta). \tag{69}$$

Note that

$$R^*_{(1)} = \widetilde{F}(-\infty), \quad R^*_{(N)} = \widetilde{F}(+\infty). \tag{70}$$

While minimizing the k-loss is more desirable than ERM in many applications, the k-loss is non-smooth (and generally non-convex), and is challenging to solve for large-scale problems (Jin et al., 2020; Nouiehed et al., 2019). TERM offers a good approximation to k-loss as well. Note that if we fix $\alpha = 1 - \frac{k}{N}$, minimizing k-loss is equivalent to minimizing γ where $\widetilde{Q}(\gamma;\theta) = \alpha$. Based on the bound of $\widetilde{\mathrm{VaR}}$, we obtain a bound on k-loss:

Corollary 7. *For all $k \in \{2,\dots,N-1\}$, and all $t \in \mathbb{R}$:*

$$R_{(k)}(\theta) \leqslant \min_t \left\{ \widetilde{F}(-\infty) + \frac{1}{t}\log\left[\frac{e^{(\tilde{R}(t;\theta)-\widetilde{F}(-\infty))t} - \frac{k}{N}}{1-\frac{k}{N}}\right]_+ \right\} \leqslant \min_{t\in\mathbb{R}^{>0}} \left\{ \widetilde{R}(t;\theta) - \frac{1}{t}\log\left(1-\frac{k}{N}\right)\right\}. \tag{71}$$

Proof. Note that

$$R_{(k)}(\theta) = \widetilde{\mathrm{VaR}}\left(\frac{k}{N}; \theta\right). \tag{72}$$

The proof completes by setting $\alpha = 1 - \frac{k}{N}$ in Eq. (65) and noting $\widetilde{VaR}(1-\alpha;\theta) \leqslant \widetilde{TiVaR}(1-\alpha;\theta) \leqslant \widetilde{EVaR}(1-\alpha;\theta)$. □

Corollary 7 optimizes over all $t \in \mathbb{R}$ over the upper bound of $R_{(k)}(\theta)$, which can be relaxed to searching over positive t's, as stated in Corollary 8 below.

Corollary 8. *For all $k \in \{2, \ldots, N-1\}$, and all $t \in \mathbb{R}^{>0}$:*

$$R_{(k)}(\theta) \leqslant \widetilde{F}(-\infty) + \frac{1}{t} \log\left(\frac{e^{(\widetilde{R}(t;\theta) - \widetilde{F}(-\infty))t} - \frac{k}{N}}{1 - \frac{k}{N}}\right). \tag{73}$$

5. Solving TERM

In this section, we develop first-order batch (Section 5.1) and stochastic (Section 5.2) optimization methods for solving TERM, and rigorously analyze the effects that t has on the convergence of these methods.

Recall that in Section 2.2, we discuss the Lipschitzness, convexity, and smoothness properties of TERM. t-tilted loss remains strongly convex for $t > 0$, so long as the original loss function is strongly convex. On the other hand, for sufficiently large negative t, the t-tilted loss becomes non-convex. Hence, while the t-tilted solutions for positive t are unique, the objective may have multiple (spurious) local minima for negative t even if the original loss function is strongly convex. For negative t, we seek the solution for which the parametric set of t-tilted solutions obtained by sweeping $t \in \mathbb{R}$ (i.e., $\breve{\theta}(t)$ defined in Eq. (10)) remains continuous (as in Figure 1a-c and Figure 2). To this end, for negative t, we solve TERM by smoothly decreasing t from 0 observing that the solutions form a continuum in $\mathbb{R}^d$ empirically. Despite the non-convexity of TERM with $t < 0$, we find that this approach produces effective solutions to multiple real-world problems in Section 7. Additionally, as the objective remains smooth, it is still relatively efficient to solve. On the toy problem studied in Figure 2, we plot the convergence with t in Figure 5 below.

5.1 First-Order Batch Methods

TERM solver in the batch setting is summarized in Algorithm 1. The main steps include running gradient descent on $\widetilde{R}(t;\theta)$, which involve computing the tilted gradients (i.e., a weighted aggregation of individual gradients (Lemma 5)) of the objective. We also provide convergence results in Theorem 9–11 below for Algorithm 1.

Theorem 9 (Convergence of Algorithm 1 for strongly-convex problems)**.** *Under Assumption 2, there exist $\beta_{\max} \leqslant C_1 < \infty$ and $C_2 < \infty$ that do not depend on t such that for any $t \in \mathbb{R}^{>0}$, setting the step size $\alpha = \frac{1}{C_1 + C_2 t}$, after k iterations:*

$$\widetilde{R}(t, \theta_k) - \widetilde{R}(t, \breve{\theta}(t)) \leqslant \left(1 - \frac{\beta_{\min}}{C_1 + C_2 t}\right)^k \left(\widetilde{R}(t, \theta_0) - \widetilde{R}(t, \breve{\theta}(t))\right). \tag{74}$$

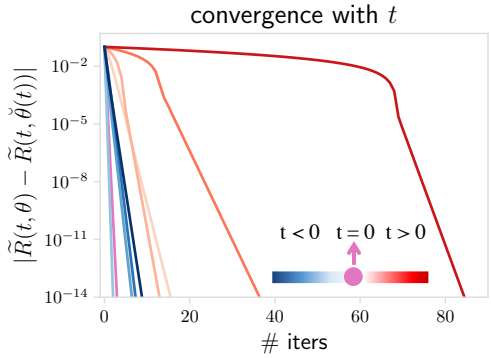


Figure 5: As $t \to +\infty$, the objective becomes less smooth in the vicinity of the final solution where smoothness can be measured by the upper bound of Hessian (see Lemma 3), hence suffering from slower convergence. For negative values of t, TERM converges fast due to the smoothness in the vicinity of solutions despite its non-convexity.

Algorithm 1: Batch (Non-Hierarchical) TERM

Input: t, α, θ

while *stopping criteria not reached* **do**

compute the loss $f(x_i;\theta)$ and gradient $\nabla_\theta f(x_i;\theta)$ for all $i \in [N]$

$\widetilde{R}(t;\theta) \leftarrow t$-tilted loss (2) on all $i \in [N]$

$w_i(t;\theta) \leftarrow e^{t(f(x_i;\theta)-\widetilde{R}(t;\theta))}$

$\theta \leftarrow \theta - \frac{\alpha}{N} \sum_{i\in[N]} w_i(t;\theta) \nabla_\theta f(x_i;\theta)$

end

Proof. First note that by Lemma 2, $\widetilde{R}(t,\theta)$ is $\beta_{\min}$-strongly convex for all $t \in \mathbb{R}^{>0}$. Next, by Lemma 3, there exist $C_1, C_2 < \infty$ such that $\widetilde{R}(t;\theta)$ has $(C_1 + C_2 t)$-Lipschitz gradients for all $t \in \mathbb{R}^{>0}$. The result follows directly from Karimi et al. (2016, Theorem 1). □

Note that under additional assumptions on L-Lipschitzness of $f(x;\theta)$, we can plug in the explicit smoothness constants established by Lowy and Razaviyayn (2021, Lemma 5.3) to obtain explicit constants in the convergence rate, i.e., $C_1 = \beta_{\max}$ and $C_2 = L^2$. Theorem 9 indicates that solving TERM to a local optimum using gradient-based methods tends to be as efficient as traditional ERM for small-to-moderate values of t (Jin et al., 2017), which we corroborate via experiments on multiple real-world datasets in Section 7. This is in contrast to solving for the min-max solution, which would be similar to solving TERM as $t \to +\infty$ (Kort and Bertsekas, 1972; Pee and Royset, 2011; Ostrovskii et al., 2020).

Theorem 10 (Convergence of Algorithm 1 for smooth problems satisfying PL conditions)**.** *Assume $f(x;\theta)$ is $\beta_{\max}$-smooth and (possibly) non-convex. Further assume $\sum_{i\in[N]} p_i f(x_i;\theta)$ is $\frac{\mu}{2}$-PL for any $\mathbf{p} \in \Delta_N$ where $\mathbf{p} := (p_1, \ldots, p_N)$. There exist $\beta_{\max} \leqslant C_1 < \infty$ and $C_2 < \infty$ that do not depend on t such that for any $t \in \mathbb{R}^{>0}$, setting the step size $\alpha = \frac{1}{C_1 + C_2 t}$, after k iterations:*

$$\widetilde{R}(t,\theta_k) - \widetilde{R}(t,\breve{\theta}(t)) \leqslant \left(1 - \frac{\mu}{C_1 + C_2 t}\right)^k \left(\widetilde{R}(t,\theta_0) - \widetilde{R}(t,\breve{\theta}(t))\right), \tag{75}$$

Proof. If $\sum_{i\in[N]} p_i f(x_i;\theta)$ is μ-PL for any $\mathbf{p}\in\Delta_N$, then $\widetilde{R}(t;\theta)$ is μ-PL (Qi et al., 2020a). $\widetilde{R}(t;\theta)$ is $\beta_{\max}$ smooth for $t<0$ and its smoothness parameter scales linearly with t for $t>0$, following the same proof as Lemma 3. □

Theorem 10 applies to both convex and non-convex smooth functions satisfying PL conditions. Again, here we can plug in explicit smoothness parameter (Lowy and Razaviyayn, 2021, Lemma 5.3) if $f(x;\theta)$ is Lipschitz. We next state results without the PL condition assumption for completeness.

Theorem 11 (Convergence of Algorithm 1 for non-convex smooth problems)**.** *Assume $f(x;\theta)$ is $\beta_{\max}$-smooth and (possibly) non-convex. Setting the step size $\alpha=\frac{1}{\beta(t)}$, after K iterations, we have:*

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla\widetilde{R}(t,\theta_k)\|^2\leqslant\frac{2\beta(t)(\widetilde{R}(t,\theta_0)-\widetilde{R}(t,\breve{\theta}(t)))}{K},\tag{76}$$

where for $t\in\mathbb{R}^{>0}$, $\beta(t)=C_1+Ct$ where C_1,C_2 are independent of t and $\beta_{\max}\leqslant C_1<\infty, C_2<\infty$, and for $t\in\mathbb{R}^{-}$, $\beta(t)=\beta_{\max}$.

Theorem 11 also covers the case of convex $f(x;\theta)$ with $t<0$. We note that for non-convex problems, when $t<0$, the convergence rate is independent of t under our assumptions. We also observe this on a toy problem in Figure 5. In all applications we studied in Section 7 with negative t's, TERM runs the same number iterations as those of ERM.

5.2 First-Order Stochastic Methods

To obtain unbiased stochastic gradients, we need to have access to the normalization weights for each sample (i.e., $\frac{1}{N}\sum_{i\in[N]}e^{tf(x_i;\theta)}$), which is often intractable to compute for large-scale problems. Hence, we use $\widetilde{\widetilde{R}}_t$, a term that incorporates stochastic dynamics, to estimate the tilted objective $\widetilde{R}_t:=\widetilde{R}(t;\theta)$, which is used for normalizing the weights as in (25). In particular, we do not use a trivial linear averaging of the current estimate and the history to update $\widetilde{\widetilde{R}}_t$. Instead, we use a tilted averaging to ensure an unbiased estimator (if θ is not being updated).

On the other hand, the TERM objective can be viewed as a composition of functions $\frac{1}{N}\sum_{i\in[N]}e^{tf(x_i;\theta)}$ and $\frac{1}{t}\log(\cdot)$, and could be optimized based on previous stochastic compositional optimization techniques (e.g., Wang et al., 2017; Qi et al., 2020b,a; Wang et al., 2016a; Ghadimi et al., 2020). Similar to Wang et al. (2017), we maintain two sequences (in our context, the model θ and the objective estimate $\widetilde{\widetilde{R}}_t$) throughout the optimization process. This (non-hierarchical) stochastic algorithm is summarized in Algorithm 2 below.

For the purpose of analysis, we sample two independent mini-batches to obtain the gradient of the original loss functions $\nabla_\theta f(x;\theta)$ and update $\widetilde{\widetilde{R}}_t$, respectively (described in Algorithm 5 for completeness). As we will see in Theorem 12, the additional randomness allows us to achieve better convergence rates compared with the algorithm proposed in Wang et al. (2017) instantiated to our objective. Our rate of this simple algorithm matches the rate of more complicated ones (Qi et al., 2020a), and developing optimal optimization procedures is out of the scope of this work. Empirically, we observe that sampling two mini-batches

yield similar performance as using the same mini-batch to query the individual losses and the weights (Figure 17 in Appendix C.2). Therefore, we employ the cheaper variant of just involving one mini-batch (Algorithm 2) in the corresponding experiments.

Algorithm 2: Stochastic (Non-Hierarchical) TERM

Initialize: $\theta, \widetilde{\widetilde{R}}_t = \frac{1}{t}\log\left(\frac{1}{N}\sum_{i\in[N]} e^{tf(x_i;\theta)}\right)$
Input: t, α, λ
while *stopping criteria not reached* **do**
  sample a minibatch B uniformly at random from $[N]$
  compute the loss $f(x;\theta)$ and gradient $\nabla_\theta f(x;\theta)$ for all $x \in B$
  $\widetilde{R}_{B,t} \leftarrow t$-tilted loss (2) on minibatch B
  $\widetilde{\widetilde{R}}_t \leftarrow \frac{1}{t}\log\left((1-\lambda)e^{t\widetilde{\widetilde{R}}_t} + \lambda e^{t\widetilde{R}_{B,t}}\right)$
  $w_{t,x} \leftarrow e^{tf(x;\theta) - t\widetilde{\widetilde{R}}_t}$
  $\theta \leftarrow \theta - \frac{\alpha}{|B|}\sum_{x\in B} w_{t,x}\nabla_\theta f(x;\theta)$
end

The stochastic algorithm developed here requires roughly the same time/space complexity as mini-batch SGD, and thus scales similarly for large-scale problems. It can also help mitigate the potential numerical issues in implementation caused by the exponential tilting operator. We find that these methods perform well empirically on a variety of tasks (Section 7).

Theorem 12 (Convergence of Algorithm 5 for strongly-convex problems)**.** *Assume* $f : \mathcal{X}\times\Theta \to [\widetilde{F}_{\min}, \widetilde{F}_{\max}]$ *is L-Lipschitz in* θ*, i.e.,* $\widetilde{F}_{\min} \leqslant f(x;\theta) \leqslant \widetilde{F}_{\max}$,[6] *and* $|f(x;\theta_i) - f(x;\theta_j)| \leqslant L\|\theta_i - \theta_j\|$ *for* $x \in \mathcal{X}$ *and* $\theta_i, \theta_j \in \Theta \subseteq \mathbb{R}^d$*. Assume* $\widetilde{R}(t;\theta)$ *has compact domain* θ*. Assume* $\widetilde{R}(t;\theta)$ *is* μ*-strongly convex (Assumption 2) with uniformly bounded stochastic gradient, i.e.,* $\|\nabla\widetilde{R}(x_i;\theta)\| := \left\|\frac{e^{tf(x_i;\theta)}}{e^{t\widetilde{R}(t;\theta)}}\nabla f(x_i;\theta)\right\| \leqslant B$ *for* $\theta \in \mathbb{R}^d$ *and* $i \in [N]$*. Denote* $k_t := \arg\max_k\left(k < \frac{2e}{\mu} + \frac{etLBe^{t(\widetilde{F}_{\max}-\widetilde{F}_{\min})}}{\mu k}\right)$*. Assume the batch size is 1. For* $k \geqslant k_t$*,*

$$\mathbb{E}[\|\theta_{k+1} - \theta^*\|^2] \leqslant \frac{V_t}{k+1}, \tag{77}$$

where

$$\theta^* := \breve{\theta}(t), \quad V_t = \max\left\{k_t\mathbb{E}[\|\theta_{k_t} - \theta^*\|^2], \frac{4B^2e^{2+2t(\widetilde{F}_{\max}-\widetilde{F}_{\min})}}{\mu^2}\right\}, \tag{78}$$

and

$$\mathbb{E}[\|\theta_{k_t} - \theta^*\|^2] \leqslant \max\left\{\mathbb{E}[\|\theta_1 - \theta^*\|^2], \frac{B^2e^{2t(\widetilde{F}_{\max}-\widetilde{F}_{\min})+1}}{\mu(1+tLBe^{t(\widetilde{F}_{\max}-\widetilde{F}_{\min})})}\right\}. \tag{79}$$

6. For notation consistency between the max-loss and min-loss for any sample and any iteration, we use $\widetilde{F}_{\min}$ to denote the lower bound of $f(x_i;\theta_k)$. We note that $\widetilde{F}_{\min} = \widetilde{F}(-\infty)$ defined in Definition 11.

Our assumptions are standard compared with those in related literature (Wang et al., 2017; Qi et al., 2020b). The uniformly bounded stochastic gradient of $\widetilde{R}(t;\theta)$ assumption can be satisfied by the bounded gradient of $f(x_i;\theta)$, which can be a limiting condition but has appeared in previous works on stochastic compositional optimization (Qi et al., 2020b; Wang et al., 2016a). If the objectives are coercive, which typically holds in practice (Bertsekas, 1997), Algorithm 2 will have bounded iterates and thus the compact domain assumption would hold. We defer full proofs to Appendix C.2. The main steps involve bounding the expected estimation error $\mathbb{E}[e^{t(\widetilde{R}_k-\widetilde{\widetilde{R}}_k)}]$ conditioning on the previous iterates $\{\theta_1,\ldots,\theta_k\}$.

Discussions. The theorem indicates that Algorithm 2 starts to make progress after k_t iterations, with convergence rate $O(e^{2t}/k)$. Both k_t and V_k could scale exponentially with t in the worst-case analysis, but it does not completely reflect the dependence of Algorithm 2 on t for modest values of t. Empirically, we observe that the stochastic TERM solver with moderate values of t can converge faster compared with stochastic min-max solvers, which has a rate of $1/\sqrt{k}$ for strongly convex problems (Levy et al., 2020). This leaves open for future work understanding the exact scaling of the convergence rate of stochastic TERM as $t\to\infty$.

Next, we present convergence results on non-convex smooth problems, without and with the assumptions of PL-conditions. We defer all proofs to Appendix C.2.

Theorem 13 (Convergence of Algorithm 5 for non-convex smooth problems)**.** *Assume $f:\mathcal{X}\times\Theta\to[\widetilde{F}_{\min},\widetilde{F}_{\max}]$ is L-Lipschitz in θ, i.e., $\widetilde{F}_{\min}\leqslant f(x;\theta)\leqslant\widetilde{F}_{\max}$, and $|f(x;\theta_i)-f(x;\theta_j)|\leqslant L\|\theta_i-\theta_j\|$ for $x\in\mathcal{X}$ and $\theta_i,\theta_j\in\Theta\subseteq\mathbb{R}^d$. Assume $\widetilde{R}(t;\theta)$ is β-smooth with uniformly bounded stochastic gradient, i.e., $\|\nabla\widetilde{R}(x_i;\theta)\|\leqslant B$ for $\theta\in\mathbb{R}^d$ and $i\in[N]$. Assume the batch size is 1. Denote $k_t:=\left\lceil\frac{2(\widetilde{F}_{\max}-\widetilde{F}_{\min})t^2L^2}{\beta e^2}\right\rceil$, then for $k\geqslant k_t$,*

$$\frac{1}{K}\sum_{k=k_t}^{K}\mathbb{E}[\|\nabla\widetilde{R}(t;\theta_k)\|^2]\leqslant\sqrt{8}Be^{t(\widetilde{F}_{\max}-\widetilde{F}_{\min})+1}\sqrt{\frac{\beta(\widetilde{F}_{\max}-\widetilde{F}_{\min})}{K}}. \tag{80}$$

Theorem 14 (Convergence of Algorithm 5 for non-convex smooth problems with PL conditions)**.** *Let the assumptions in Theorem 13 hold. Further assume that $\sum_{i\in[N]}p_if(x_i;\theta)$ satisfies $\frac{\mu}{2}$-PL conditions for any $\mathbf{p}\in\Delta_N$ where $\mathbf{p}:=(p_1,\ldots,p_N)$. Assume the batch size is 1. Denote $k_t:=\operatorname{argmax}_k\left(k<\frac{4e}{\mu}+\frac{4etLBE^{t(\widetilde{F}_{\max}-\widetilde{F}_{\min})}}{\mu k}\right)$, then for $t\in\mathbb{R}^{>0}$ and $k\geqslant k_t$,*

$$\mathbb{E}[\widetilde{R}(t;\theta_{k+1})-\widetilde{R}(t;\breve{\theta}(t))]\leqslant\frac{V_t}{k+1}, \tag{81}$$

where

$$V_t=\max\left\{k_t\mathbb{E}[\widetilde{R}(t;\theta_{k_t})-\widetilde{R}(t;\breve{\theta}(t))],\frac{8\beta B^2e^{2t(\widetilde{F}_{\max}-\widetilde{F}_{\min})+2}}{\mu^2}\right\}. \tag{82}$$

6. TERM Extended: Hierarchical Multi-Objective Tilting

We consider an extension of TERM that can be used to address practical applications requiring multiple objectives, e.g., simultaneously achieving robustness to noisy data *and*

ensuring fair performance across subgroups. Existing approaches typically aim to address such problems in isolation. To handle multiple objectives with TERM, let each sample x be associated with a group $g \in [G]$, i.e., $x \in g$. These groups could be related to the labels (e.g., classes in a classification task), or may depend only on features. For any $t,\tau \in \mathbb{R}$, we define multi-objective TERM as:

$$\widetilde{J}(t,\tau;\theta) := \frac{1}{t}\log\left(\frac{1}{N}\sum_{g\in[G]}|g|e^{t\widetilde{R}_g(\tau;\theta)}\right), \text{where} \widetilde{R}_g(\tau;\theta) := \frac{1}{\tau}\log\left(\frac{1}{|g|}\sum_{x\in g}e^{\tau f(x;\theta)}\right), \tag{83}$$

and $|g|$ is the size of group g. We evaluate the gradient of the hierarchical multi-objective tilt objective in Lemma 10 below.

Lemma 10 (Hierarchical multi-objective tilted gradient)**.** *Under Assumption 1,*

$$\nabla_\theta \widetilde{J}(t,\tau;\theta) = \sum_{g\in[G]}\sum_{x\in g} w_{g,x}(t,\tau;\theta)\nabla_\theta f(x;\theta) \tag{84}$$

where

$$w_{g,x}(t,\tau;\theta) := \frac{\left(\frac{1}{|g|}\sum_{y\in g}e^{\tau f(y;\theta)}\right)^{(\frac{t}{\tau}-1)}}{\sum_{g'\in[G]}|g'|\left(\frac{1}{|g'|}\sum_{y\in g'}e^{\tau f(y;\theta)}\right)^{\frac{t}{\tau}}} \quad e^{\tau f(x;\theta)}. \tag{85}$$

Similar to the tilted gradient (25), Lemma 10 indicates that the multi-objective tilted gradient is a weighted sum of the gradients, making TERM similarly efficient to solve. Multi-objective TERM recovers sample-level TERM as a special case for $\tau = t$ (Lemma 11), and reduces to group-level TERM with $\tau \to 0$.

Lemma 11 (Sample-level TERM is a special case of hierarchical multi-objective TERM)**.** *Under Assumption 1, hierarchical multi-objective TERM recovers TERM as a special case for $t=\tau$. That is*

$$\widetilde{J}(t,t;\theta) = \widetilde{R}(t;\theta). \tag{86}$$

Proof. The proof is completed by noticing that setting $t=\tau$ in (85) recovers the original sample-level tilted gradient. □

Note that all properties discussed in Section 2 carry over to group-level TERM. We validate the effectiveness of hierarchical tilting empirically in Section 7.3, where we show that TERM can significantly outperform baselines to handle class imbalance *and* noisy outliers simultaneously, while underperforming a much more complicated method in their setup. Note that hierarchical tilting could be extended to hierarchies of greater depths (than two) to simultaneously handle more than two objectives at the cost of one extra tilting hyperparameter per each additional optimization objective. For instance, we state the multi-objective tilting for a hierarchy of depth three in Appendix C.1.

6.1 Solving Hierarchical TERM

To solve hierarchical TERM in the batch setting, we can directly use gradient-based methods with tilted gradients defined for the hierarchical objective in Lemma 10. Note that Batch hierarchical TERM with $t{=}\tau$ reduces to solving the sample-level tilted objective (2). We summarize this method in Algorithm 3.

Algorithm 3: Batch Hierarchical TERM

Input: t,τ,α
while *stopping criteria not reached* **do**
  for $g{\in}[G]$ **do**
    compute the loss $f(x;\theta)$ and gradient $\nabla_\theta f(x;\theta)$ for all $x{\in}g$
    $\widetilde{R}_{g,\tau}{\leftarrow}\tau$-tilted loss (83) on group g
    $\nabla_\theta \widetilde{R}_{g,\tau}{\leftarrow}\frac{1}{|g|}\sum_{x\in g} e^{\tau f(x;\theta)-\tau\widetilde{R}_{g,\tau}}\nabla_\theta f(x;\theta)$
  end
  $\widetilde{J}_{t,\tau}{\leftarrow}\frac{1}{t}\log\Big(\frac{1}{N}\sum_{g\in[G]}|g|e^{t\widetilde{R}_g(\tau;\theta)}\Big)$
  $w_{t,\tau,g}{\leftarrow}|g|e^{t\widetilde{R}_{\tau,g}-t\widetilde{J}_{t,\tau}}$
  $\theta{\leftarrow}\theta{-}\frac{\alpha}{N}\sum_{g\in[G]} w_{t,\tau,g}\nabla_\theta\widetilde{R}_{g,\tau}$
end

We next discuss stochastic solvers for hierarchical multi-objective tilting. We extend Algorithm 2 to the multi-objective setting, presented in Algorithm 4. At a high level, at each iteration, group-level tilting is addressed by choosing a group based on the tilted weight vector. Sample-level tilting is then incorporated by re-weighting the samples in a uniformly drawn mini-batch. Similarly, we estimate the tilted objective $\widetilde{R}_{g,\tau}$ for each group g via a tilted average of the current estimate and the history. While we sample the group from which we draw the minibatch, for small number of groups, one might want to draw one minibatch per each group and weight the resulting gradients accordingly.

Algorithm 4: Stochastic Hierarchical TERM

Initialize: $\widetilde{\widetilde{R}}_{g,\tau}{=}0\ \forall g{\in}[G]$
Input: t,τ,α,λ
while *stopping criteria not reached* **do**
  sample g on $[G]$ from a Gumbel-Softmax distribution with logits $\widetilde{\widetilde{R}}_{g,\tau}{+}\frac{1}{t}\log|g|$ and temperature $\frac{1}{t}$
  sample minibatch B uniformly at random within group g
  compute the loss $f(x;\theta)$ and gradient $\nabla_\theta f(x;\theta)$ for all $x{\in}B$
  $\widetilde{R}_{B,\tau}{\leftarrow}\tau$-tilted loss (2) on minibatch B
  $\widetilde{\widetilde{R}}_{g,\tau}{\leftarrow}\frac{1}{\tau}\log\Big((1{-}\lambda)e^{\tau\widetilde{\widetilde{R}}_{g,\tau}}+\lambda e^{\tau\widetilde{R}_{B,\tau}}\Big)$
  $w_{\tau,x}{\leftarrow}e^{\tau f(x;\theta)-\tau\widetilde{\widetilde{R}}_{g,\tau}}$
  $\theta{\leftarrow}\theta{-}\frac{\alpha}{|B|}\sum_{x\in B} w_{\tau,x}\nabla_\theta f(x;\theta)$
end

Group-level tilting can be recovered from Algorithm 3 and 4 by setting the inner-level tilt parameter $\tau=0$. We apply TERM to a variety of machine learning problems; for clarity, we summarize the applications and their corresponding algorithms in Table 10 in the appendix.

7. TERM in Practice: Use Cases

We now showcase the flexibility, wide applicability, and competitive performance of the TERM framework through empirical results on a variety of real-world problems such as handling outliers (Section 7.1), ensuring fairness and improving generalization (Section 7.2), and addressing compound issues (Section 7.3). Despite the relatively straightforward modification TERM makes to traditional ERM, we show that t-tilted losses not only outperform ERM, but either outperform or are competitive with state-of-the-art, problem-specific tailored baselines on a wide range of applications. We provide implementation details in Appendix D.2. All code, datasets, and experiments are publicly available at `github.com/litian96/TERM`. The applications explored are summarized in Table 1 below.

Table 1: Summary of TERM applications.

Applications		**Sections**
Mitigating noisy outliers ($t<0$)	Robust regression	Sec. 7.1.1
	Robust classification	Sec. 7.1.2
	Low-quality annotators	Sec. 7.1.3
Fairness and generalization ($t>0$)	Fair PCA	Sec. 7.2.1
	Fair federated learning	Sec. 7.2.2
	Fair meta-learning	Sec. 7.2.3
	Handling class imbalance	Sec. 7.2.4
	Improving generalization via variance reduction	Sec. 7.2.5
Hierarchical multi-objective tilting	Class imbalance and random noise	Sec. 7.3.1
	Class imbalance and adversarial noise	Sec. 7.3.2

Choosing t. In applications when we consider tradeoffs between different objectives (e.g., fair meta-learning and federated learning), we perform a grid search over t from {0.1, 1, 2, 5, 10, 50, 100, 200} on the validation set and pick the one with the best fairness performance while not degrading mean performance. When there is not a single t dominating other values (e.g., fair PCA), we report results under different values of t. In our initial robust regression experiments, we find that the performance is robust to various t's, and we thus use a fixed $t=-2$ for all experiments involving negative t (Section 7.1 and Section 7.3). For all values of t tested, the number of iterations required to solve TERM is within $2\times$ that of standard ERM, with the same per-iteration complexity.

7.1 Mitigating Noisy Outliers ($t<0$)

We begin by investigating TERM's ability to find robust solutions that reduce the effect of noisy outliers. We note that we specifically focus on the setting of 'robustness' involving random additive noise; the applicability of TERM to more adversarial forms of robustness would be an interesting direction of future work. We do not compare with approaches that

require additional clean validation data (e.g., Roh et al., 2020; Veit et al., 2017; Hendrycks et al., 2018; Ren et al., 2018), as such data can be costly to obtain in practice.

7.1.1 Robust Regression

Label noise. We first consider a regression task with noise corrupted targets, where we aim to minimize the root mean square error (RMSE) on samples from the Drug Discovery dataset (Olier et al., 2018; Diakonikolas et al., 2019). The task is to predict the bioactivities given a set of chemical compounds. We compare against linear regression with an L_2 loss, which we view as the 'standard' ERM solution for regression, as well as with losses commonly used to mitigate outliers—the L_1 loss and Huber loss (Huber, 1964). We also compare with consistent robust regression (CRR) (Bhatia et al., 2017) and STIR (Mukhoty et al., 2019), recent state-of-the-art methods specifically designed for label noise in robust regression. In this particular problem, TERM is equivalent to exponential squared loss, studied in (Wang et al., 2013). We apply TERM at the sample level with an L_2 loss, and generate noisy outliers by assigning random targets drawn from $\mathcal{N}(5,5)$ on a fraction of the samples.

In Table 2, we report RMSE on clean test data for each objective and under different noise levels. We also present the performance of an oracle method (Genie ERM) which has access to all of the clean data samples with the noisy samples removed. *Note that Genie ERM is not a practical algorithm and is solely presented to set the expected performance limit in the noisy setting.* The results indicate that TERM is competitive with baselines on the 20% noise level, and achieves better robustness with moderate-to-extreme noise. We observe similar trends in scenarios involving both noisy features and targets (Appendix D.1). CRR tends to run slowly as it scales cubicly with the number of dimensions (Bhatia et al., 2017), while solving TERM is roughly as efficient as ERM.

Table 2: TERM is competitive with robust *regression* baselines, particularly in high noise regimes.

objectives	**test RMSE** (`Drug Discovery`)		
	20% noise	40% noise	80% noise
ERM	1.87 (.05)	2.83 (.06)	4.74 (.06)
L_1	**1.15** (.07)	1.70 (.12)	4.78 (.08)
Huber (Huber, 1964)	**1.16** (.07)	1.78 (.11)	4.74 (.07)
STIR (Mukhoty et al., 2019)	**1.16** (.07)	1.75 (.12)	4.74 (.06)
CRR (Bhatia et al., 2017)	**1.10** (.07)	1.51 (.08)	4.07 (.06)
TERM	**1.08** (.05)	**1.10** (.04)	**1.68** (.03)
Genie ERM	1.02 (.04)	1.07 (.04)	1.04 (.03)

Label and feature noise. Here, we present results involving both feature noise and target noise. We investigate the performance of TERM on two datasets (cal-housing (Pace and Barry, 1997) and abalone (Dua and Graff, 2019)) used in Yu et al. (2012). Both datasets have features with 8 dimensions. We generate noisy samples following the setup in Yu et al. (2012)—sampling 100 training samples, and randomly corrupting 5% of them by multiplying their features by 100 and multiply their targets by 10,000. From Table 3 below, we see that TERM significantly outperforms the baseline objectives in the noisy regime on both datasets.

Table 3: An alternative noise setup involving both feature and label noise. Similarly, TERM with $t=-2$ significantly outperforms several baseline objectives for noisy outlier mitigation.

objectives	**test RMSE** (`cal-housing`)		**test RMSE** (`abalone`)	
	clean	noisy	clean	noisy
ERM	0.766 (0.023)	239 (9)	2.444 (0.105)	1013 (72)
L_1	0.759 (0.019)	139 (11)	2.435 (0.021)	1008 (117)
Huber (Huber, 1964)	0.762 (0.009)	163 (7)	2.449 (0.018)	922 (45)
CRR (Bhatia et al., 2017)	0.766 (0.024)	245 (8)	2.444 (0.021)	986 (146)
TERM	0.745 (0.007)	**0.753** (0.016)	2.477 (0.041)	**2.449** (0.028)
Genie ERM	0.766 (0.023)	0.766 (0.028)	2.444 (0.105)	2.450 (0.109)

Unstructured random v.s. adversarial noise. As a word of caution, we note that the experiments thus far have focused on random noise. This makes it possible for the methods to find the underlying structure of clean data even if the majority of the samples are noisy outliers. To gain more intuition on these cases, we generate synthetic two-dimensional data points and test the performance of TERM under 0%, 20%, 40%, and 80% noise for linear regression. TERM with $t=-2$ performs well in all noise levels (Figure 6 and 7). However, as one might expect, TERM with negative t's could potentially overfit to outliers if they are constructed in an adversarial way. In the examples shown in Figure 8, under 40% noise and 80% noise, TERM has a high error measured on the clean data (green dots).

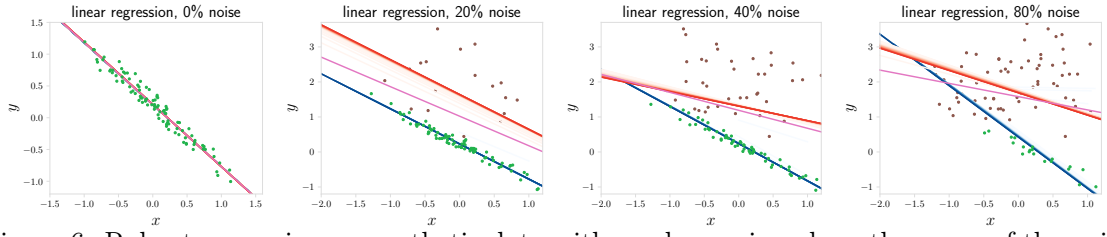


Figure 6: Robust regression on synthetic data with random noise where the mean of the noisy samples is different from that of clean ones. TERM with negative t's (blue, $t=-2$) can fit structured clean data at all noise levels, while ERM (purple) and TERM with positive t's (red) overfit to corrupted data. We color inliers in green and outliers in brown for visualization.

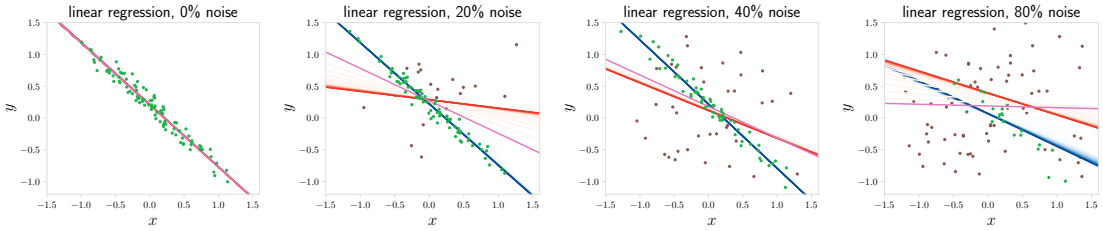


Figure 7: In the presence of random noise with the same mean as that of clean data, TERM with negative t's (blue) can still surpass outliers in all cases, while ERM (purple) and TERM with positive t's (red) overfit to corrupted data. While the performance drops for 80% noise, TERM can still learn useful information, and achieves much lower error than ERM.

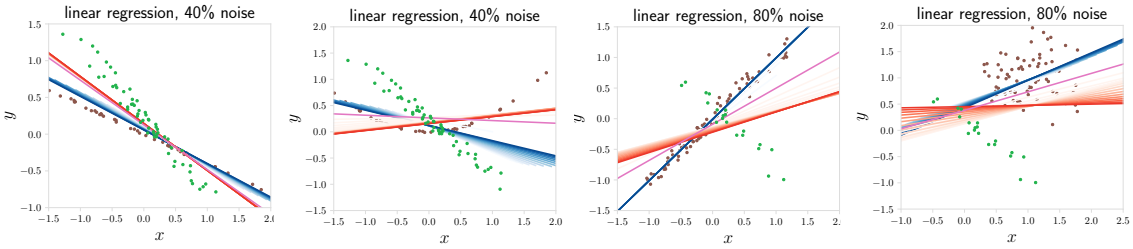


Figure 8: TERM with negative t's (blue) cannot fit clean data if the noisy samples (brown) are adversarial or structured in a manner that differs substantially from the underlying true distribution.

7.1.2 Robust Classification

Deep neural networks can easily overfit to corrupted labels (e.g., Zhang et al., 2017). While the theoretical properties we study for TERM (Section 2) do not directly cover objectives with neural network function approximations, we show that TERM can be applied empirically to DNNs to achieve robustness to noisy training labels. MentorNet (Jiang et al., 2018) is a popular method in this setting, which learns to assign weights to samples based on feedback from a student net. Following the setup in Jiang et al. (2018), we explore classification on CIFAR10 (Krizhevsky et al., 2009) when a fraction of the training labels are corrupted with uniform noise—comparing TERM with ERM and several state-of-the-art approaches (Kumar et al., 2010; Ren et al., 2018; Zhang and Sabuncu, 2018; Krizhevsky et al., 2009). As shown in Table 4, TERM performs competitively with 20% noise, and outperforms all baselines in the high noise regimes. We use MentorNet-PD as a baseline since it does not require clean validation data. In Appendix D.1, we show that TERM also matches the performance of MentorNet-DD, which requires clean validation data. To help reason about the performance of TERM, we also explore a simpler, two-dimensional logistic regression problem in Figure 19, Appendix D.1, finding that TERM with $t=-2$ is similarly robust across the considered noise regimes.

Table 4: TERM is competitive with robust *classification* baselines, and is superior in high noise regimes.

objectives	**test accuracy** (`CIFAR10`, Inception)		
	20% noise	40% noise	80% noise
ERM	0.775 (.004)	0.719 (.004)	0.284 (.004)
RandomRect (Ren et al., 2018)	0.744 (.004)	0.699 (.005)	0.384 (.005)
SelfPaced (Kumar et al., 2010)	0.784 (.004)	0.733 (.004)	0.272 (.004)
MentorNet-PD (Jiang et al., 2018)	0.798 (.004)	0.731 (.004)	0.312 (.005)
GCE (Zhang and Sabuncu, 2018)	**0.805** (.004)	0.750 (.004)	0.433 (.005)
TERM	0.795 (.004)	**0.768** (.004)	**0.455** (.005)
Genie ERM	0.828 (.004)	0.820 (.004)	0.792 (.004)

7.1.3 Low-Quality Annotators

It is not uncommon for practitioners to obtain human-labeled data for their learning tasks from crowd-sourcing platforms. However, these labels are usually noisy in part due to the varying quality of the human annotators. Given a collection of labeled samples from crowd-workers, we aim to learn statistical models that are robust to the potentially low-quality annotators. As a case study, following the setup of (Khetan et al., 2018), we take the CIFAR-10 dataset and simulate 100 annotators where 20 of them are *hammers* (i.e., always correct) and 80 of them are *spammers* (i.e., assigning labels uniformly at random). We apply TERM at the annotator group level in (83), which is equivalent to assigning annotator-level weights based on the aggregate value of their loss. As shown in Figure 9, TERM is able to achieve the test accuracy limit set by *Genie ERM*, i.e., *the ideal performance obtained by completely removing the known outliers*. We note in particular that the accuracy reported by (Khetan et al., 2018) (0.777) is lower than TERM (0.825) in the same setup, even though their approach is a two-pass algorithm requiring at least to double the training time. We provide full empirical details and investigate additional noisy annotator scenarios in Appendix D.1.

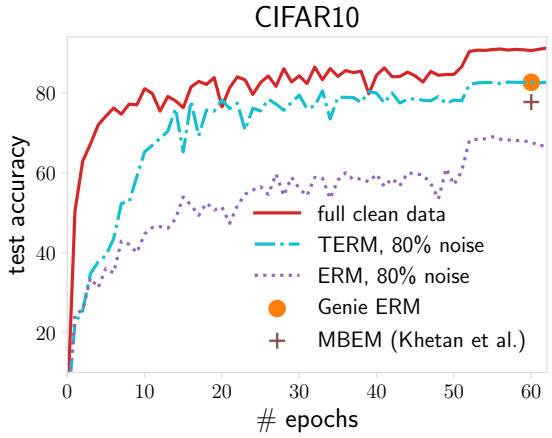


Figure 9: TERM (t=−2) completely removes the impact of noisy annotators, reaching the performance limit set by Genie ERM.

7.2 Fairness and Generalization (t>0)

In this section, we show that positive values of t in TERM can help promote fairness via learning fair representations and enforcing fairness during optimization, and offer variance reduction for better generalization.

7.2.1 Fair Principal Component Analysis (PCA)

We explore the flexibility of TERM in learning fair representations using PCA. In fair PCA, the goal is to learn low-dimensional representations which are fair to all considered subgroups (e.g., yielding similar reconstruction errors) (Samadi et al., 2018; Tantipongpipat et al., 2019; Kamani et al., 2019). Despite the non-convexity of the fair PCA problem, we apply

TERM to this task, referring to the resulting objective as TERM-PCA. We tilt the same loss function as in Samadi et al. (2018): $f(X;U)=\frac{1}{|X|}\left(\|X-XUU^\top\|_F^2-\|X-\hat{X}\|_F^2\right)$, where $X\in\mathbb{R}^{n\times d}$ is a subset (group) of data, $U\in\mathbb{R}^{d\times r}$ is the current projection, and $\hat{X}\in\mathbb{R}^{n\times d}$ is the optimal rank-r approximation of X. Instead of solving a more complex min-max problem using semi-definite programming as in Samadi et al. (2018), which scales poorly with problem dimension, we apply gradient-based methods, re-weighting the gradients at each iteration based on the loss on each group. In Figure 10, we plot the aggregate loss for two groups (high vs. low education) in the Default Credit dataset (Yeh and Lien, 2009) for different target dimensions r. By varying t, we achieve varying degrees of performance improvement on different groups—TERM (t=200) recovers the min-max results of (Samadi et al., 2018) by forcing the losses on both groups to be (almost) identical, while TERM (t=10) offers the flexibility of reducing the performance gap less aggressively. We also provide convergence plots for different values of t in this application (Figure 11), and observe slower convergence for larger values of t, which is consistent with our analyses in Section 2 and 5. However, we do not observe exponential dependence on t from the convergence curves, which suggest that the theoretical dependence on t in convergence proofs for the solvers may be an artifact of our proof techniques, and might possibly be further improved by other analysis techniques for typical practical use cases.

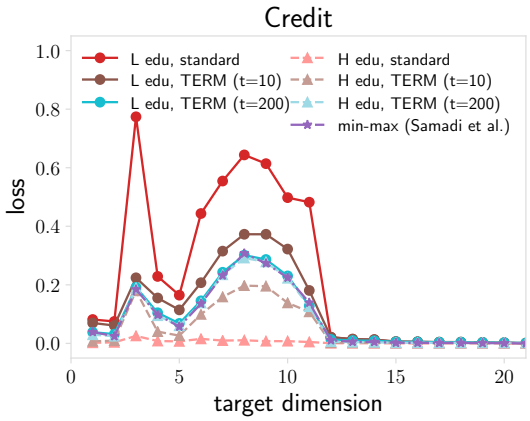


Figure 10: TERM-PCA flexibly trades the performance on the high (H) edu group for the performance on the low (L) edu group.

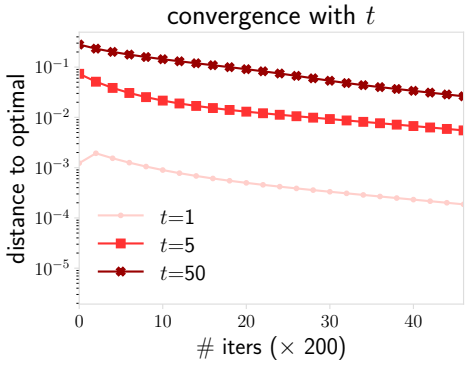


Figure 11: Convergence of TERM with respect to t in fair PCA (target dimension=7). We tune optimal learning rates separately for each t. As t increases, the convergence becomes slower, which validates our analyses in Section 2 and 5.

7.2.2 Fair Federated Learning

Federated learning involves learning statistical models across massively distributed networks of remote devices or isolated organizations (McMahan et al., 2017; Li et al., 2020a). Ensuring fair (i.e., uniform) performance distribution across the devices is a major concern in federated settings (Mohri et al., 2019; Li et al., 2020b), as using current approaches for federated learning (FedAvg (McMahan et al., 2017)) may result in highly variable performance across

the network. Li et al. (2020b) consider solving an alternate objective for federated learning, called q-FFL, to dynamically emphasize the worst-performing devices, which is conceptually similar to the goal of TERM, though it is applied specifically to the problem of federated learning and limited to the case of positive t. Here, we compare TERM with q-FFL in their setup on the vehicle dataset (Duarte and Hu, 2004) consisting of data collected from 23 distributed sensors (hence 23 devices). We tilt the L_2 regularized linear SVM objective at the device level. At each communication round, we re-weight the accumulated local model updates from each selected device based on the weights estimated via Algorithm 4. From Figure 12, we see that similar to q-FFL, TERM (t=0.1) can also significantly promote the accuracy on the worst device while maintaining the overall performance. The statistics of the accuracy distribution are reported in Table 5 below.

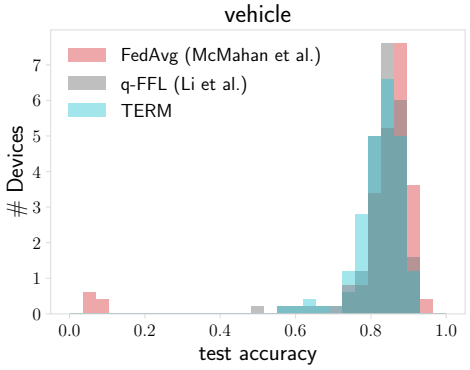


Figure 12: TERM FL (t=0.1) significantly increases the accuracy on the worst-performing device (similar to q-FFL) while obtaining a similar average accuracy.

Table 5: Both q-FFL and TERM can encourage more uniform accuracy distributions across the devices in federated networks while maintaining similar average performance. Numbers in the parentheses correspond to the standard error of each metric across 5 runs.

objectives	**test accuracy**		
	average	worst 10%	stdev
FedAvg	0.853 (.078)	0.421 (.007)	0.173 (.001)
q-FFL (q=5)	0.862 (.029)	**0.704** (.033)	**0.064** (.005)
TERM (t=0.1)	0.853 (.027)	**0.707** (.009)	**0.061** (.003)

7.2.3 Fair Meta-Learning

Meta-learning aims to learn a shared initialization across all tasks such that the initialization can quickly adapt to unseen tasks (i.e., meta-testing tasks) using a few samples. In practice, the resulting performance across meta-testing tasks can vary due to different data distributions associated with these tasks. One of the popular meta-learning methods is MAML (Finn et al., 2017), whose objective is to minimize the sum of empirical losses across tasks $\{\mathcal{T}_i\}$ generated from $p(\mathcal{T})$ after one step of adaptation, i.e., $\min_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} f(\mathcal{T}_i; \theta - \alpha \nabla_\theta f(\mathcal{T}_i; \theta))$. Previous works have proposed a min-max variant of MAML to encourage a more fair (uniform) performance distribution by optimizing the worst meta-training task called TR-MAML (Collins et al., 2020). We apply TERM to MAML by replacing the ERM formulation with tilted losses. Following the setup in Collins et al. (2020), we evaluate TERM on the popular sin wave regression problem. For a fair comparison, we perform task-level tilting for TERM, and operates on task-level reweighting for TR-MAML. From Table 6, we see that TERM with t=2 not only decreases the standard deviation of test errors, but also achieves lower mean errors than MAML. As the number of tasks is large (5,000), solving the min-max variant (TR-MAML) is challenging, and results in slightly worse performance than TERM.

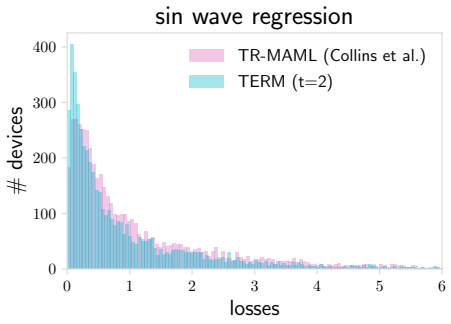


Figure 13: Loss distribution of TERM compared with the TR-MAML baseline.

Table 6: TERM (t=2) results in fairer and lower test errors across meta-test tasks after adaptation compared with MAML (Finn et al., 2017). TERM also outperforms a recently proposed min-max task-robust MAML method (TR-MAML) (Collins et al., 2020).

methods	mean	std	max	worst 10%
MAML	1.23	1.63	19.1	5.16
TR-MAML	1.25	1.51	14.31	4.85
TERM (t=2)	**1.14**	**1.33**	**13.59**	**4.29**

7.2.4 Handling Class Imbalance

Next, we show that TERM can reduce the performance variance across classes with extremely imbalanced data when training deep neural networks. We compare TERM with several baselines which re-weight samples during training, including assigning weights inversely proportional to the class size (InverseRatio), focal loss (Lin et al., 2017), HardMine (Malisiewicz et al., 2011), and LearnReweight (Ren et al., 2018). Following the setting of Ren et al. (2018), the datasets are composed of imbalanced 4 and 9 digits from MNIST (LeCun et al., 1998). In Figure 14, we see that TERM obtains similar (or higher) final accuracy on the clean test data as the state-of-the-art methods. We note that compared with LearnReweight, which optimizes the model over an additional balanced validation set and requires three gradient calculations for each update, TERM neither requires such balanced validation data nor does it increase the per-iteration complexity.

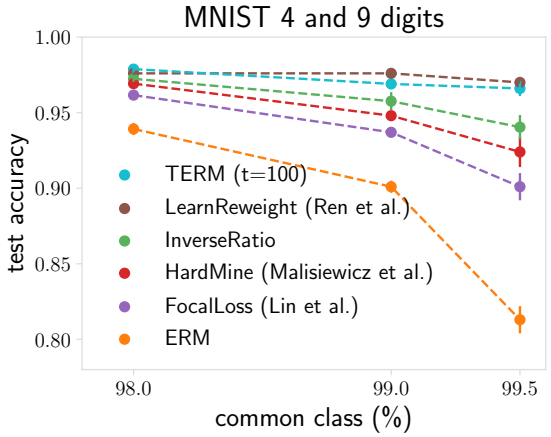


Figure 14: TERM (t=100) is competitive with state-of-the-art methods for classification with imbalanced classes.

7.2.5 Improving Generalization via Variance Reduction

A common alternative to ERM is to consider a distributionally robust objective, which optimizes for the worst-case training loss over a set of distributions, and has been shown to offer variance-reduction properties that benefit generalization (e.g., Duchi and Namkoong, 2019; Sinha et al., 2018; Chen and Paschalidis, 2020; Duchi and Namkoong, 2018). While not directly developed for distributional robustness, TERM also enables variance reduction for positive values of t (Theorem 3), which can be used to strike a better bias-variance trade-off for generalization. We compare TERM with several baselines including robustly regularized risk (RobustRegRisk) (Duchi and Namkoong, 2019), linear SVM (Ren et al., 2018), Conditional Value-at-Risk (CVaR) (Rockafellar et al., 2000; Soma and Yoshida, 2020), LearnRewight (Ren et al., 2018), FocalLoss (Lin et al., 2017), and HRM (Leqi et al., 2019) on the HIV-1 dataset (Rögnvaldsson, 2013; Dua and Graff, 2019) originally investigated by Duchi and Namkoong (2019). We examine the accuracy on the rare class ($Y=0$), the common class ($Y=1$), and overall accuracy.

The mean and standard error of accuracies are reported in Table 7. RobustRegRisk and TERM offer similar performance improvements compared with other baselines, such as linear SVM, CVaR, LearnRewight, FocalLoss, and HRM. Note that here RobustRegRisk (Duchi and Namkoong, 2019) and CVaR (Rockafellar et al., 2000) can both be viewed as specific instances of the distributionally robust optimization framework, with different uncertainty sets. For larger t, TERM achieves similar accuracy in both classes, while RobustRegRisk does not show similar trends by sweeping its hyperparameters. It is common to adjust the decision threshold to boost the accuracy on the rare class. We do this for ERM and RobustRegRisk and optimize the threshold so that ERM_+ and RobustRegRisk_+ result in the same validation accuracy on the rare class as TERM (t=50). TERM achieves similar performance to RobustRegRisk_+, without the need for an extra tuned hyperparameter.

Table 7: TERM (t=0.1) is competitive with strong baselines in generalization. TERM (t=50) outperforms ERM_+ (with decision threshold changed for providing fairness) and is competitive with RobustRegRisk_+ with no need for extra hyperparameter tuning.

objectives	**accuracy** ($Y=0$)		**accuracy** ($Y=1$)		**overall accuracy** (%)	
	train	test	train	test	train	test
ERM	0.841 (.005)	0.822 (.009)	0.971 (.000)	0.966 (.002)	0.944 (.000)	0.934 (.003)
Linear SVM	0.873 (.003)	**0.838** (.013)	0.965 (.000)	0.964 (.002)	0.951 (.001)	**0.937** (.004)
CVaR (Rockafellar et al., 2000)	0.877 (.004)	**0.844** (.013)	0.972 (.000)	0.964 (.003)	0.952 (.001)	**0.937** (.003)
LearnReweight (Ren et al., 2018)	0.860 (.004)	**0.841** (.014)	0.960 (.002)	0.961 (.004)	0.940 (.001)	0.934 (.004)
FocalLoss (Lin et al., 2017)	0.871 (.003)	**0.834** (.013)	0.970 (.000)	0.966 (.003)	0.949 (.001)	**0.937** (.004)
HRM (Leqi et al., 2019)	0.875 (.003)	**0.839** (.012)	0.972 (.000)	0.965 (.003)	0.952 (.001)	**0.937** (.003)
RobustRegRisk (Duchi et al., 2019)	0.875 (.003)	**0.844** (.010)	0.971 (.000)	0.966 (.003)	0.951 (.001)	**0.939** (.004)
TERM (t=0.1)	0.864 (.003)	**0.840** (.011)	0.970 (.000)	0.964 (.003)	0.949 (.001)	**0.937** (.004)
ERM_+ (thresh = 0.26)	0.943 (.001)	0.916 (.008)	0.919 (.001)	0.917 (.003)	0.924 (.001)	0.917 (.002)
RobustRegRisk_+ (thresh=0.49)	0.943 (.000)	0.917 (.005)	0.928 (.001)	**0.928** (.002)	0.931 (.001)	**0.924** (.001)
TERM (t=50)	0.942 (.001)	0.917 (.005)	0.926 (.001)	**0.925** (.002)	0.929 (.001)	**0.924** (.001)

7.3 Solving Compound Issues: Hierarchical Multi-Objective Tilting

Finally, in this section, we focus on settings where multiple issues, e.g., class imbalance and label noise, exist in the data simultaneously. We discuss two possible instances of hierarchical multi-objective TERM to tackle such problems. One can think of other variants in this hierarchical tilting space which could be useful depending on applications at hand.

7.3.1 Class Imbalance and Random Noise

We explore the HIV-1 dataset (Rögnvaldsson, 2013), as in Section 7.2. We report both overall accuracy and accuracy on the rare class in four scenarios: **(a) clean and 1:4**, the original dataset that is naturally slightly imbalanced with rare samples represented 1:4 with respect to the common class; **(b) clean and 1:20**, where we subsample to introduce a 1:20 imbalance ratio; **(c) noisy and 1:4**, which is the original dataset with labels associated with 30% of the samples randomly reshuffled; and **(d) noisy and 1:20**, where 30% of the labels of the 1:20 imbalanced dataset are reshuffled.

Table 8: Hierarchical TERM can address both class imbalance and noisy samples.

objectives	**test accuracy** (HIV-1)							
	clean data				30% noise			
	1:4		1:20		1:4		1:20	
	Y=0	overall	Y=0	overall	Y=0	overall	Y=0	overall
ERM	0.822 (.009)	0.934 (.003)	0.503 (.013)	0.888 (.006)	0.656 (.014)	0.911 (.006)	0.240 (.018)	0.831 (.011)
CVaR (Rockafellar et al., 2000)	**0.844** (.013)	**0.937** (.003)	0.621 (.011)	0.906 (.005)	0.651 (.015)	0.909 (.006)	0.252 (.014)	0.834 (.010)
GCE (Zhang and Sabuncu, 2018)	0.822 (.009)	0.934 (.003)	0.503 (.013)	0.888 (.006)	0.732 (.021)	**0.925** (.005)	0.324 (.017)	0.849 (.008)
LearnReweight (Ren et al., 2018)	**0.841** (.014)	0.934 (.004)	0.800 (.022)	0.904 (.003)	0.721 (.034)	0.856 (.008)	0.532 (.054)	0.856 (.013)
RobustRegRisk (Duchi et al., 2019)	**0.844** (.010)	**0.939** (.004)	0.622 (.011)	0.906 (.005)	0.634 (.014)	0.907 (.006)	0.051 (.014)	0.792 (.012)
FocalLoss (Lin et al., 2017)	**0.834** (.013)	**0.937** (.004)	**0.806** (.020)	**0.918** (.003)	0.638 (.008)	0.908 (.005)	0.565 (.027)	**0.890** (.009)
HAR (Cao et al., 2021)	**0.842** (.011)	0.936 (.004)	0.817 (.013)	**0.926** (.004)	**0.870** (.010)	0.915 (.004)	**0.800** (.016)	0.867 (.012)
TERM$_{sc}$	**0.840** (.010)	**0.937** (.004)	**0.836** (.018)	**0.921** (.002)	**0.852** (.010)	**0.924** (.004)	**0.778** (.008)	**0.900** (.005)
TERM$_{ca}$	**0.844** (.014)	**0.938** (.004)	**0.834** (.021)	**0.918** (.003)	**0.846** (.015)	**0.933** (.003)	**0.806** (.020)	**0.901** (.010)

In Table 8, hierarchical TERM is applied at the sample level and class level (TERM$_{sc}$), where we use the sample-level tilt of τ=-2 for noisy data. We use class-level tilt of t=0.1 for the 1:4 case and t=50 for the 1:20 case. We compare against baselines for robust classification and class imbalance (discussed previously in Sections 7.1 and 7.2), where we tune them for best performance (Appendix D.2). Similar to the experiments in Section 7.1, we avoid using baselines that require clean validation data (e.g., Roh et al., 2020). We compare TERM with an additional baseline of HAR (Cao et al., 2021), a recent work addressing the issues of noisy and rare samples simultaneously with adaptive Lipschitz regularization. While different baselines (except HAR) perform well in their respective problem settings, TERM and HAR are far superior to all baselines when considering noisy samples and class imbalance simultaneously (rightmost column in Table 8). Finally, in the last row of Table 8, we simulate the noisy annotator setting of Section 7.1.3 assuming that the data is coming from 10 annotators, i.e., in the 30% noise case we have 7 hammers and 3 spammers. In this case, we apply hierarchical TERM at both class and annotator levels (TERM$_{ca}$), where we perform the higher level tilt at the annotator (group) level and the lower level tilt at the class level (with no sample-level tilting). We show that this approach can benefit noisy/imbalanced

data even further (far right, Table 8), while suffering only a small performance drop on the clean and noiseless data (far left, Table 8).

7.3.2 Class Imbalance and Adversarial Noise

We evaluate hierarchical tilting on a more difficult task involving more adversarial noise with deep neural network models. We take the setup studied in Cao et al. (2021). The noise is created by exchanging labels of 40% samples which come from similar classes ('cat' and 'dog', 'vehicle' and 'automobile') in the CIFAR10 dataset. To simulate class imbalance, only 10% of the training data from these four noisy classes are subsampled. For TERM, we apply group-level positive tilting by linearly scaling t from 0 to 3, and perform sample-level negative tilting within each class with τ scaling from 0 to -2. Table 9 reports the results of hierarchical TERM (TERM$_{sc}$) compared with HAR (Cao et al., 2021) and other baselines. We see that TERM underperforms HAR, and outperforms all other approaches. Note that HAR is a more complicated method which requires to perform end-to-end training for two times with higher per-iteration complexity (involving second-order information), while TERM is a simple method and enjoys the same training time as that of ERM on this problem.

Table 9: TERM outperforms most baselines addressing the co-existence of noisy samples and class imbalance by a large margin, and is worse than a more complicated method HAR.

objectives	**test accuracy** (`CIFAR10`, ResNet32)	
	noisy, rare class	clean, common class
ERM	0.529 (.012)	**0.944** (.001)
GCE (Zhang and Sabuncu, 2018)	0.482 (.006)	0.916 (.003)
MentorNet (Jiang et al., 2018)	0.541 (.010)	0.903 (.005)
MW-Net (Shu et al., 2019b)	0.554 (.011)	0.917 (.005)
HAR (Cao et al., 2021)	**0.635** (.008)	**0.943** (.002)
TERM$_{sc}$	0.585 (.014)	0.913 (.003)

8. Related Approaches in Machine Learning

Here we discuss related problem-specific works in machine learning addressing deficiencies of ERM. We roughly group them into alternate aggregation schemes, alternate loss functions, and sample re-weighting schemes.

Alternate aggregation schemes. A common alternative to the standard average loss in empirical risk minimization is to consider a min-max objective, which aims to minimize the max-loss. Min-max objectives are commonplace in machine learning, and have been used for a wide range of applications, such as ensuring fairness across subgroups (Hashimoto et al., 2018; Mohri et al., 2019; Stelmakh et al., 2019; Samadi et al., 2018; Tantipongpipat et al., 2019; Lahoti et al., 2020), enabling robustness under small perturbations (Sinha et al., 2018), or generalizing to unseen domains (Volpi et al., 2018). As discussed in Section 2, the TERM objective can be viewed as a minimax smoothing (Kort and Bertsekas, 1972; Pee and Royset, 2011) with the added flexibility of a tunable t to allow the user to optimize

utility for different quantiles of loss similar to superquantile approaches (Rockafellar et al., 2000; Laguel et al., 2021), directly trading off between robustness/fairness and utility for positive and negative values of t (see Section 2 for these connections). However, the TERM objective remains smooth (and efficiently solvable) for moderate values of t, resulting in faster convergence even when the resulting solutions are effectively the same as the min-max solution or other desired quantiles of the loss (as we demonstrate in the experiments of Section 7). Interestingly, Cohen et al. introduce Simnets (Cohen and Shashua, 2014; Cohen et al., 2016), with a similar exponential smoothing operator, though for a differing purpose of achieving layer-wise operations *between* sum and max in deep neural networks.

Alternate loss functions. Rather than modifying the way the losses are aggregated, as in (smoothed) min-max or superquantile methods, it is also quite common to modify the losses themselves. For example, in robust regression, it is common to consider losses such as the L_1 loss, Huber loss, or general M-estimators (Holland and Ikeda, 2019) as a way to mitigate the effect of outliers (Bhatia et al., 2015). Wang et al. (2013) study a similar exponentially tilted loss for robust regression and characterize the break down point, though it is limited to the squared loss and only corresponds to $t<0$. Losses can also be modified to address outliers by favoring small losses (Yu et al., 2012; Zhang and Sabuncu, 2018) or gradient clipping (Menon et al., 2020). Some works mitigate label noise by explicitly modeling noise distributions into end-to-end training combined with an additional noise model regularizer (Jindal et al., 2016, 2019). On the other extreme, the largest losses can be magnified to encourage focus on hard samples (Lin et al., 2017; Wang et al., 2016b; Li et al., 2020b), which is a popular approach for curriculum learning. Constraints could also be imposed to promote fairness during the optimization procedure (Hardt et al., 2016; Donini et al., 2018; Rezaei et al., 2020; Zafar et al., 2017; Baharlouei et al., 2020; Cotter et al., 2019; Lowy et al., 2021; Alghamdi et al., 2020; Zafar et al., 2019; Prost et al., 2019). A line of work proposes α-loss, which is able to promote fairness or robustness for classification tasks (Sypherd et al., 2019). Ignoring the log portion of the objective in (2), TERM can be viewed as an alternate loss function exponentially shaping the loss to achieve both of these goals with a single objective, i.e., magnifying hard examples with $t>0$ and suppressing outliers with $t<0$. In addition, we show that TERM can even achieve both goals simultaneously with hierarchical multi-objective optimization (Section 7.3).

Sample re-weighting schemes. There exist approaches that implicitly modify the underlying ERM objective by re-weighting the influence of the samples themselves. These re-weighting schemes can be enforced in many ways. A simple and widely used example is to subsample training points in different classes. Alternatively, one can re-weight examples according to their loss function when using a stochastic optimizer, which can be used to put more emphasis on "hard" or "unfair" examples (Shrivastava et al., 2016; Jiang et al., 2019; Katharopoulos and Fleuret, 2017; Leqi et al., 2019; Abernethy et al., 2022). Re-weighting can also be implicitly enforced via the inclusion of a regularization parameter (Abdelkarim et al., 2020), loss clipping (Yang et al., 2010), or modelling crowd-worker qualities (Khetan et al., 2018). Such an explicit re-weighting has been explored for other applications (e.g., Lin et al., 2017; Jiang et al., 2018; Shu et al., 2019a; Chang et al., 2017; Gao et al., 2015; Ren et al., 2018), though in contrast to these methods, TERM is applicable to a general class of loss functions, with theoretical guarantees. TERM is equivalent to a dynamic re-weighting

of the samples based on the values of the objectives (Lemma 5), which could be viewed as a convexified version of loss clipping. We note that such view holds more generally for all distributionally robust objectives (Słowik and Bottou, 2022). We compare to several sample re-weighting schemes empirically in Section 7.

9. Discussion and Conclusion

In this manuscript, we have explored the use of exponential tilting in risk minimization, examining tilted empirical risk minimization (TERM) as a flexible extension to the ERM framework. We rigorously established connections between TERM and related objectives including VaR, CVaR, and DRO. We explored, both theoretically and empirically, TERM's ability to handle various known issues with ERM, such as robustness to noise, class imbalance, fairness, and generalization, as well as more complex issues like the simultaneous existence of class imbalance and noisy outliers. Despite the straightforward modification TERM makes to traditional ERM objectives, the framework consistently outperforms ERM and delivers competitive performance with state-of-the-art, problem-specific methods on a wide range of applications.

Our work highlights the effectiveness and versatility of tilted objectives in machine learning. As such, our framework (TERM) could be widely used for applications both positive and negative. However, our hope is that the TERM framework will allow machine learning practitioners to easily modify the ERM objective to handle practical concerns such as enforcing fairness amongst subgroups, mitigating the effect of outliers, and ensuring robust performance on new, unseen data. One potential downside of the TERM objective is that if the underlying dataset is *not* well-understood, incorrectly tuning t could have the unintended consequence of *magnifying* the impact of biased/corrupted data in comparison to traditional ERM. Indeed, critical to the success of such a framework is understanding the implications of the modified objective, both theoretically and empirically. The goal of this work is therefore to explore these implications so that it is clear when such a modified objective would be appropriate.

In terms of the use-cases explored with the TERM framework, we relied on benchmark datasets that have been commonly explored in prior work (e.g., Yang et al., 2010; Samadi et al., 2018; Tantipongpipat et al., 2019; Yu et al., 2012). However, we note that some of these common benchmarks, such as cal-housing (Pace and Barry, 1997) and Credit (Yeh and Lien, 2009), contain potentially sensitive information. While the goal of our experiments was to showcase that the TERM framework could be useful in learning fair representations that suppress membership bias and hence promote fairer performance, developing an understanding for—and removing—such membership biases requires a more comprehensive treatment of the problem that is outside the scope of this work.

In the future, in addition to generalization bounds of TERM, it would be interesting to further explore applications of tilted losses in machine learning. We note that since the early TERM work (Li et al., 2021) was made public, there are several subsequent works applying (variants of) TERM to handle other real-world ML applications (Szabo et al., 2021; Zhou et al., 2021), or exploring risk bounds on differential private TERM (Lowy and Razaviyayn, 2021), which suggest rich implications and wide applicability of TERM, beyond what is studied in this work.

Appendix

In this appendix, we provide full statements and proofs of the analyses presented in Section 2-Section 4 (Appendix A and B); details and convergence proof on the methods we propose for solving TERM (Appendix C), and complete empirical results and details of our empirical setup (Appendix D). We provide a table of contents below for easier navigation.

Contents

Appendix A. Properties and Interpretations (Proofs and Additional Results)

In this section, we provide the proofs of the main results in the paper, along with additional results on the properties of TERM objective, its solution, as well as the corresponding solvers.

A.1 Proofs of Basic Properties of the TERM Objective

We first provide proofs for the basic properties of the TERM objective.

Proof of Lemma 1. The conclusion follows by noting that for any $\theta_1,\theta_2\in\Theta$,

$$\left|\widetilde{R}(t;\theta_1)-\widetilde{R}(t;\theta_2)\right|=\left|\frac{1}{t}\log\left(\frac{1}{N}\sum_{i\in[N]}e^{tf(x_i;\theta_1)}\right)-\frac{1}{t}\log\left(\frac{1}{N}\sum_{i\in[N]}e^{tf(x_i;\theta_2)}\right)\right| \tag{87}$$

$$=\left|\frac{1}{t}\log\left(\frac{\sum_{i\in[N]}e^{tf(x_i;\theta_1)}}{\sum_{i\in[N]}e^{tf(x_i;\theta_2)}}\right)\right| \tag{88}$$

$$\leqslant\left|\frac{1}{t}\log\left(e^{tL\|\theta_1-\theta_2\|_2}\frac{\sum_{i\in[N]}e^{tf(x_i;\theta_2)}}{\sum_{i\in[N]}e^{tf(x_i;\theta_2)}}\right)\right| \tag{89}$$

$$=L\|\theta_1-\theta_2\|_2. \tag{90}$$

□

Proof of Lemma 2. Recall that

$$\nabla_\theta\widetilde{R}(t;\theta)=\frac{\sum_{i\in[N]}\nabla_\theta f(x_i;\theta)e^{tf(x_i;\theta)}}{\sum_{i\in[N]}e^{tf(x_i;\theta)}} \tag{91}$$

$$=\frac{1}{N}\sum_{i\in[N]}\nabla_\theta f(x_i;\theta)e^{t(f(x_i;\theta)-\widetilde{R}(t;\theta))}. \tag{92}$$

The proof of the first part is completed by differentiating again with respect to θ, followed by algebraic manipulation. To prove the second part, notice that the term in (13) is positive semi-definite, whereas the term in (14) is positive definite and lower bounded by $\beta_{\min}\mathbf{I}$ (see Assumption 2, Eq. (6)). □

Proof of Lemma 3. Let us first provide a proof for $t\in\mathbb{R}^-$. Invoking Lemma 2 and Weyl's inequality (Weyl, 1912), we have

$$\lambda_{\max}\left(\nabla^2_{\theta\theta^\top}\widetilde{R}(t;\theta)\right)$$

$$\leqslant\lambda_{\max}\left(\frac{t}{N}\sum_{i\in[N]}(\nabla_\theta f(x_i;\theta)-\nabla_\theta\widetilde{R}(t;\theta))(\nabla_\theta f(x_i;\theta)-\nabla_\theta\widetilde{R}(t;\theta))^\top e^{t(f(x_i;\theta)-\widetilde{R}(t;\theta))}\right) \tag{93}$$

$$+\lambda_{\max}\left(\frac{1}{N}\sum_{i\in[N]}\nabla^2_{\theta\theta^\top}f(x_i;\theta)e^{t(f(x_i;\theta)-\widetilde{R}(t;\theta))}\right) \tag{94}$$

$$\leqslant\beta_{\max}, \tag{95}$$

where we have used the fact that the term in (13) is negative semi-definite for $t<0$, and that the term in (14) is positive definite for all t with smoothness bounded by $\beta_{\max}$ (which would hold from smoothness of $f(x_i;\theta)$; see Assumption 2, Eq. (6)).

For $t\in\mathbb{R}^{>0}$, following Lemma 2 and Weyl's inequality (Weyl, 1912), we have

$$\left(\frac{1}{t}\right)\lambda_{\max}\left(\nabla^2_{\theta\theta^\top}\widetilde{R}(t;\theta)\right)$$

$$\leqslant\lambda_{\max}\left(\frac{1}{N}\sum_{i\in[N]}(\nabla_\theta f(x_i;\theta)-\nabla_\theta\widetilde{R}(t;\theta))(\nabla_\theta f(x_i;\theta)-\nabla_\theta\widetilde{R}(t;\theta))^\top e^{t(f(x_i;\theta)-\widetilde{R}(t;\theta))}\right) \tag{96}$$

$$+\left(\frac{1}{t}\right)\lambda_{\max}\left(\frac{1}{N}\sum_{i\in[N]}\nabla^2_{\theta\theta^\top}f(x_i;\theta)e^{t(f(x_i;\theta)-\widetilde{R}(t;\theta))}\right). \tag{97}$$

Due to Weyl's inequality, the smoothness of $f(x_i;\theta)$, and the fact that $\frac{1}{N}\sum_{i\in[N]}e^{t(f(x_i;\theta)-\widetilde{R}(t;\theta))}=1$, $\sum_{i\in[N]}\nabla^2_{\theta\theta^\top}f(x_i;\theta)e^{t(f(x_i;\theta)-\widetilde{R}(t;\theta))}$ is bounded. Consequently,

$$\lim_{t\to+\infty}\left(\frac{1}{t}\right)\lambda_{\max}\left(\nabla^2_{\theta\theta^\top}\widetilde{R}(t;\theta)\right)<+\infty. \tag{98}$$

On the other hand, following Weyl's inequality (Weyl, 1912),

$$\lambda_{\max}\left(\nabla^2_{\theta\theta^\top}\widetilde{R}(t;\theta)\right)$$

$$\geqslant t\lambda_{\max}\left(\frac{1}{N}\sum_{i\in[N]}(\nabla_\theta f(x_i;\theta)-\nabla_\theta\widetilde{R}(t;\theta))(\nabla_\theta f(x_i;\theta)-\nabla_\theta\widetilde{R}(t;\theta))^\top e^{t(f(x_i;\theta)-\widetilde{R}(t;\theta))}\right), \tag{99}$$

and hence,

$$\lim_{t\to+\infty}\left(\frac{1}{t}\right)\lambda_{\max}\left(\nabla^2_{\theta\theta^\top}\widetilde{R}(t;\theta)\right)>0, \tag{100}$$

where we have used the fact that no solution θ exists that would make all f_i's vanish (Assumption 2). □

Under the strict saddle property (Assumption 4), it is known that gradient-based methods would converge to a local minimum (Ge et al., 2015), i.e., $\breve{\theta}(t)$ would be obtained using gradient descent (GD). The rate of convergence of GD scales linearly with the smoothness parameter of the optimization landscape, which is characterized by Lemma 3.

Proof of Lemma 4. For $t\to0$,

$$\lim_{t\to0}\widetilde{R}(t;\theta)=\lim_{t\to0}\frac{1}{t}\log\left(\frac{1}{N}\sum_{i\in[N]}e^{tf(x_i;\theta)}\right)$$

$$=\lim_{t\to0}\frac{\sum_{i\in[N]}f(x_i;\theta)e^{tf(x_i;\theta)}}{\sum_{i\in[N]}e^{tf(x_i;\theta)}} \tag{101}$$

$$=\frac{1}{N}\sum_{i\in[N]}f(x_i;\theta), \tag{102}$$

where (101) is due to L'Hôpital's rule applied to t as the denominator and $\log\left(\frac{1}{N}\sum_{i\in[N]}e^{tf(x_i;\theta)}\right)$ as the numerator.

For $t\to-\infty$, we proceed as follows:

$$\begin{aligned}\lim_{t\to-\infty}\widetilde{R}(t;\theta)&=\lim_{t\to-\infty}\frac{1}{t}\log\left(\frac{1}{N}\sum_{i\in[N]}e^{tf(x_i;\theta)}\right)\\&\leqslant\lim_{t\to-\infty}\frac{1}{t}\log\left(\frac{1}{N}\sum_{i\in[N]}e^{t\min_{j\in[N]}f(x_j;\theta)}\right)&(103)\\&=\min_{i\in[N]}f(x_i;\theta).&(104)\end{aligned}$$

On the other hand,

$$\begin{aligned}\lim_{t\to-\infty}\widetilde{R}(t;\theta)&=\lim_{t\to-\infty}\frac{1}{t}\log\left(\frac{1}{N}\sum_{i\in[N]}e^{tf(x_i;\theta)}\right)\\&\geqslant\lim_{t\to-\infty}\frac{1}{t}\log\left(\frac{1}{N}e^{t\min_{j\in[N]}f(x_j;\theta)}\right)&(105)\\&=\min_{i\in[N]}f(x_i;\theta)-\lim_{t\to-\infty}\left\{\frac{1}{t}\log N\right\}&(106)\\&=\min_{i\in[N]}f(x_i;\theta).&(107)\end{aligned}$$

Hence, the proof follows by putting together (104) and (107).

The proof proceeds similarly to $t\to-\infty$ for $t\to+\infty$ and is omitted for brevity. □

A.2 General Properties of the Objective for GLMs

In this section, even if not explicitly stated, all results are derived under Assumption 3 with a generalized linear model and loss function of the form (7), effectively assuming that the loss function is the negative log-likelihood of an exponential family (Wainwright and Jordan, 2008).

Definition 4 (Empirical cumulant generating function)**.** *Let*

$$\widetilde{\Lambda}(t;\theta):=t\widetilde{R}(t;\theta).\tag{108}$$

Definition 5 (Empirical log-partition function (Wainwright et al., 2005))**.** *Let* $\Gamma(t;\theta)$ *be*

$$\Gamma(t;\theta):=\log\left(\frac{1}{N}\sum_{i\in[N]}e^{-t\theta^\top T(x_i)}\right).\tag{109}$$

Thus, we have

$$\widetilde{R}(t;\theta)=A(\theta)+\frac{1}{t}\log\left(\frac{1}{N}\sum_{i\in[N]}e^{-t\theta^\top T(x_i)}\right)=A(\theta)+\frac{1}{t}\Gamma(t;\theta).\tag{110}$$

Definition 6 (Tilted empirical mean and empirical variance of the sufficient statistic)**.** *Let $\mathcal{M}$ and $\mathcal{V}$ denote the mean and the variance of the sufficient statistic, and be given by*

$$\mathcal{M}(t;\theta):=\frac{1}{N}\sum_{i\in[N]} T(x_i)e^{-t\theta^\top T(x_i)-\Gamma(t;\theta)}, \tag{111}$$

$$\mathcal{V}(t;\theta):=\frac{1}{N}\sum_{i\in[N]} (T(x_i)-\mathcal{M}(t;\theta))(T(x_i)-\mathcal{M}(t;\theta))^\top e^{-t\theta^\top T(x_i)-\Gamma(t;\theta)}. \tag{112}$$

We notice that $\mathcal{M}(t;\theta)$ and $\mathcal{V}(t;\theta)$ defined here are equivalent to tilted empirical mean/variance in the main text (Eq. (29) and Eq. (31)) over sufficient statistic, i.e.,

$$\mathcal{M}(t;\theta)=\sum_{i\in[N]} w_i(t;\theta)T(x_i), \tag{113}$$

$$\mathcal{V}(t;\theta)=\sum_{i\in[N]} w_i(t;\theta)(T(x_i)-\mathcal{M}(t;\theta))(T(x_i)-\mathcal{M}(t;\theta))^\top. \tag{114}$$

Similarly, as a special case of t-tilted empirical mean/variance (Eq. (30) and Eq. (32)), t-tilted empirical mean/variance over sufficient statistic are defined as

$$\mathcal{M}_t:=\mathcal{M}(t;\breve{\theta}(t)), \tag{115}$$

$$\mathcal{V}_t:=\mathcal{V}(t;\breve{\theta}(t)). \tag{116}$$

The quantities $\mathcal{M}(t;\theta)$,$\mathcal{V}(t;\theta)$,$\mathcal{M}_t$, and $\mathcal{V}_t$ will be used for proving general properties of TERM solutions in this section.

Lemma 12. *For all $t\in\mathbb{R}$, we have $\mathcal{V}(t;\theta)\succeq 0$.*

Next we state a few key relationships that we will use in our characterizations. The proofs are straightforward and omitted for brevity.

Lemma 13 (Partial derivatives of Γ)**.** *For all $t\in\mathbb{R}$ and all $\theta\in\Theta$,*

$$\frac{\partial}{\partial t}\Gamma(t;\theta)=-\theta^\top\mathcal{M}(t;\theta), \tag{117}$$

$$\nabla_\theta\Gamma(t;\theta)=-t\mathcal{M}(t;\theta). \tag{118}$$

Lemma 14 (Partial derivatives of $\mathcal{M}$)**.** *For all $t\in\mathbb{R}$ and all $\theta\in\Theta$,*

$$\frac{\partial}{\partial t}\mathcal{M}(t;\theta)=-\mathcal{V}(t;\theta)\theta, \tag{119}$$

$$\nabla_\theta\mathcal{M}(t;\theta)=-t\mathcal{V}(t;\theta). \tag{120}$$

The next few lemmas characterize the partial derivatives of the cumulant generating function.

Lemma 15. *(Derivative of $\widetilde{\Lambda}$ with t) For all $t\in\mathbb{R}$ and all $\theta\in\Theta$,*

$$\frac{\partial}{\partial t}\widetilde{\Lambda}(t;\theta)=A(\theta)-\theta^\top\mathcal{M}(t;\theta). \tag{121}$$

Proof. The proof is carried out by

$$\frac{\partial}{\partial t}\widetilde{\Lambda}(t;\theta)=A(\theta)-\theta^\top\sum_{i\in[N]}T(x_i)e^{-t\theta^\top T(x_i)-\Gamma(t;\theta)}=A(\theta)-\theta^\top\mathcal{M}(t;\theta). \tag{122}$$

□

Lemma 16 (Second derivative of $\widetilde{\Lambda}$ with t)**.** *For all $t\in\mathbb{R}$ and all $\theta\in\Theta$,*

$$\frac{\partial^2}{\partial t^2}\widetilde{\Lambda}(t;\theta)=\theta^\top\mathcal{V}(t;\theta)\theta. \tag{123}$$

Lemma 17 (Gradient of $\widetilde{\Lambda}$ with θ)**.** *For all $t\in\mathbb{R}$ and all $\theta\in\Theta$,*

$$\nabla_\theta\widetilde{\Lambda}(t;\theta)=t\nabla_\theta A(\theta)-t\mathcal{M}(t;\theta). \tag{124}$$

Lemma 18 (Hessian of $\widetilde{\Lambda}$ with θ)**.** *For all $t\in\mathbb{R}$ and all $\theta\in\Theta$,*

$$\nabla^2_{\theta\theta^\top}\widetilde{\Lambda}(t;\theta)=t\nabla^2_{\theta\theta^\top}A(\theta)+t^2\mathcal{V}(t;\theta). \tag{125}$$

Lemma 19 (Gradient of $\widetilde{\Lambda}$ with respect to t and θ)**.** *For all $t\in\mathbb{R}$ and all $\theta\in\Theta$,*

$$\frac{\partial}{\partial t}\nabla_\theta\widetilde{\Lambda}(t;\theta)=\nabla_\theta A(\theta)-\mathcal{M}(t;\theta)+t\mathcal{V}(t;\theta)\theta. \tag{126}$$

Proof of Theorem 1. Following (110),

$$\frac{\partial}{\partial t}\widetilde{R}(t;\theta)=\frac{\partial}{\partial t}\left\{\frac{1}{t}\Gamma(t;\theta)\right\} \tag{127}$$

$$=-\frac{1}{t^2}\Gamma(t;\theta)-\frac{1}{t}\theta^\top\mathcal{M}(t;\theta), \tag{128}$$

$$=:g(t;\theta), \tag{129}$$

where (128) follows from Lemma 13, and (129) defines $g(t;\theta)$.

Let $g(0;\theta):=\lim_{t\to0}g(t;\theta)$ Notice that

$$g(0;\theta)=\lim_{t\to0}\left\{-\frac{1}{t^2}\Gamma(t;\theta)-\frac{1}{t}\theta^\top\mathcal{M}(t;\theta)\right\} \tag{130}$$

$$=-\lim_{t\to0}\left\{\frac{\frac{1}{t}\Gamma(t;\theta)+\theta^\top\mathcal{M}(t;\theta)}{t}\right\} \tag{131}$$

$$=\theta^\top\mathcal{V}(0;\theta)\theta, \tag{132}$$

where (132) is due to L'Hôpital's rule and Lemma 16. Now consider

$$\frac{\partial}{\partial t}\{t^2g(t;\theta)\}=\frac{\partial}{\partial t}\{-\Gamma(t;\theta)-t\theta^\top\mathcal{M}(t;\theta)\} \tag{133}$$

$$=\theta^\top\mathcal{M}(t;\theta) \tag{134}$$

$$-\theta^\top\mathcal{M}(t;\theta)+t\theta^\top\mathcal{V}(t;\theta)\theta \tag{135}$$

$$=t\theta^\top\mathcal{V}(t;\theta)\theta, \tag{136}$$

where $g(t;\theta)=\frac{\partial}{\partial t}\widetilde{R}(t;\theta)$, (134) follows from Lemma 13, (135) follows from the chain rule and Lemma 14. Hence, $t^2g(t;\theta)$ is an increasing function of t for $t\in\mathbb{R}^{>0}$, and a decreasing function of t for $t\in\mathbb{R}^{-}$, taking its minimum at $t=0$. Hence, $t^2g(t;\theta)\geqslant0$ for all $t\in\mathbb{R}$. This implies that $g(t;\theta)\geqslant0$ for all $t\in\mathbb{R}$, which in conjunction with (129) implies the statement of the theorem.

A.3 General Properties of TERM Solutions for GLMs

Next, we characterize some of the general properties of the solutions of TERM objectives. Note that these properties are established under Assumptions 3 and 4.

Lemma 20. *For all* $t\in\mathbb{R}$,

$$\nabla_\theta \widetilde{\Lambda}(t;\breve{\theta}(t))=0. \tag{137}$$

Proof. The proof follows from definition and the assumption that Θ is an open set. □

Lemma 21. *For all* $t\in\mathbb{R}$,

$$\nabla_\theta A(\breve{\theta}(t))=\mathcal{M}(t;\breve{\theta}(t)). \tag{138}$$

Proof. The proof is completed by noting Lemma 20 and Lemma 17. □

Lemma 22 (Derivative of the solution with respect to tilt)**.** *Under Assumption 4, for all* $t\in\mathbb{R}$,

$$\frac{\partial}{\partial t}\breve{\theta}(t)=-\Big(\nabla^2_{\theta\theta^\top}A(\breve{\theta}(t))+t\mathcal{V}(t;\breve{\theta}(t))\Big)^{-1}\mathcal{V}(t;\breve{\theta}(t))\breve{\theta}(t), \tag{139}$$

where

$$\nabla^2_{\theta\theta^\top}A(\breve{\theta}(t))+t\mathcal{V}(t;\breve{\theta}(t))\succ 0 \tag{140}$$

is a symmetric positive definite matrix.

Proof. By noting Lemma 20, and further differentiating with respect to t, we have

$$\begin{aligned}
0&=\frac{\partial}{\partial t}\nabla_\theta\widetilde{\Lambda}(t;\breve{\theta}(t)) && (141)\\
&=\frac{\partial}{\partial \tau}\nabla_\theta\widetilde{\Lambda}(\tau;\breve{\theta}(t))\Big|_{\tau=t}+\nabla^2_{\theta\theta^\top}\widetilde{\Lambda}(t;\breve{\theta}(t))\left(\frac{\partial}{\partial t}\breve{\theta}(t)\right) && (142)\\
&=t\mathcal{V}(t;\breve{\theta}(t))\breve{\theta}(t)+\big(t\nabla^2_{\theta\theta^\top}A(\theta)+t^2\mathcal{V}(t;\theta)\big)\left(\frac{\partial}{\partial t}\breve{\theta}(t)\right), && (143)
\end{aligned}$$

where (142) follows from the chain rule, (143) follows from Lemmas 19 and 21 and 18. The proof is completed by noting that $\nabla^2_{\theta\theta^\top}\widetilde{\Lambda}(t;\breve{\theta}(t))$ is symmetric positive definite for all $t\in\mathbb{R}$ under Assumption 4. □

Finally, we state an auxiliary lemma that will be used in the proof of the main theorem.

Lemma 23. *For all* $t,\tau\in\mathbb{R}$ *and all* $\theta\in\Theta$,

$$\mathcal{M}(\tau;\theta)-\mathcal{M}(t;\theta)=-\left(\int_t^\tau\mathcal{V}(\nu;\theta)d\nu\right)\theta. \tag{144}$$

Proof. The proof is completed by noting that

$$\mathcal{M}(\tau;\theta)-\mathcal{M}(t;\theta)=\int_t^\tau\frac{\partial}{\partial\nu}\mathcal{M}(\nu;\theta)d\nu=-\left(\int_t^\tau\mathcal{V}(\nu;\theta)d\nu\right)\theta. \tag{145}$$

□

Proof of Theorem 2. Notice that for all θ, and all $\epsilon\in\mathbb{R}^{>0}$,

$$\widetilde{R}(t+\epsilon;\theta)\geqslant\widetilde{R}(t;\theta) \tag{146}$$

$$\geqslant\widetilde{R}(t;\breve{\theta}(t)), \tag{147}$$

where (146) follows from Theorem 1 and (147) follows from the definition of $\breve{\theta}(t)$. Hence,

$$\widetilde{R}(t+\epsilon;\breve{\theta}(t+\epsilon))=\min_{\theta\in B(\breve{\theta}(t),r)}\widetilde{R}(t+\epsilon;\theta)\geqslant\widetilde{R}(t;\breve{\theta}(t)), \tag{148}$$

which completes the proof. □

Proof of Theorem 3. Recall that $f(x_i;\theta)=A(\theta)-\theta^\top T(x_i)$. Thus,

$$\widehat{E}_t(\mathbf{f}(\theta))=\sum_{i\in[N]}w_i(t;\breve{\theta}(t))f(x_i;\theta)=A(\theta)-\theta^\top\sum_{i\in[N]}w_i(t;\breve{\theta}(t))T(x_i)=A(\theta)-\theta^\top\mathcal{M}_t, \tag{149}$$

where $\mathcal{M}_t$ is defined in (115). Consequently,

$$\widehat{\mathrm{var}}_t(\mathbf{f}(\theta))=\widehat{E}_t\Big(f(x_i;\theta)-\widehat{E}_t(\mathbf{f}(\theta)\Big)^2 \tag{150}$$

$$=\widehat{E}_t\big(\theta^\top T(x_i)-\theta^\top\mathcal{M}_t\big)^2 \tag{151}$$

$$=\theta^\top\widehat{E}_t\big((T(x_i)-\mathcal{M}_t)(T(x_i)-\mathcal{M}_t)^\top\big)\theta \tag{152}$$

$$=\theta^\top\mathcal{V}_t\theta, \tag{153}$$

where $\mathcal{V}_t$ is defined in (116). Hence,

$$\frac{\partial}{\partial\tau}\Big\{\widehat{\mathrm{var}}_t(\mathbf{f}(\breve{\theta}(\tau)))\Big\}=\left(\frac{\partial}{\partial\tau}\breve{\theta}(\tau)\right)^\top\nabla_\theta\Big\{\widehat{\mathrm{var}}_t(\mathbf{f}(\breve{\theta}(\tau)))\Big\} \tag{154}$$

$$=2\left(\frac{\partial}{\partial\tau}\breve{\theta}(\tau)\right)^\top\mathcal{V}_t\breve{\theta}(\tau) \tag{155}$$

$$=-2\breve{\theta}^\top(\tau)\mathcal{V}(\tau;\breve{\theta}(\tau))\Big(\nabla^2_{\theta\theta}A(\breve{\theta}(\tau))+\tau\mathcal{V}(\tau;\breve{\theta}(\tau))\Big)^{-1}\mathcal{V}_t\breve{\theta}(\tau), \tag{156}$$

and in turn

$$\frac{\partial}{\partial\tau}\Big\{\widehat{\mathrm{var}}_t(\mathbf{f}(\breve{\theta}(\tau)))\Big\}\Big|_{\tau=t}\leqslant 0, \tag{157}$$

where we have used the fact that $\mathcal{V}_\tau\Big(\nabla^2_{\theta\theta}A(\breve{\theta}(\tau))+\tau\mathcal{V}_\tau\Big)^{-1}\mathcal{V}_\tau$ is a symmetric positive semidefinite matrix (due to Lemma 12), hence completing the proof. □

Proof of Theorem 4. Notice that

$$H(\mathbf{w}(t;\theta))=-\sum_{i\in[N]}w_i(t;\theta)\log w_i(t;\theta) \tag{158}$$

$$=-\frac{1}{N}\sum_{i\in[N]}(tf(x_i;\theta)-\widetilde{\Lambda}(t;\theta))e^{tf(x_i;\theta)-\widetilde{\Lambda}(t;\theta)} \tag{159}$$

$$=\widetilde{\Lambda}(t;\theta)-t\frac{1}{N}\sum_{i\in[N]}f(x_i;\theta)e^{tf(x_i;\theta)-\widetilde{\Lambda}(t;\theta)} \tag{160}$$

$$=\widetilde{\Lambda}(t;\theta)-tA(\theta)+t\theta^\top\mathcal{M}(t;\theta). \tag{161}$$

Thus,

$$\nabla_\theta H(\mathbf{w}(t;\theta))=\nabla_\theta\Big(\widetilde{\Lambda}(t;\theta)-tA(\theta)+t\theta^\top\mathcal{M}(t;\theta)\Big) \tag{162}$$

$$=t\nabla_\theta A(\theta)-t\mathcal{M}(t;\theta)-t\nabla_\theta A(\theta)+t\mathcal{M}(t;\theta)-t^2\mathcal{V}(t;\theta)\theta \tag{163}$$

$$=-t^2\mathcal{V}(t;\theta)\theta. \tag{164}$$

Hence,

$$\frac{\partial}{\partial\tau}H\Big(\mathbf{w}(t;\breve{\theta}(\tau))\Big)=\left(\frac{\partial}{\partial\tau}\breve{\theta}(\tau)\right)^\top\nabla_\theta H\Big(\mathbf{w}(t;\breve{\theta}(\tau))\Big) \tag{165}$$

$$=\left(\frac{\partial}{\partial\tau}\breve{\theta}(\tau)\right)^\top\nabla_\theta\Big(\widetilde{\Lambda}(t;\theta)-tA(\theta)+t\theta^\top\mathcal{M}(t;\theta)\Big) \tag{166}$$

$$=t^2\breve{\theta}^\top(\tau)\mathcal{V}(\tau;\breve{\theta}(\tau))\Big(\nabla^2_{\theta\theta}A(\breve{\theta}(\tau))+\tau\mathcal{V}(\tau;\breve{\theta}(\tau))\Big)^{-1}\mathcal{V}(t;\breve{\theta}(\tau))\breve{\theta}(\tau) \tag{167}$$

and

$$\frac{\partial}{\partial\tau}H\Big(\mathbf{w}(t;\breve{\theta}(\tau))\Big)\Big|_{t=\tau}\geqslant 0, \tag{168}$$

completing the proof. □

There are different ways to define performance uniformity. In Theorem 15, we further prove that the tilted cosine similarity between the scaled loss vector and the all-ones vector increases as t decreases by a small amount, which shows that larger t promotes a more *uniform* performance across all losses and can have implications for fairness defined as representation disparity (Hashimoto et al., 2018) (Section 7.2).

Definition 7 (t-tilted cosine similarity)**.** *For $\mathbf{u},\mathbf{v}\in\mathbb{R}^N$, let cosine similarity be defined as*

$$s(\mathbf{u},\mathbf{v}):=\frac{\mathbf{u}^\top\mathbf{v}}{\|\mathbf{u}\|_2\|\mathbf{v}\|_2}. \tag{169}$$

For a weight vector $\mathbf{w}$, let the weighted cosine similarity be defined as

$$s_{\mathbf{w}}(\mathbf{u},\mathbf{v}):=s\Big(\sqrt{\mathbf{W}}\mathbf{u},\sqrt{\mathbf{W}}\mathbf{v}\Big), \tag{170}$$

where $\mathbf{W}:=diag(\mathbf{w})$. In particular, we call $s_{\mathbf{w}(t;\breve{\theta}(t))}(\cdot,\cdot)$ the t-tilted cosine similarity.

Theorem 15 (t-tilted cosine similarity of the scaled loss vector and the all-ones vector increases with t)**.** *Let*

$$\mathbf{f}^+(\theta):=\Big\{f(x_i;\theta)-\widetilde{F}(-\infty)\Big\}_{i\in[N]}, \tag{171}$$

where $\widetilde{F}(-\infty)$ is defined in Eq. (10), and let $\mathbf{1}_N$ denote the all-one N-vector. Then, under Assumption 3 and Assumption 4, for any $t\in\mathbb{R}$,

$$\frac{\partial}{\partial t}\Big\{s_{\mathbf{w}(\tau,\breve{\theta}(\tau))}\Big(\mathbf{f}^+(\breve{\theta}(t)),\mathbf{1}_N\Big)\Big\}\Big|_{\tau=t}>0, \tag{172}$$

where $\mathbf{w}(t;\breve{\theta}(t))$ is the tilted weight vector defined in Eq. (26).

Proof. Notice that

$$s_{\mathbf{w}(t;\breve{\theta}(t))}(\mathbf{f}^+(\theta),\mathbf{1}_N)=\frac{\widehat{E}_t f(x_i;\theta)-\widetilde{F}(-\infty)}{\sqrt{\widehat{E}_t(f(x_i;\theta)-\widetilde{F}(-\infty))^2}}. \tag{173}$$

Hence,

$$\widehat{E}_t f(x_i;\theta)-\widetilde{F}(-\infty)=A(\theta)-\theta^\top\mathcal{M}_t-\widetilde{F}(-\infty), \tag{174}$$

$$\widehat{E}_t(f(x_i;\theta)-\widetilde{F}(-\infty))^2=(A(\theta)-\theta^\top\mathcal{M}_t-\widetilde{F}(-\infty))^2+\theta^\top\mathcal{V}_t\theta, \tag{175}$$

where $\mathcal{M}_t$ and $\mathcal{V}_t$ are defined in (115) and (116), respectively. Notice that

$$\nabla_\theta\Big\{s^2_{\mathbf{w}(t;\breve{\theta}(t))}(\mathbf{f}^+(\theta),\mathbf{1}_N)\Big\} \tag{176}$$

$$=\nabla_\theta\left\{\frac{\Big(\widehat{E}_t f(x_i;\theta)-\widetilde{F}(-\infty)\Big)^2}{\widehat{E}_t(f(x_i;\theta)-\widetilde{F}(-\infty))^2}\right\} \tag{177}$$

$$=\nabla_\theta\left\{\frac{(A(\theta)-\theta^\top\mathcal{M}_t-\widetilde{F}(-\infty))^2}{(A(\theta)-\theta^\top\mathcal{M}_t-\widetilde{F}(-\infty))^2+\theta^\top\mathcal{V}_t\theta}\right\} \tag{178}$$

$$=\frac{2(A(\theta)-\theta^\top\mathcal{M}_t-\widetilde{F}(-\infty))(\nabla_\theta A(\theta)-\mathcal{M}_t)\theta^\top\mathcal{V}_t\theta-2(A(\theta)-\theta^\top\mathcal{M}_t-\widetilde{F}(-\infty))^2\mathcal{V}_t\theta}{\Big((A(\theta)-\theta^\top\mathcal{M}_t-\widetilde{F}(-\infty))^2+\theta^\top\mathcal{V}_t\theta\Big)^2} \tag{179}$$

$$=\frac{2(A(\theta)-\theta^\top\mathcal{M}_t-\widetilde{F}(-\infty))\Big(\theta^\top(\nabla_\theta A(\theta)-\mathcal{M}_t)-A(\theta)+\theta^\top\mathcal{M}_t+\widetilde{F}(-\infty)\Big)\mathcal{V}_t\theta}{\Big((A(\theta)-\theta^\top\mathcal{M}_t-\widetilde{F}(-\infty))^2+\theta^\top\mathcal{V}_t\theta\Big)^2} \tag{180}$$

$$=\frac{2(A(\theta)-\theta^\top\mathcal{M}_t-\widetilde{F}(-\infty))\Big(\theta^\top\nabla_\theta A(\theta)-A(\theta)+\widetilde{F}(-\infty)\Big)\mathcal{V}_t\theta}{\Big((A(\theta)-\theta^\top\mathcal{M}_t-\widetilde{F}(-\infty))^2+\theta^\top\mathcal{V}_t\theta\Big)^2}. \tag{181}$$

Hence,

$$\frac{\partial}{\partial\tau}\Big\{s^2_{\mathbf{w}(t;\breve{\theta}(t))}(\mathbf{f}^+(\breve{\theta}(\tau)),\mathbf{1}_N)\Big\} \tag{182}$$

$$=\left(\frac{\partial}{\partial\tau}\breve{\theta}(\tau)\right)^\top\nabla_\theta\Big\{s^2_{\mathbf{w}(t;\breve{\theta}(t))}(\mathbf{f}^+(\breve{\theta}(\tau)),\mathbf{1}_N)\Big\} \tag{183}$$

$$\begin{aligned}&=-\breve{\theta}^\top(\tau)\mathcal{V}(\tau;\breve{\theta}(\tau))\Big(\nabla^2_{\theta\theta}A(\breve{\theta}(\tau))+\tau\mathcal{V}(\tau;\breve{\theta}(\tau))\Big)^{-1}\\&\quad\times-\frac{2(A(\breve{\theta}(\tau))-\breve{\theta}(\tau)^\top\mathcal{M}_t-\widetilde{F}(-\infty))(A(\breve{\theta}(\tau))-\breve{\theta}(\tau)^\top\mathcal{M}(\tau;\breve{\theta}(\tau))-\widetilde{F}(-\infty))}{\Big((A(\breve{\theta}(\tau))-\breve{\theta}(\tau)^\top\mathcal{M}_t-\widetilde{F}(-\infty))^2+\breve{\theta}(\tau)^\top\mathcal{V}_t\theta\Big)^2}\mathcal{V}_t\breve{\theta}(\tau).\end{aligned} \tag{184}$$

Note that $2(A(\breve{\theta}(\tau))-\breve{\theta}(\tau)^\top\mathcal{M}_t-\widetilde{F}(-\infty))(A(\breve{\theta}(\tau))-\breve{\theta}(\tau)^\top\mathcal{M}(\tau;\breve{\theta}(\tau))-\widetilde{F}(-\infty))>0$ by definition, and $\mathcal{V}_\tau\Big(\nabla^2_{\theta\theta}A(\breve{\theta}(\tau))+\tau\mathcal{V}_\tau\Big)^{-1}\mathcal{V}_\tau$ is a symmetric positive semi-definite matrix. Therefore, The proof is completed following that the quantity in Eq. (184) is non-negative for $t=\tau$. □

Appendix B. Connections to Other Objectives (Proofs and Additional Results)

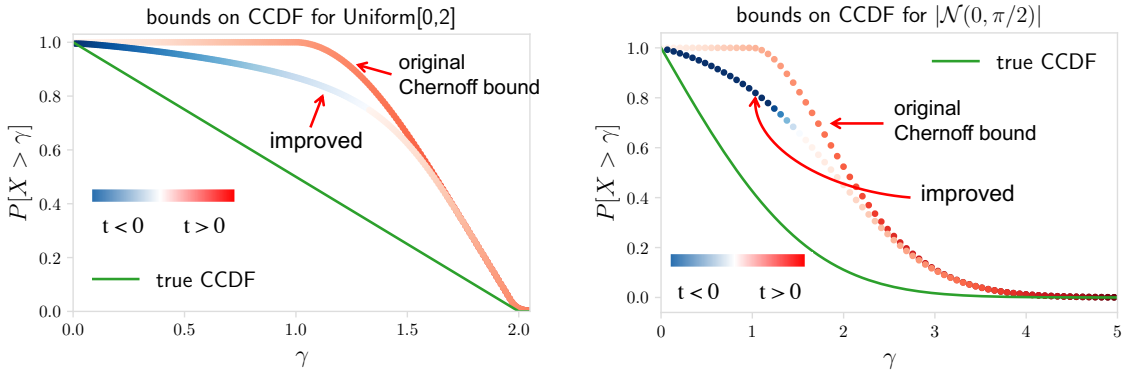


Figure 15: Comparing the new Chernoff bound on complementary CDF (CCDF) (i.e., $P[X \geqslant \gamma]$) proposed in Theorem 5 (denoted as 'improved') with the original Chernoff bound in two cases: $X \sim$Uniform[0,2] and $X \sim |\mathcal{N}(0,\pi/2)|$. We see that by sweeping t from all real numbers, our bound is significantly tighter than the generic Chernoff bound which optimizes over $t \in \mathbb{R}^+$, especially in the small deviations regime.

Lemma 24. *If $a < \widetilde{F}(-\infty)$ then $\widetilde{Q}^0(\gamma)=1$. Further, if $\gamma > \widetilde{F}(+\infty)$ then $\widetilde{Q}^0(\gamma)=0$, where $\widetilde{F}(\cdot)$ is defined in Definition 11, and is reproduced here:*

$$\widetilde{F}(-\infty) = \lim_{t \to -\infty} \widetilde{R}(t; \breve{\theta}(t)) = \min_{\theta} \min_{i \in [N]} f(x_i; \theta), \tag{185}$$

$$\widetilde{F}(+\infty) = \lim_{t \to +\infty} \widetilde{R}(t; \breve{\theta}(t)) = \min_{\theta} \max_{i \in [N]} f(x_i; \theta). \tag{186}$$

Next, we present our main result on the connection between tail distribution of losses and TERM, using Theorem 5.

Theorem 16. *For all $t \in \mathbb{R}$, and all θ, and all $\gamma \in (\widetilde{F}(-\infty), \widetilde{F}(+\infty))$,*[7]

$$\widetilde{Q}(\gamma; \theta) \leqslant \overline{Q}(\gamma; t, \theta) := \frac{e^{\widetilde{R}(t;\theta)t} - e^{\widetilde{F}(-\infty)t}}{e^{\gamma t} - e^{\widetilde{F}(-\infty)t}}. \tag{187}$$

Proof. The proof is a direct application of Theorem 5 to the non-negative random variable $(f(X;\theta) - \widetilde{F}(-\infty))$, where X is distributed according to the empirical distribution. □

Recall that optimizing $\widetilde{\text{VaR}}$ is equivalent to optimizing $\widetilde{Q}$. Next we show how TERM is related to optimizing $\widetilde{Q}$. Recall that $\widetilde{Q}^0(\gamma)$ denotes the optimal value of $\widetilde{Q}(\gamma;\theta)$ optimized over θ. Let

$$\widetilde{Q}^1(\gamma) := \inf_{t \in \mathbb{R}} \left\{ \widetilde{Q}(\gamma; \breve{\theta}(t)) \right\}, \tag{188}$$

which denotes the value at risk optimized over the t-tilted solutions.

7. We define the RHS at $t=0$ via continuous extension.

Theorem 17. *For all* $\gamma\in(\widetilde{F}(-\infty),\widetilde{F}(+\infty))$*, we have*

$$\widetilde{Q}^0(\gamma)\leqslant\widetilde{Q}^1(\gamma)\leqslant\widetilde{Q}^2(\gamma)\leqslant\widetilde{Q}^3(\gamma)=\inf_{t\in\mathbb{R}}\big\{\overline{Q}(\gamma,t)\big\}, \tag{189}$$

where

$$\overline{Q}(\gamma,t):=\frac{e^{\widetilde{F}(t)t}-e^{\widetilde{F}(-\infty)t}}{e^{\gamma t}-e^{\widetilde{F}(-\infty)t}}, \tag{190}$$

$$\widetilde{t}^3(\gamma):=\operatorname*{arginf}_{t\in\mathbb{R}}\big\{\overline{Q}(\gamma,t)\big\}, \tag{191}$$

$$\widetilde{Q}^2(\gamma):=\widetilde{Q}(\gamma;\breve{\theta}(\widetilde{t}^3(\gamma))), \tag{192}$$

$$\widetilde{Q}^3(\gamma):=\overline{Q}(\gamma,\widetilde{t}^3(\gamma)). \tag{193}$$

Proof. The only non-trivial step is to show that $Q^2(\gamma)\leqslant Q^3(\gamma)$. Following Theorem 16,

$$Q^2(\gamma)=\widetilde{Q}(\gamma;\breve{\theta}(\widetilde{t}(\gamma))\leqslant\inf_{t\in\mathbb{R}}\overline{Q}(\gamma;t,\breve{\theta}(t))=Q^3(\gamma), \tag{194}$$

which completes the proof. □

Theorem 17 motivates us with the following approximation on the solutions of the minimizing the tail distribution of losses (Definition 2).

Approximation 1. *For all* $\gamma\in(\widetilde{F}(-\infty),\widetilde{F}(+\infty))$*,*

$$\widetilde{Q}(\gamma;\theta^0(\gamma))=\widetilde{Q}^0(\gamma)\approx\widetilde{Q}^2(\gamma)=\widetilde{Q}(\gamma;\breve{\theta}(\widetilde{t}(\gamma))), \tag{195}$$

and hence, $\breve{\theta}(\widetilde{t}(\gamma))$ *is an approximate solution to the tail probability optimization problem.*

While we have not characterized how tight this approximation is for $\gamma\in(\widetilde{F}(-\infty),\widetilde{F}(+\infty))$, we believe that Approximation 1 provides a reasonable solution to the tail distribution optimization problem in general. This is evidenced empirically when the approximation is evaluated on the toy examples of Figure 1, and compared with the global solutions of the tail distribution optimization method, as shown in Figure 16. As can be seen, $\widetilde{Q}^0(\gamma)\approx\widetilde{Q}^2(\gamma)$ as suggested by Approximation 1. Also, we can see that while the bound in Theorem 17 ($\widetilde{Q}^3(\gamma)$) is not tight, the solution that is obtained from solving it ($\widetilde{Q}^2(\gamma)$) results in a good approximation to the tail distribution minimization ($\widetilde{Q}^0(\gamma)$).

Inverse CVaR. We note that while the most popular form of CVaR focuses on upper quantiles (as discussed in the main text), one may explore 'inverse' CVaR that can focus on lower quantiles, with its empirical form $\widetilde{\mathrm{CVaR}}_{\mathrm{inv}}(1-\alpha;\theta)$ for $\alpha\in[0,1)$ defined as

$$\widetilde{\mathrm{CVaR}}_{\mathrm{inv}}(1-\alpha;\theta):=-\min_{\gamma}\left\{\gamma+\frac{1}{1-\alpha}\frac{1}{N}\sum_{i\in[N]}[-f(x_i;\theta)-\gamma]_+\right\}. \tag{196}$$

As α ranges from 0 to 1, optimizing $\widetilde{\mathrm{CVaR}}_{\mathrm{inv}}(1-\alpha;\theta)$ transitions from solving avg-loss to min-loss. However, different from TiVaR or CVaR, $\mathrm{CVaR}_{\mathrm{inv}}$ is not a valid upper bound of

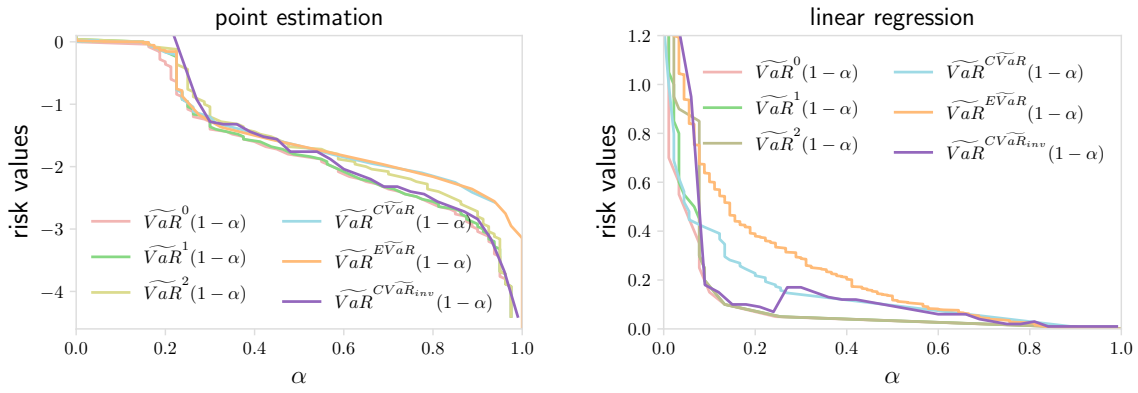


Figure 16: Comparing the solutions of different risks in terms of how well they solve VaR. For $i\in\{0,1,2\}$, $\widetilde{\mathrm{VaR}}^i(1-\alpha):=\min_\gamma\{\gamma|\widetilde{Q}^i(\gamma)\leqslant\alpha\}$. $\widetilde{\mathrm{VaR}}^0(1-\alpha):=\min_\gamma\{\gamma|\widetilde{Q}^0(\gamma)\leqslant\alpha\}$ is the optimal $\widetilde{\mathrm{VaR}}(1-\alpha;\theta)$. By definition, $\widetilde{\mathrm{VaR}}^2(1-\alpha)$ is the risk value of $\widetilde{\mathrm{VaR}}(1-\alpha;\theta)$ with θ being the solutions of $\widetilde{\mathrm{TiVaR}}(1-\alpha;\theta)$. $\widetilde{\mathrm{VaR}}^{\widetilde{\mathrm{CVaR}}}(1-\alpha)$ denotes the value of $\widetilde{\mathrm{VaR}}(1-\alpha;\theta)$ evaluated at $\mathrm{argmin}_\theta\widetilde{\mathrm{CVaR}}(1-\alpha;\theta)$, and $\widetilde{\mathrm{VaR}}^{\widetilde{\mathrm{CVaR}}_{\mathrm{inv}}}(1-\alpha)$ and $\widetilde{\mathrm{VaR}}^{\widetilde{\mathrm{EVaR}}}(1-\alpha)$ are defined in the similar way. We see that $\widetilde{\mathrm{VaR}}^1(1-\alpha)$ and $\widetilde{\mathrm{VaR}}^2(1-\alpha)$ are close to $\widetilde{\mathrm{VaR}}^0(1-\alpha)$, which indicates VaR with the solutions obtained from solving $\widetilde{\mathrm{TiVaR}}(1-\alpha;\theta)$ (which is $\widetilde{\mathrm{VaR}}^2(1-\alpha)$) is a tight upper bound of the globally optimal $\widetilde{\mathrm{VaR}}(1-\alpha;\theta)$. $\widetilde{\mathrm{VaR}}^2(1-\alpha)$ is also tighter than VaR under EVaR solutions when α is not small.

VaR. Despite this, we optimize $\min_\theta\widetilde{\mathrm{CVaR}}_{\mathrm{inv}}(1-\alpha;\theta)$, plug in the optimal model parameters to evaluate VaR values, and compare with the approximate VaR values under the solutions of other risks including TiVaR. From Figure 16, we see that VaR values under TiVaR solutions can be smaller than those under $\mathrm{CVaR}_{\mathrm{inv}}$ solutions on linear regression. Given any α, our proposed TiVaR objective approximates VaR, ranging from min-loss to max-loss smoothly *in a single formulation*, which can be more desirable than optimizing two objectives.

Proof of Theorem 6. We first prove $\widetilde{\mathrm{TiVaR}}(1-\alpha;\theta)\leqslant\widetilde{\mathrm{EVaR}}(1-\alpha;\theta)$.

$$\widetilde{\mathrm{EVaR}}(1-\alpha;\theta)-\widetilde{F}(-\infty)=\min_{t\in\mathbb{R}^{>0}}\frac{1}{t}\log\left(\frac{\frac{1}{N}\sum_{i\in[N]}e^{tf(x_i;\theta)}}{\alpha}\right)-\widetilde{F}(-\infty) \tag{197}$$

$$=\min_{t\in\mathbb{R}^{>0}}\frac{1}{t}\log\left(\frac{e^{(\widetilde{R}(t;\theta)-\widetilde{F}(-\infty))t}}{\alpha}\right) \tag{198}$$

$$\geqslant\min_{t\in\mathbb{R}^{>0}}\frac{1}{t}\log\left[\frac{e^{(\widetilde{R}(t;\theta)-\widetilde{F}(-\infty))t}-(1-\alpha)}{\alpha}\right]_+ \tag{199}$$

$$\geqslant\min_{t\in\mathbb{R}}\frac{1}{t}\log\left[\frac{e^{(\widetilde{R}(t;\theta)-\widetilde{F}(-\infty))t}-(1-\alpha)}{\alpha}\right]_+. \tag{200}$$

We next prove $\widetilde{\mathrm{VaR}}(1-\alpha;\theta)\leqslant\widetilde{\mathrm{TiVaR}}(1-\alpha;\theta)$. From Theorem 16, we know that for any t,θ,

$$\widetilde{Q}(\gamma;\theta)\leqslant\min_{t\in\mathbb{R}}\left\{\frac{e^{\widetilde{R}(t;\theta)t}-e^{-\widetilde{F}(-\infty)t}}{e^{\gamma t}-e^{\widetilde{F}(-\infty)t}}\right\} \tag{201}$$

Let $\widetilde{Q}(\gamma;\theta)=\alpha$, and $\gamma^*=\widetilde{\mathrm{VaR}}(1-\alpha;\theta)$. We have $\min_{t\in\mathbb{R}}\left\{\frac{e^{\widetilde{R}(t;\theta)t}-e^{-\widetilde{F}(-\infty)t}}{e^{\gamma^* t}-e^{\widetilde{F}(-\infty)t}}\right\}\geqslant\alpha$. We also note

$$\min_{t\in\mathbb{R}}\left\{\frac{e^{\widetilde{R}(t;\theta)t}-e^{-\widetilde{F}(-\infty)t}}{e^{\widetilde{\mathrm{TiVaR}}(1-\alpha;\theta)t}-e^{\widetilde{F}(-\infty)t}}\right\}=\alpha. \tag{202}$$

Hence,

$$\widetilde{\mathrm{TiVaR}}(1-\alpha;\theta)\geqslant\gamma^*=\widetilde{\mathrm{VaR}}(1-\alpha;\theta). \tag{203}$$

TERM and Entropic Value-at-Risk. Let $\breve{\theta}_X(t)$ be the minimizer of entropic risk $R_X(t;\theta)$:

$$\breve{\theta}_X(t):=\operatorname*{argmin}_{\theta\in\Theta}R_X(t;\theta). \tag{204}$$

Further, let $F_X(t)$ be the optimum value of entropic risk, i.e.,

$$F_X(t):=R_X(t;\breve{\theta}_X(t)). \tag{205}$$

Our next result will relate EVaR to entropic risk.

Lemma 25 (Relations between entropic risk and EVaR)**.** *Assume that for $t\in\mathbb{R}^{>0}$, $F_X(t)$ is a strongly convex function of $\frac{1}{t}$. Further, let*

$$\breve{t}_X(\alpha)\in\arg\min_{t\in\mathbb{R}^{>0}}\left\{F_X(t)-\frac{1}{t}\log\alpha\right\}, \tag{206}$$

then

$$\operatorname*{argmin}_{\theta}\{EVaR_X(1-\alpha;\theta)\}=\operatorname*{argmin}_{\theta}\left\{R_X(\breve{t}_X(\alpha);\theta)\right\}:=\breve{\theta}_X(\breve{t}_X(\alpha)), \tag{207}$$

$$R_X\left(\breve{t}_X(\alpha);\breve{\theta}_X(\breve{t}_X(\alpha))\right)=F_X(\breve{t}_X(\alpha))\leqslant EVaR_X\left(1-\alpha;\breve{\theta}_X(\breve{t}_X(\alpha))\right). \tag{208}$$

Proof. Consider the any minimizer of $R_X(\breve{t}_X(\alpha),\theta)$, i.e.,

$$\breve{\theta}_X(\breve{t}_X(\alpha))\in\operatorname*{argmin}_{\theta}R_X(\breve{t}_X(\alpha);\theta), \tag{209}$$

we next prove

$$\breve{\theta}_X(\breve{t}_X(\alpha))\in\operatorname*{argmin}_{\theta}\left(\min_{t>0}\left(\frac{1}{t}\log\mathbb{E}[e^{tf(X;\theta)}]-\frac{1}{t}\log\alpha\right)\right). \tag{210}$$

Denote $\min_{t>0}\left(\frac{1}{t}\log\mathbb{E}[e^{tf(X;\theta)}]-\frac{1}{t}\log\alpha\right)$ as $h(\theta;\alpha)$. Let

$$\theta_v^*\in\operatorname*{argmin}_{\theta}h(\theta;\alpha), \tag{211}$$

$$t_v(\theta_v^*)\in\operatorname*{argmin}_{t>0}\left(\frac{1}{t}\log\mathbb{E}[e^{tf(X;\theta_v^*)}]-\frac{1}{t}\log\alpha\right). \tag{212}$$

By the definition of $\breve{t}_X(\alpha)$ and $\breve{\theta}_X(t)$, we have

$$\frac{1}{\breve{t}_X(\alpha)}\log\mathbb{E}[e^{\breve{t}_X(\alpha)f(X;\breve{\theta}_X(\breve{t}_X(\alpha)))}]-\frac{1}{\breve{t}_X(\alpha)}\log\alpha \tag{213}$$

$$\leqslant\frac{1}{t_v(\theta_v^*)}\log\mathbb{E}[e^{t_v(\theta_v^*)f(X;\breve{\theta}_X(t_v(\theta_v^*)))}]-\frac{1}{t_v(\theta_v^*)}\log\alpha \tag{214}$$

$$\leqslant\frac{1}{t_v(\theta_v^*)}\log\mathbb{E}[e^{t_v(\theta_v^*)f(X;\theta_v^*)}]-\frac{1}{t_v(\theta_v^*)}\log\alpha. \tag{215}$$

By the definition of θ_v^*, $h(\theta_v^*;\alpha)\leqslant h(\breve{\theta}_X(\breve{t}_X(\alpha));\alpha)$, i.e.,

$$\min_{t>0}\left(\frac{1}{t}\log\mathbb{E}[e^{tf(X;\theta_v^*)}]-\frac{1}{t}\log\alpha\right)\leqslant\min_{t>0}\left(\frac{1}{t}\log\mathbb{E}[e^{tf(X;\breve{\theta}_X(\breve{t}_X(\alpha)))}]-\frac{1}{t}\log\alpha\right). \tag{216}$$

We have

$$\frac{1}{t_v(\theta_v^*)}\log\mathbb{E}[e^{t_v(\theta_v^*)f(X;\theta_v^*)}]-\frac{1}{t_v(\theta_v^*)}\log\alpha \tag{217}$$

$$\leqslant\min_{t>0}\left(\frac{1}{t}\log\mathbb{E}[e^{tf(X;\breve{\theta}_X(\breve{t}_X(\alpha)))}]-\frac{1}{t}\log\alpha\right) \tag{218}$$

$$\leqslant\frac{1}{\breve{t}_X(\alpha)}\log\mathbb{E}[e^{\breve{t}_X(\alpha)f(X;\breve{\theta}_X(\breve{t}_X(\alpha)))}]-\frac{1}{\breve{t}_X(\alpha)}\log\alpha, \tag{219}$$

Hence, $\breve{\theta}_X(\breve{t}_X(\alpha))\in\operatorname{argmin}_\theta\big(\min_{t>0}\big(\frac{1}{t}\log\mathbb{E}[e^{tf(X;\theta)}]-\frac{1}{t}\log\alpha\big)\big)$.

For the other direction, consider any minimizer of $\mathrm{EVaR}_X(1-\alpha;\theta)$, i.e.,

$$\theta_v^*\in\operatorname*{argmin}_\theta h(\theta;\alpha) \tag{220}$$

$$t_v(\theta_v^*)\in\operatorname*{argmin}_{t>0}\left(\frac{1}{t}\log\mathbb{E}[e^{f(X;\theta_v^*)}]-\frac{1}{t}\log\alpha\right). \tag{221}$$

We next prove $\theta_v^*\in\operatorname{argmin}_\theta\frac{1}{\breve{t}_X(\alpha)}\log\mathbb{E}[e^{\breve{t}_X(\alpha)f(X;\theta)}]$. By the definition of $\breve{\theta}_X(t)$ and $\breve{t}_X(\alpha)$,

$$\frac{1}{t_v(\theta_v^*)}\log\mathbb{E}[e^{t_v(\theta_v^*)f(X;\theta_v^*)}]-\frac{1}{t_v(\theta_v^*)}\log\alpha \tag{222}$$

$$\geqslant\frac{1}{t_v(\theta_v^*)}\log\mathbb{E}[e^{t_v(\theta_v^*)f(X;\breve{\theta}_X(t_v(\theta_v^*)))}]-\frac{1}{t_v(\theta_v^*)}\log\alpha \tag{223}$$

$$\geqslant\frac{1}{\breve{t}_X(\alpha)}\log\mathbb{E}[e^{\breve{t}_X(\alpha)f(X;\breve{\theta}_X(\breve{t}_X(\alpha)))}]-\frac{1}{\breve{t}_X(\alpha)}\log\alpha. \tag{224}$$

On the other hand, by the definition of θ_v^* and $t_v(\theta_v^*)$,

$$\frac{1}{t_v(\theta_v^*)}\log\mathbb{E}[e^{t_v(\theta_v^*)f(X;\theta_v^*)}]-\frac{1}{t_v(\theta_v^*)}\log\alpha\leqslant\min_{t>0}\left(\frac{1}{t}\log\mathbb{E}[e^{tf(X;\breve{\theta}_X(\breve{t}_X(\alpha)))}]-\frac{1}{t}\log\alpha\right) \tag{225}$$

$$\leqslant\frac{1}{\breve{t}_X(\alpha)}\log\mathbb{E}[e^{\breve{t}_X(\alpha)f(X;\breve{\theta}_X(\breve{t}_X(\alpha)))}]-\frac{1}{\breve{t}_X(\alpha)}\log\alpha. \tag{226}$$

Therefore,

$$\frac{1}{t_v(\theta_v^*)}\log\mathbb{E}[e^{t_v(\theta_v^*)f(X;\theta_v^*)}]-\frac{1}{t_v(\theta_v^*)}\log\alpha \tag{227}$$

$$=\frac{1}{t_v(\theta_v^*)}\log\mathbb{E}[e^{t_v(\theta_v^*)f(X;\breve{\theta}_X(t_v(\theta_v^*)))}]-\frac{1}{t_v(\theta_v^*)}\log\alpha \tag{228}$$

$$=\frac{1}{\breve{t}_X(\alpha)}\log\mathbb{E}[e^{\breve{t}_X(\alpha)f(X;\breve{\theta}_X(\breve{t}_X(\alpha)))}]-\frac{1}{\breve{t}_X(\alpha)}\log\alpha. \tag{229}$$

If

$$s(t):=\frac{1}{t}\log\mathbb{E}[e^{tf(X;\breve{\theta}_X(t))}]-\frac{1}{t}\log\alpha \tag{230}$$

has a unique minimizer,

$$\breve{t}_X(\alpha)=t_v(\theta_v^*), \tag{231}$$

and

$$\theta_v^*\in\operatorname{argmin}_\theta\frac{1}{\breve{t}_X(\alpha)}\log\mathbb{E}[e^{\breve{t}_X(\alpha)f(X;\theta)}]. \tag{232}$$

Hence, we have proved

$$\operatorname{argmin}_\theta\mathrm{EVaR}_X(1-\alpha;\theta)=\operatorname{argmin}_\theta R_X(\breve{t}_X(\alpha);\theta), \tag{233}$$

and

$$R_X(\breve{t}_X(\alpha);\breve{\theta}_X(\breve{t}_X(\alpha)))\leqslant\mathrm{EVaR}_X(1-\alpha;\breve{\theta}_X(\breve{t}_X(\alpha))). \tag{234}$$

□

The lemma relates the solution and the optimal value of EVaR with those of entropic risk. We can extend Lemma 25 to the empirical version below.

Lemma 26 (Relations between empirical entropic risk and empirical EVaR)**.** *Assume that $\widetilde{F}(t)$ is a strongly convex function of $\frac{1}{t}$. For $\alpha\in\{\frac{k}{N}\}_{k\in[N]}$, let*

$$\breve{t}(\alpha)\in\operatorname{argmin}_{t>0}\left\{\widetilde{F}(t)-\frac{1}{t}\log\alpha\right\}, \tag{235}$$

then

$$\operatorname{argmin}_\theta\widetilde{EVaR}(1-\alpha;\theta)=\operatorname{argmin}_\theta\widetilde{R}(\breve{t}(\alpha);\theta), \tag{236}$$

$$\widetilde{F}(\breve{t}(\alpha))\leqslant\widetilde{EVaR}(1-\alpha;\breve{\theta}(\breve{t}(\alpha))). \tag{237}$$

Appendix C. Solving TERM (Proofs and Details)

C.1 Hierarchical Multi-Objective Tilting

We state the hierarchical multi-objective tilting for a hierarchy of depth 3. While we don't directly use this form, it is stated to clarify the experiments in Section 7 where tilting is done at class level and annotator level, and the sample-level tilt value could be understood to be 0.

$$\widetilde{J}(m,t,\tau;\theta):=\frac{1}{m}\log\left(\frac{1}{N}\sum_{G\in[GG]}\left(\sum_{g\in[G]}|g|\right)e^{m\widetilde{J}_G(\tau;\theta)}\right) \tag{238}$$

$$\widetilde{J}_G(t,\tau;\theta):=\frac{1}{t}\log\left(\frac{1}{\sum_{g\in[G]}|g|}\sum_{g\in[G]}|g|e^{t\widetilde{R}_g(\tau;\theta)}\right) \tag{239}$$

$$\widetilde{R}_g(\tau;\theta):=\frac{1}{\tau}\log\left(\frac{1}{|g|}\sum_{x\in g}e^{\tau f(x;\theta)}\right), \tag{240}$$

Proof of Lemma 10. We proceed as follows. First notice that by invoking Lemma 5,

$$\nabla_\theta\widetilde{J}(t,\tau;\theta)=\sum_{g\in[G]}w_g(t,\tau;\theta)\nabla_\theta\widetilde{R}_g(\tau;\theta) \tag{241}$$

where

$$w_g(t,\tau;\theta):=\frac{|g|e^{t\widetilde{R}_g(\tau;\theta)}}{\sum_{g'\in[G]}|g'|e^{t\widetilde{R}_{g'}(\tau;\theta)}}. \tag{242}$$

where $\widetilde{R}_g(\tau;\theta)$ is defined in (83), and is reproduced here:

$$\widetilde{R}_g(\tau;\theta):=\frac{1}{\tau}\log\left(\frac{1}{|g|}\sum_{x\in g}e^{\tau f(x;\theta)}\right). \tag{243}$$

On the other hand, by invoking Lemma 5,

$$\nabla_\theta\widetilde{R}_g(\tau;\theta)=\sum_{x\in g}w_{g,x}(\tau;\theta)\nabla_\theta f(x;\theta) \tag{244}$$

where

$$w_{g,x}(\tau;\theta):=\frac{e^{\tau f(x;\theta)}}{\sum_{y\in g}e^{\tau f(y;\theta)}}. \tag{245}$$

Hence, combining (241) and (244),

$$\nabla_\theta\widetilde{J}(t,\tau;\theta)=\sum_{g\in[G]}\sum_{x\in g}w_g(t,\tau;\theta)w_{g,x}(\tau;\theta)\nabla_\theta f(x;\theta). \tag{246}$$

The proof is completed by algebraic manipulations to show that

$$w_{g,x}(t,\tau;\theta)=w_g(t,\tau;\theta)w_{g,x}(\tau;\theta). \tag{247}$$

□

C.2 Proofs of Convergence for TERM Solvers

Algorithm 5: Stochastic Non-Hierarchical TERM with two mini-batches

Initialize: $\theta, \widetilde{\widetilde{R}}_t = \frac{1}{t}\log\Big(\frac{1}{N}\sum_{i\in[N]} e^{tf(x_i;\theta)}\Big)$
Input: t, α, λ
while *stopping criteria not reached* **do**
  sample two independent minibatches B_1, B_2 uniformly at random from $[N]$
  compute the loss $f(x;\theta)$ and gradient $\nabla_\theta f(x;\theta)$ for all $x \in B_1$
  $\widetilde{R}_{B,t} \leftarrow t$-tilted loss (2) on minibatch B_2
  $\widetilde{\widetilde{R}}_t \leftarrow \frac{1}{t}\log\Big((1-\lambda)e^{t\widetilde{\widetilde{R}}_t} + \lambda e^{t\widetilde{R}_{B,t}}\Big)$
  $w_{t,x} \leftarrow e^{tf(x;\theta) - t\widetilde{\widetilde{R}}_t}$
  $\theta \leftarrow \theta - \frac{\alpha}{|B_1|}\sum_{x\in B_1} w_{t,x}\nabla_\theta f(x;\theta)$
end

To prove our convergence results in Theorem 12, we first prove a lemma below.

Lemma 27. *Denote* $k_t := \operatorname{argmax}_k\Big(k < \frac{2e}{\mu} + \frac{etLBe^{t(\widetilde{F}_{\max}-\widetilde{F}_{\min})}}{\mu k}\Big)$. *Let* $\lambda = 1 - \frac{1}{2e}$, *and*

$$\alpha_k = \begin{cases} \frac{1}{tLBe^{t(\widetilde{F}_{\max}-\widetilde{F}_{\min})}+1}, & \text{if } k \leqslant k_t \\ \frac{2e}{\mu k}, & \text{otherwise,} \end{cases} \tag{248}$$

then for any k,

$$\mathbb{E}[e^{t(\widetilde{\widetilde{R}}_k - \widetilde{R}_k)} | \theta_1, ..., \theta_k] \leqslant 2e, \tag{249}$$

where $\widetilde{R}_k := \widetilde{R}(t;\theta_k) = \frac{1}{t}\log\Big(\frac{1}{N}\sum_{i\in[N]} e^{tf(x_i;\theta_k)}\Big)$.

Proof. We have the updating rule

$$e^{t\widetilde{\widetilde{R}}_{k+1}} = \lambda e^{tf(\xi_k,\theta_k)} + (1-\lambda)e^{t\widetilde{\widetilde{R}}_k}. \tag{250}$$

Taking conditional expectation $\mathbb{E}[\cdot|\theta_1,...,\theta_{k+1}]$ on both sides of (250) gives

$$\mathbb{E}[e^{t(\widetilde{\widetilde{R}}_{k+1} - \widetilde{R}_k)} | \theta_1, ..., \theta_{k+1}] \tag{251}$$
$$= \lambda\mathbb{E}[e^{t(f(\xi_k;\theta_k) - \widetilde{R}_k)} | \theta_1, ..., \theta_{k+1}] + (1-\lambda)\mathbb{E}[e^{t(\widetilde{\widetilde{R}}_k - \widetilde{R}_k)} | \theta_1, ..., \theta_{k+1}] \tag{252}$$
$$= \lambda + (1-\lambda)\mathbb{E}[e^{t(\widetilde{\widetilde{R}}_k - \widetilde{R}_k)} | \theta_1, ..., \theta_k]. \tag{253}$$

For any k, we have

$$\|\theta_{k+1} - \theta_k\| = \alpha_k \left\| \frac{e^{t\widetilde{R}_k}}{e^{t\widetilde{\widetilde{R}}_k}} \nabla \widetilde{R}_k \right\| \leqslant \alpha_k e^{t(\widetilde{F}_{\max} - \widetilde{F}_{\min})} B. \tag{254}$$

Therefore,

$$|f(x_i;\theta_{k-1})-f(x_i;\theta_k)|\leqslant L\|\theta_{k-1}-\theta_k\|\leqslant\alpha_k LBe^{t(\tilde{F}_{\max}-\tilde{F}_{\min})}, \tag{255}$$

and

$$e^{-t\alpha_k LBe^{t(\tilde{F}_{\max}-\tilde{F}_{\min})}}\leqslant e^{t(\widetilde{R}_k-\widetilde{R}_{k+1})}=\frac{\sum_{i\in[N]}e^{tf(x_i;\theta_k)}}{\sum_{i\in[N]}e^{tf(x_i;\theta_{k+1})}}\leqslant e^{t\alpha_k LBe^{t(\tilde{F}_{\max}-\tilde{F}_{\min})}}, \tag{256}$$

$$\begin{aligned}
&e^{-t\alpha_k LBe^{t(\tilde{F}_{\max}-\tilde{F}_{\min})}}\mathbb{E}[e^{t(\tilde{\widetilde{R}}_{k+1}-\widetilde{R}_{k+1})}|\theta_1,...,\theta_{k+1}] && (257)\\
&\leqslant\mathbb{E}[e^{t(\tilde{\widetilde{R}}_{k+1}-\widetilde{R}_k)}|\theta_1,...,\theta_{k+1}]\\
&\leqslant e^{t\alpha_k LBe^{t(\tilde{F}_{\max}-\tilde{F}_{\min})}}\mathbb{E}[e^{t(\tilde{\widetilde{R}}_{k+1}-\widetilde{R}_{k+1})}|\theta_1,...,\theta_{k+1}]. && (258)
\end{aligned}$$

Hence,

$$\begin{aligned}
&e^{-t\alpha_k LBe^{t(\tilde{F}_{\max}-\tilde{F}_{\min})}}\mathbb{E}[e^{t(\tilde{\widetilde{R}}_{k+1}-\widetilde{R}_{k+1})}|\theta_1,...,\theta_{k+1}] && (259)\\
&\leqslant\lambda+(1-\lambda)\mathbb{E}[e^{t(\tilde{\widetilde{R}}_k-\widetilde{R}_k)}|\theta_1,...,\theta_k] && (260)\\
&\leqslant e^{t\alpha_k LBe^{t(\tilde{F}_{\max}-\tilde{F}_{\min})}}\mathbb{E}[e^{t(\tilde{\widetilde{R}}_{k+1}-\widetilde{R}_{k+1})}|\theta_1,...,\theta_{k+1}]. && (261)
\end{aligned}$$

(i) When $k\leqslant k_t$, under the learning rate α_k set as in Eq. (248), we have

$$\alpha_k LBe^{t(\tilde{F}_{\max}-\tilde{F}_{\min})}<1. \tag{262}$$

Hence,

$$\begin{aligned}
\mathbb{E}[e^{t(\tilde{\widetilde{R}}_k-\widetilde{R}_k)}|\theta_1,...,\theta_k]&\leqslant e(\lambda+(1-\lambda)\mathbb{E}[e^{t(\tilde{\widetilde{R}}_{k-1}-\widetilde{R}_{k-1})}|\theta_1,...,\theta_{k-1}]) && (263)\\
&\leqslant e+\frac{1}{2}\mathbb{E}[e^{t(\tilde{\widetilde{R}}_{k-1}-\widetilde{R}_{k-1})}|\theta_1,...,\theta_{k-1}] && (264)\\
&\leqslant\cdots\leqslant e\left(2-\frac{1}{2^{k-2}}\right)+\frac{1}{2^{k-1}}\mathbb{E}[e^{t(\tilde{\widetilde{R}}_1-\widetilde{R}_1)}|\theta_1]\leqslant 2e. && (265)
\end{aligned}$$

(ii) When $k>k_t$,

$$\alpha_k=\frac{2e}{\mu k}<\frac{k}{k+tLBe^{t(\tilde{F}_{\max}-\tilde{F}_{\min})}}. \tag{266}$$

Similarly, we have

$$\begin{aligned}
\mathbb{E}[e^{t(\tilde{\widetilde{R}}_k-\widetilde{R}_k)}|\theta_1,...,\theta_k]&\leqslant e^{t\alpha_k LBe^{t(\tilde{F}_{\max}-\tilde{F}_{\min})}}\left(\lambda+(1-\lambda)\mathbb{E}[e^{t(\tilde{\widetilde{R}}_{k-1}-\widetilde{R}_{k-1})}|\theta_1,...,\theta_{k-1}]\right) && (267)\\
&\leqslant\cdots\leqslant 2e, && (268)
\end{aligned}$$

which completes the proof. □

Proof of Theorem 12 Denote the empirical optimal solution $\breve{\theta}(t)$ as θ^*. Denote the tilted stochastic gradient on data ζ_k as g_k, where

$$g_k=\frac{e^{tf(\zeta_k;\theta_k)}}{e^{t\widetilde{\widetilde{R}}_k}}\nabla f(\zeta_k;\theta_k)=\frac{e^{t\widetilde{R}_k}}{e^{t\widetilde{\widetilde{R}}_k}}\frac{e^{tf(\zeta_k;\theta_k)}}{e^{t\widetilde{R}_k}}\nabla f(\zeta_k;\theta_k)=\frac{e^{t\widetilde{R}_k}}{e^{t\widetilde{\widetilde{R}}_k}}\nabla\widetilde{R}_k(\zeta_k). \tag{269}$$

Therefore, for any $k\geqslant 1$,

$$\begin{aligned}
\mathbb{E}[\langle\theta_k-\theta^*,g_k\rangle]&=\mathbb{E}[\mathbb{E}[\langle\theta_k-\theta^*,g_k\rangle|\theta_1,...,\theta_k]] && (270)\\
&=\mathbb{E}[\langle\theta_k-\theta^*,\mathbb{E}[g_k|\theta_1,...,\theta_k]\rangle] && (271)\\
&=\mathbb{E}[\langle\theta_k-\theta^*,\mathbb{E}[e^{t(\widetilde{R}_k-\widetilde{\widetilde{R}}_k)}|\theta_1,...,\theta_k]\mathbb{E}[\nabla\widetilde{R}_k(\zeta_k)|\theta_1,...,\theta_k]\rangle] && (272)\\
&\geqslant\frac{1}{2e}\mathbb{E}[\langle\theta_k-\theta^*,\nabla\widetilde{R}(\theta_k)\rangle]\quad(\mathbb{E}[e^{t(\widetilde{R}_k-\widetilde{\widetilde{R}}_k)}|\theta_1,...,\theta_k]\geqslant 1/\mathbb{E}[e^{t(\widetilde{\widetilde{R}}_k-\widetilde{R}_k)}|\theta_1,...,\theta_k]) && (273)\\
&\geqslant\frac{\mu}{2e}\mathbb{E}[\|\theta_k-\theta^*\|^2]\quad(\mu\text{-strong convexity of }\widetilde{R}), && (274)
\end{aligned}$$

where (272) follows from the fact that $e^{t(\widetilde{R}_k-\widetilde{\widetilde{R}}_k)}$ and $\nabla\widetilde{R}_k(\zeta_k)$ are independent given $\{\theta_1,...,\theta_k\}$. For $k\geqslant k_t$ with $\alpha_k=\frac{2e}{\mu k}$,

$$\begin{aligned}
\mathbb{E}[\|\theta_{k+1}-\theta^*\|^2]&=\mathbb{E}[\|\theta_k-\alpha_k g_k-\theta^*\|^2] && (275)\\
&=\mathbb{E}[\|\theta_k-\theta^*\|^2]-2\alpha_k\mathbb{E}[\langle\theta_k-\theta^*,g_k\rangle]+\alpha_k^2\mathbb{E}[\|g_k\|^2] && (276)\\
&\leqslant\Big(1-\frac{\alpha_k\mu}{e}\Big)\mathbb{E}[\|\theta_k-\theta^*\|^2]+\alpha_k^2\mathbb{E}[\|e^{t(\widetilde{R}_k-\widetilde{\widetilde{R}}_k)}\nabla\widetilde{R}_k(\zeta_k)\|^2] && (277)\\
&\leqslant\left(1-\frac{2}{k}\right)\mathbb{E}[\|\theta_k-\theta^*\|^2]+\frac{4e^2B^2e^{2t(\widetilde{F}_{\max}-\widetilde{F}_{\min})}}{\mu^2k^2}. && (278)
\end{aligned}$$

When $k\leqslant k_t$ with $\alpha_k=\frac{1}{1+tLBe^{t(\widetilde{F}_{\max}-\widetilde{F}_{\min})}}$,

$$\mathbb{E}[\|\theta_k-\theta^*\|^2]\leqslant\left(1-\frac{\mu}{e(tLBe^{t(\widetilde{F}_{\max}-\widetilde{F}_{\min})}+1)}\right)\mathbb{E}[\|\theta_{k-1}-\theta^*\|^2]+\frac{B^2e^{2t(\widetilde{F}_{\max}-\widetilde{F}_{\min})}}{(1+tLBe^{t(\widetilde{F}_{\max}-\widetilde{F}_{\min})})^2}. \tag{279}$$

$$\tag{280}$$

We can thus prove

$$\mathbb{E}[\|\theta_{k_t}-\theta^*\|^2]\leqslant\max\left\{\mathbb{E}[\|\theta_1-\theta^*\|^2],\frac{B^2e^{2t(\widetilde{F}_{\max}-\widetilde{F}_{\min})+1}}{\mu(1+tLBe^{t(\widetilde{F}_{\max}-\widetilde{F}_{\min})})}\right\} \tag{281}$$

Let

$$V_t=\max\left\{k_t\mathbb{E}[\|\theta_{k_t}-\theta^*\|^2],\frac{4B^2e^{2+2t(\widetilde{F}_{\max}-\widetilde{F}_{\min})}}{\mu^2}\right\}. \tag{282}$$

We next prove for $k \geqslant k_t$,

$$\mathbb{E}[\|\theta_k - \theta^*\|^2] \leqslant \frac{V_t}{k}. \tag{283}$$

Suppose $\mathbb{E}[\|\theta_k - \theta^*\|^2] \leqslant \frac{V_t}{k}$. From (278), we have

$$\mathbb{E}[\|\theta_{k+1} - \theta^*\|^2] \leqslant \left(1 - \frac{2}{k}\right)\mathbb{E}[\|\theta_k - \theta^*\|^2] + \frac{4e^2 B^2 e^{2t(\widetilde{F}_{\max} - \widetilde{F}_{\min})}}{k^2 \mu^2} \tag{284}$$

$$\leqslant \left(1 - \frac{2}{k}\right)\frac{V_t}{k} + \frac{V_t^2}{k^2} \tag{285}$$

$$\leqslant \frac{V_t}{k+1}, \tag{286}$$

where $k \geqslant k_t = \left\lceil \frac{e + \sqrt{e^2 + \mu t L B e^{t(\widetilde{F}_{\max} - \widetilde{F}_{\min})} + 1}}{\mu} \right\rceil$. This completes the proof. □

Proof of Theorem 13. Assume $\widetilde{R}(t;\theta)$ is non-convex and β-smooth, we have

$$\widetilde{R}_{k+1} - \widetilde{R}_k - \langle \nabla \widetilde{R}_k, \theta_{k+1} - \theta_k \rangle \leqslant \frac{\beta}{2}\|\theta_{k+1} - \theta_k\|^2, \tag{287}$$

where $\widetilde{R}_k := \widetilde{R}(t;\theta_k) = \frac{1}{t}\log\left(\frac{1}{N}\sum_{i\in[N]} e^{t f(x_i;\theta_k)}\right)$. Plugging in the updating rule

$$\theta_{k+1} - \theta_k = -\alpha_k \frac{e^{t(\zeta_k;\theta_k)}}{e^{t\widetilde{\widetilde{R}}_k}} \nabla f(\zeta_k;\theta_k) = -\alpha_k e^{t(\widetilde{R}_k - \widetilde{\widetilde{R}}_k)} \nabla \widetilde{R}_k(\zeta_k) \tag{288}$$

gives

$$\widetilde{R}_{k+1} - \widetilde{R}_k + \alpha_k \langle \nabla \widetilde{R}_k, e^{t(\widetilde{R}_k - \widetilde{\widetilde{R}}_k)} \nabla \widetilde{R}_k(\zeta_k) \rangle \leqslant \frac{\beta}{2} \left\| \alpha_k e^{t(\widetilde{R}_k - \widetilde{\widetilde{R}}_k)} \nabla \widetilde{R}_k(\zeta_k) \right\|^2. \tag{289}$$

First, we note

$$\left\| \alpha_k^2 e^{t(\widetilde{R}_k - \widetilde{\widetilde{R}}_k)} \nabla \widetilde{R}_k(\zeta_k) \right\|^2 \leqslant \alpha_k^2 e^{2t(\widetilde{F}_{\max} - \widetilde{F}_{\min})} \|\nabla \widetilde{R}_k(\zeta_k)\|^2. \tag{290}$$

Take expectation on both sides of (289),

$$\mathbb{E}[\widetilde{R}_{k+1}] - \mathbb{E}[\widetilde{R}_k] + \alpha_k \mathbb{E}[\langle \nabla \widetilde{R}_k, e^{t(\widetilde{R}_k - \widetilde{\widetilde{R}}_k)} \nabla \widetilde{R}_k(\zeta_k) \rangle] \leqslant \frac{\beta \alpha_k^2 e^{2t(\widetilde{F}_{\max} - \widetilde{F}_{\min})} B^2}{2}. \tag{291}$$

Let

$$k_t := \left\lceil \frac{2(\widetilde{F}_{\max} - \widetilde{F}_{\min}) t^2 L^2}{\beta e^2} \right\rceil. \tag{292}$$

For any $k \geqslant k_t$, let

$$\alpha_k = \frac{\sqrt{2(\widetilde{F}_{\max} - \widetilde{F}_{\min})}}{e^{t(\widetilde{F}_{\max} - \widetilde{F}_{\min})}\sqrt{\beta B^2 K}}. \tag{293}$$

For $k < k_t$, let

$$\alpha_k = \frac{1}{tLBe^{t(\widetilde{F}_{\max} - \widetilde{F}_{\min})} + 1}. \tag{294}$$

We have for any $k \geqslant 1$,

$$\alpha_k tLBe^{t(\widetilde{F}_{\max} - \widetilde{F}_{\min})} \leqslant 1. \tag{295}$$

Therefore, for any $k \geqslant 1$,

$$\mathbb{E}[e^{t(\widetilde{\widetilde{R}}_k - \widetilde{R}_k)} | \theta_1, ..., \theta_k] \leqslant 2e. \tag{296}$$

Thus, for any $k \geqslant 1$,

$$\mathbb{E}[\langle \nabla \widetilde{R}_k, e^{t(\widetilde{R}_k - \widetilde{\widetilde{R}}_k)} \nabla \widetilde{R}_k(\zeta_k) \rangle] = \mathbb{E}[\mathbb{E}[\langle \nabla \widetilde{R}_k, e^{t(\widetilde{R}_k - \widetilde{\widetilde{R}}_k)} \nabla \widetilde{R}_k(\zeta_k) \rangle | \theta_1, ..., \theta_k]] \tag{297}$$

$$= \mathbb{E}[\langle \nabla \widetilde{R}_k, \mathbb{E}[e^{t(\widetilde{R}_k - \widetilde{\widetilde{R}}_k)} \nabla \widetilde{R}_k(\zeta_k) | \theta_1, ..., \theta_k] \rangle] \tag{298}$$

$$= \mathbb{E}[\langle \nabla \widetilde{R}_k, \mathbb{E}[e^{t(\widetilde{R}_k - \widetilde{\widetilde{R}}_k)} | \theta_1, ..., \theta_k] \mathbb{E}[\nabla \widetilde{R}_k(\zeta_k) | \theta_1, ..., \theta_k] \rangle] \tag{299}$$

$$= \mathbb{E}[\langle \nabla \widetilde{R}_k, \mathbb{E}[e^{t(\widetilde{R}_k - \widetilde{\widetilde{R}}_k)} | \theta_1, ..., \theta_k] \nabla \widetilde{R}_k \rangle] \tag{300}$$

$$\geqslant \frac{1}{2e} \mathbb{E}[\|\nabla \widetilde{R}_k\|^2]. \tag{301}$$

Plug (301) into (291),

$$\mathbb{E}[\|\nabla \widetilde{R}_k\|^2] + \frac{2e}{\alpha_k}(\mathbb{E}[\widetilde{R}_{k+1}] - \mathbb{E}[\widetilde{R}_k]) \leqslant \beta \alpha_k e^{2t(\widetilde{F}_{\max} - \widetilde{F}_{\min})} eB^2. \tag{302}$$

Apply telescope sum from $k_t + 1$ to K and divide both sides by K,

$$\frac{1}{K} \sum_{k=k_t}^{K} \mathbb{E}[\|\nabla \widetilde{R}_k\|^2] + \frac{2e(\mathbb{E}[\widetilde{R}_{K+1}] - \mathbb{E}[\widetilde{R}_{k_t}])}{\alpha_k K} \leqslant \beta \alpha_k e^{2t(\widetilde{F}_{\max} - \widetilde{F}_{\min})} eB^2. \tag{303}$$

$$\frac{1}{K} \sum_{k=k_t}^{K} \mathbb{E}[\|\nabla \widetilde{R}_k\|^2] \leqslant \beta \alpha_k e^{2t(\widetilde{F}_{\max} - \widetilde{F}_{\min})} eB^2 + \frac{2e(\mathbb{E}[\widetilde{R}_{k_t} - \widetilde{R}_{K+1}])}{\alpha_k K} \tag{304}$$

$$\leqslant \beta \alpha_k e^{2t(\widetilde{F}_{\max} - \widetilde{F}_{\min})} eB^2 + \frac{2e(\widetilde{F}_{\max} - \widetilde{F}_{\min})}{\alpha_k K} \tag{305}$$

Consider that $\alpha_k = \frac{\sqrt{2(\widetilde{F}_{\max} - \widetilde{F}_{\min})}}{e^{t(\widetilde{F}_{\max} - \widetilde{F}_{\min})}\sqrt{\beta B^2 K}}$,

$$\frac{1}{K} \sum_{k=k_t}^{K} \mathbb{E}[\|\nabla \widetilde{R}_k\|^2] \leqslant \sqrt{8} B e^{t(\widetilde{F}_{\max} - \widetilde{F}_{\min}) + 1} \sqrt{\frac{\beta(\widetilde{F}_{\max} - \widetilde{F}_{\min})}{K}}, \tag{306}$$

completing the proof. □

Proof of Theorem 14. From the assumptions, we have $\widetilde{R}(t;\theta)$ is $\frac{\mu}{2}$-PL, i.e.,

$$\mu(\widetilde{R}(t;\theta)-\widetilde{R}^*)\leqslant\|\nabla\widetilde{R}(t;\theta)\|^2, \tag{307}$$

where $\widetilde{R}^*:=\widetilde{R}(t;\breve{\theta}(t))$. Let

$$k_t:=\operatorname*{argmax}_k\left(k<\frac{4e}{\mu}+\frac{4etLBe^{t(\widetilde{F}_{\max}-\widetilde{F}_{\min})}}{\mu k}\right), \tag{308}$$

and

$$\alpha_k=\begin{cases}\frac{1}{tLBe^{t(\widetilde{F}_{\max}-\widetilde{F}_{\min})}+1}, & \text{if } k\leqslant k_t\\ \frac{4e}{\mu k}, & \text{otherwise.}\end{cases} \tag{309}$$

Similarly, we can prove for any $k\geqslant 1$,

$$\mathbb{E}[e^{t(\widetilde{\widetilde{R}}_k-\widetilde{R}_k)}|\theta_1,\ldots,\theta_k]\leqslant 2e. \tag{310}$$

Similarly,

$$\mathbb{E}[\widetilde{R}_{k+1}]-\mathbb{E}[\widetilde{R}_k]+\frac{\alpha_k}{2e}\mathbb{E}[\|\nabla\widetilde{R}_k\|^2]\leqslant\frac{\beta\alpha_k^2e^{2t(\widetilde{F}_{\max}-\widetilde{F}_{\min})}B^2}{2}. \tag{311}$$

Therefore,

$$\mathbb{E}[\widetilde{R}_{k+1}]-\mathbb{E}[\widetilde{R}_k]+\frac{\alpha_k}{2e}\mu\mathbb{E}[\widetilde{R}_k-\widetilde{R}^*]\leqslant\frac{\beta\alpha_k^2e^{2t(\widetilde{F}_{\max}-\widetilde{F}_{\min})}B^2}{2} \tag{312}$$

$$\mathbb{E}[\widetilde{R}_{k+1}-\widetilde{R}^*]-\mathbb{E}[\widetilde{R}_k-\widetilde{R}^*]+\frac{\alpha_k}{2e}\mu\mathbb{E}[\widetilde{R}_k-\widetilde{R}^*]\leqslant\frac{\beta\alpha_k^2e^{2t(\widetilde{F}_{\max}-\widetilde{F}_{\min})}B^2}{2} \tag{313}$$

$$\mathbb{E}[\widetilde{R}_{k+1}-\widetilde{R}^*]\leqslant\left(1-\frac{\alpha_k}{2e}\mu\right)\mathbb{E}[\widetilde{R}_k-\widetilde{R}^*]+\frac{\beta\alpha_k^2e^{2t(\widetilde{F}_{\max}-\widetilde{F}_{\min})}B^2}{2} \tag{314}$$

Let $\alpha_k=\frac{4e}{\mu k}$, and

$$V_t=\max\left\{k_t\mathbb{E}[\widetilde{R}_{k_t}-\widetilde{R}^*],\frac{8\beta B^2e^{2t(\widetilde{F}_{\max}-\widetilde{F}_{\min})+2}}{\mu^2}\right\}. \tag{315}$$

We next prove $\mathbb{E}[\widetilde{R}_k-\widetilde{R}^*]\leqslant\frac{1}{k}$ $(k\geqslant k_t)$ by induction. Suppose $\mathbb{E}[\widetilde{R}_k-\widetilde{R}^*]\leqslant\frac{V_t}{k}$, then

$$\mathbb{E}[\widetilde{R}_{k+1}-\widetilde{R}^*]\leqslant\left(1-\frac{2}{k}\right)\mathbb{E}[\widetilde{R}_k-\widetilde{R}^*]+\frac{V_t}{k^2} \tag{316}$$

$$\leqslant\left(1-\frac{2}{k}\right)\frac{V_t}{k}+\frac{V_t}{k^2} \tag{317}$$

$$\leqslant\frac{V_t}{k+1}, \tag{318}$$

which concludes the proof. □

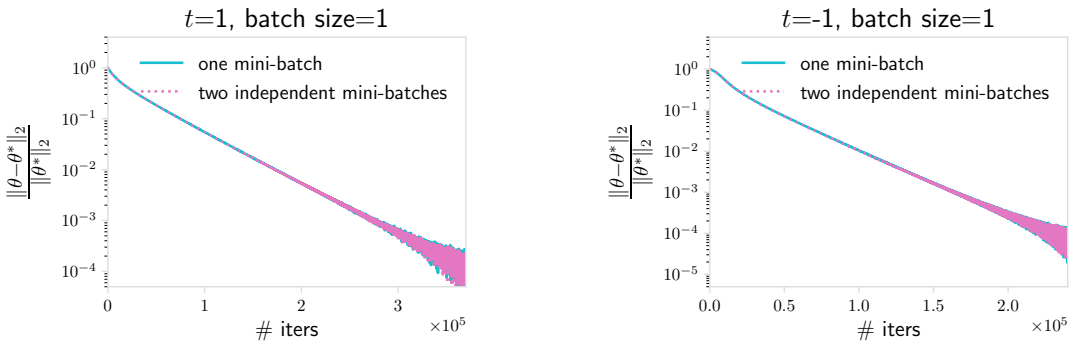


Figure 17: Convergence of Algorithm 2 using two independent mini-batches to update $\widetilde{\widetilde{R}}_t$ and calculate $e^{tf(x;\theta)}\nabla_\theta f(x;\theta)$ and a simpler variant using only one mini-batch to query $w_{t,x}\nabla_\theta f(x;\theta)$. We plot the optimality gap versus the number of iterations on the point estimation example (Figure 1 (a)) with batch size being 1. While Algorithm 2 allows us to get better convergence guarantees theoretically, we find that these two variants perform similarly empirically.

Table 10: TERM Applications and their corresponding solvers.

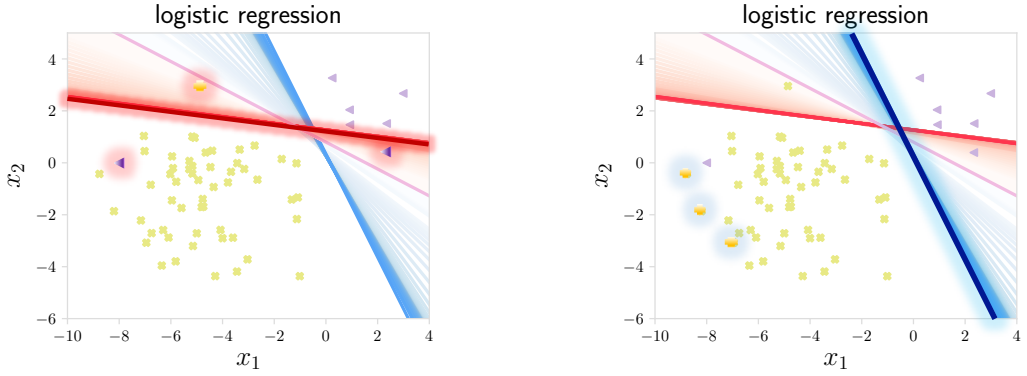
Three toy examples (Figure 1)	Algorithm 1
Robust regression (Table 2)	Algorithm 1
Robust classification (Table 4)	Algorithm 2
Low-quality annotators (Figure 9)	Algorithm 4 (τ=0)
Fair PCA (Figure 10)	Algorithm 3 (τ=0)
Class imbalance (Figure 14)	Algorithm 4 (τ=0)
Variance reduction (Table 7)	Algorithm 3 (τ=0)
Hierarchical TERM (Table 8)	Algorithm 3

Appendix D. Additional Experiments and Experimental Details

In Appendix D.1, we provide complete experimental results on the properties or the use-cases of TERM. Details on how the experiments in Section 7 were executed are provided in Appendix D.2.

D.1 Complete Results

Recall that in Section 2, Interpretation 1 is that TERM can be tuned to re-weight samples to magnify or suppress the influence of outliers. In Figure 18 below, we visually show this effect by highlighting the samples with the largest weight for $t \to +\infty$ and $t \to -\infty$ on the logistic regression example previously described in Figure 1.



(a) Samples with the largest weights as $t \to +\infty$. (b) Samples with the largest weights as $t \to -\infty$.

Figure 18: For positive values of t, TERM focuses on the samples with relatively large losses (rare instances). When $t \to +\infty$ (left), a few misclassified samples have the largest weights and are highlighted. On the other hand, for negative values of t, TERM suppresses the effect of the outliers, and as $t \to -\infty$ (right), samples with the smallest losses hold the the largest weights.

Next, we provide complete results of applying TERM to a diverse set of applications.

Robust classification. Recall that in Section 7.1, for classification in the presence of label noise, we only compare with baselines which do not require clean validation data. In Table 11 below, we report the complete results of comparing TERM with all baselines, including MentorNet-DD (Jiang et al., 2018) which needs additional clean data. In particular, in contrast to the other methods, MentorNet-DD uses 5,000 clean validation images. TERM is competitive with the performance of MentorNet-DD, even though it does not have access to this clean data.

To interpret the noise more easily, we provide a toy logistic regression example with synthetic data here. In Figure 19, we see that TERM with $t=-2$ (blue) can converge to the correct classifier under 20%, 40%, and 80% noise.

Table 11: A complete comparison including two MentorNet variants. TERM is able to match the performance of MentorNet-DD, which needs additional clean labels.

objectives	**test accuracy** (CIFAR-10, Inception)		
	20% noise	40% noise	80% noise
ERM	0.775 (.004)	0.719 (.004)	0.284 (.004)
RandomRect (Ren et al., 2018)	0.744 (.004)	0.699 (.005)	0.384 (.005)
SelfPaced (Kumar et al., 2010)	0.784 (.004)	0.733 (.004)	0.272 (.004)
MentorNet-PD (Jiang et al., 2018)	0.798 (.004)	0.731 (.004)	0.312 (.005)
GCE (Zhang and Sabuncu, 2018)	**0.805** (.004)	0.750 (.004)	0.433 (.005)
MentorNet-DD (Jiang et al., 2018)	**0.800** (.004)	**0.763** (.004)	**0.461**(.005)
TERM	0.795 (.004)	**0.768** (.004)	**0.455** (.005)
Genie ERM	0.828 (.004)	0.820 (.004)	0.792 (.004)

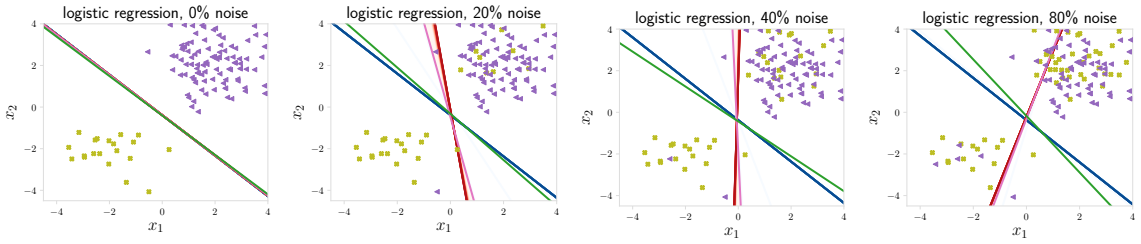


Figure 19: Robust classification using synthetic data. On this toy problem, we show that TERM with negative t's (blue) can be robust to random noisy samples. The green line corresponds to the solution of the generalized cross entropy (GCE) baseline (Zhang and Sabuncu, 2018). Note that on this toy problem, GCE is as good as TERM with negative t's, despite its inferior performance on the real-world CIFAR10 dataset.

Low-quality annotators. In Section 7.1.3, we demonstrate that TERM can be used to mitigate the effect of noisy annotators, and we assume each annotator is either always correct, or always uniformly assigning random labels. Here, we explore a different and possibly more practical scenario where there are four noisy annotators who corrupt 0%, 20%, 40%, and 100% of their data by assigning labels uniformly at random, and there is one additional adversarial annotator who always assigns wrong labels. We assume the data points labeled by each annotator do not overlap, since (Khetan et al., 2018) show that obtaining one label per sample is optimal for the data collectors under a fixed annotation budget. We compare TERM with several baselines: (a) training without the data coming from the adversarial annotator, (b) training without the data coming from the worst two annotators, and (c) training with all the clean data combined (Genie ERM). The results are shown in Figure 20. We see that TERM outperforms the strong baselines of removing one or two noisy annotators, and closely matches the performance of training with all the available clean data.

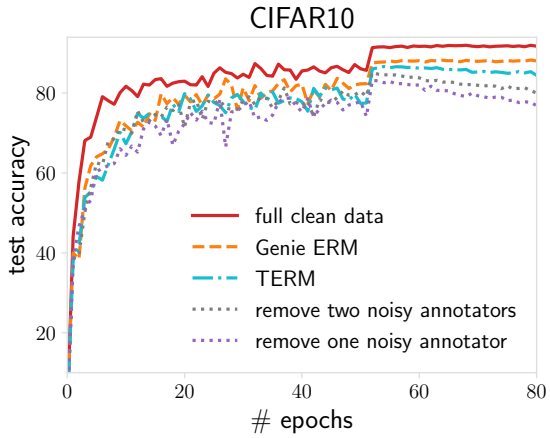


Figure 20: TERM achieves higher test accuracy than the baselines, and can match the performance of Genie ERM (i.e., training on all the clean data combined).

D.2 Experimental Details

We first describe the datasets and models used in each experiment presented in Section 7, and then provide a detailed setup including the choices of hyperparameters. All code and datasets are publicly available at github.com/litian96/TERM.

D.2.1 Datasets and models

In Section 7.1, for regression tasks, we use the drug discovery data extracted from Diakonikolas et al. (2019) which is originally curated from Olier et al. (2018) and train linear regression models with different losses. There are 4,085 samples in total with each having 411 features. We randomly split the dataset into 80% training set, 10% validation set, and 10% testing set. For mitigating noise on classification tasks, we use the standard CIFAR-10 data and their standard train/val/test partitions along with a standard inception network (Szegedy et al., 2016). For experiments regarding mitigating noisy annotators, we again use the CIFAR-10 data and their standard partitions with a ResNet20 model. The noise generation procedure is described in Section 7.1.3.

In Section 7.2, for fair PCA experiments, we use the complete Default Credit data to learn low-dimensional approximations and the loss is computed on the full training set. We follow the exact data processing steps described in the work (Samadi et al., 2018) we compare with. There are 30,000 total data points with 21-dimensional features (after preprocessing). Among them, the high education group has 24,629 samples and the low education group has 5,371 samples. For meta-learning experiments, one the popular sine wave regression problem (Finn et al., 2017), we generate 5,000 meta-training and 5,000 meta-testing tasks. Following Collins et al. (2020), there are 250 hard meta-training tasks with amplitudes drawn from [4.95,5] and 4,750 easy meta-training tasks with amplitudes drawn from [0.01,1]. The amplitudes of meta-testing tasks are drawn uniformly from [0.1,5]. The phase values are drawn uniformly from $[0,\pi]$ for all tasks. For class imbalance experiments, we directly take the unbalanced data extracted from MNIST (LeCun et al., 1998) used in Ren et al. (2018). When demonstrating

the variance reduction of TERM, we use the HIV-1 dataset (Rögnvaldsson, 2013) as in (Duchi and Namkoong, 2019) and randomly split it into 80% train, 10% validation, and 10% test set. There are 6,590 total samples and each has 160 features. We report results based on five such random partitions of the data. We train logistic regression models (without any regularization) for this binary classification task for TERM and the baseline methods. We also investigate the performance of a linear SVM.

In Section 7.3, the HIV-1 data are the same as that in Section 7.2. We also manually subsample the data to make it more imbalanced, or inject random noise, as described in Section 7.3. The CIFAR10 dataset used in this section is a standard benchmark, and we follow the same procedures in Cao et al. (2021) to generate a noisy and imbalanced variant.

D.2.2 Hyperparameters

Selecting t. In Section 7.2 where we consider positive t's, we select t from a limited candidate set of {0.1,1,2,5,10,50,100,200} on the held-out validation set. For initial robust regression experiments, RMSE changed by only 0.08 on average across t; we thus used $t=-2$ for all experiments involving noisy training samples (Section 7.1 and Section 7.3).

Other parameters. For all experiments, we tune all other hyperparameters (the learning rates, the regularization parameters, the decision threshold for ERM_+, ρ for (Duchi and Namkoong, 2019), the quantile value for CVaR (i.e., α in Eq. (62)) (Rockafellar et al., 2000), α and γ for focal loss (Lin et al., 2017)) based on a validation set, and select the best one. For experiments regarding focal loss (Lin et al., 2017), we select the class balancing parameter (α in the original focal loss paper) from `range`(0.05,0.95,0.05) and select the main parameter γ from {0.5,1,2,3,4,5}. We tune ρ in (Duchi and Namkoong, 2019) such that $\frac{\rho}{n}$ is selected from {0.5,1,2,3,4,5,10} where n is the training set size. We tune α for CVaR from {0.1,0.3,0.5,0.7,0.9}. All regularization parameters including regularization for linear SVM are selected from {0.0001,0.01,0.1,1,2}. For all experiments on the baseline methods, we use the default hyperparameters in the original paper (or the open-sourced code).

We summarize a complete list of main hyperparameter values as follows.

Section 7.1:

- Robust regression. The threshold parameter δ for Huber loss for all noisy levels is 1, the corruption parameter k for CRR is: 500 (20% noise), 1000 (40% noise), and 3000 (80% noise); and TERM uses $t=-2$.

- Robust classification. The results are all based on the default hyperparameters provided by the open-sourced code of MentorNet (Jiang et al., 2018), if applicable. We tune the q parameter for generalized cross entropy (GCE) from {0.4,0.8,1.0} and select a best one for each noise level. For TERM, we scale t linearly as the number of iterations from 0 to -2 for all noise levels.

- Low-quality annotators. For all methods, we use the same set of hyperparameters. The initial step-size is set to 0.1 and decayed to 0.01 at epoch 50. The batch size is 100.

Section 7.2:

- Fair PCA. We use the default hyperparameters and directly run the public code of (Samadi et al., 2018) to get the results on the min-max fairness baseline. We use a learning rate of 0.001 for our gradient-based solver for all target dimensions.

- Fair meta-learning. We use a fixed learning rate of 0.01 for all methods, and tune a best learning rate for the task weights for the work of Collins et al. (2020). Similar as Collins et al. (2020), for all methods, we run one step of mini-batch SGD for inner optimization.

- Handling class imbalance. We take the open-sourced code of LearnReweight (Ren et al., 2018) and use the default hyperparameters for the baselines of LearnReweight, HardMine, and ERM. We implement focal loss, and select α=0.05,γ=2.

- Variance reduction. The regularization parameter for linear SVM is 1. γ for focal loss is 2. We perform binary search on the decision thresholds for ERM_+ and RobustRegRisk_+, and choose 0.26 and 0.49, respectively.

Section 7.3:

- Logistic regression on HIV. We tune the q parameter for GCE based on validation data. We use q=0,0,0.7,0.3 respectively for the four scenarios we consider. For RobustlyRegRisk, we use $\frac{\rho}{n}$=10 (where n is the training sample size) and we find that the performance is not sensitive to the choice of ρ. For CVaR, the tuned α value is 0.5 when the data imbalance ratio is 1:4, and 0.1 when the imbalance ratio is 1:20. For focal loss, we tune the hyperparameters for best performance and select γ=2, α=0.5, 0.1, 0.5, and 0.2 for four scenarios. For HAR, we tune the regularization parameter λ via grid search from {0.1,1,2,5,10} and select the best one. We use $t=-2$ for TERM in the presence of noise, and tune the positive t's based on validation data. In particular, the values of tilts under four cases are: (0, 0.1), (0, 50), (-2, 5), and (-2, 10) for TERM_{sc} and (0.1, 0), (50, 0), (1, -2) and (50, -2) for TERM_{ca}.

- ResNet32 on CIFAR10. We reproduce (and then directly take) the results from (Cao et al., 2021) for all baseline methods. For hierarchical TERM, we scale t from 0 to 3 for group-level tilting, and scale t from 0 to -2 for sample-level tilting within each group. λ is set to 0.2. We use the default hyperparameters (batch size, learning rate, etc) in the open-sourced code of HAR (Cao et al., 2021) for TERM.

References

Sherif Abdelkarim, Panos Achlioptas, Jiaji Huang, Boyang Li, Kenneth Church, and Mohamed Elhoseiny. Long-tail visual relationship recognition with a visiolinguistic hubless loss. *arXiv preprint arXiv:2004.00436*, 2020.

Jacob Abernethy, Pranjal Awasthi, Matthäus Kleindessner, Jamie Morgenstern, Chris Russell, and Jie Zhang. Active sampling for min-max fairness. In *International Conference on Machine Learning*, 2022.

Amir Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 2012.

Wael Alghamdi, Shahab Asoodeh, Hao Wang, Flavio P Calmon, Dennis Wei, and Karthikeyan Natesan Ramamurthy. Model projection: Theory and applications to fair machine learning. In *IEEE International Symposium on Information Theory*, 2020.

Erdal Arıkan. An inequality on guessing and its application to sequential decoding. *IEEE Transactions on Information Theory*, 1996.

Philippe Artzner. Thinking coherently. *Risk*, 1997.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 1999.

Sina Baharlouei, Maher Nouiehed, Ahmad Beirami, and Meisam Razaviyayn. Rényi fair inference. *International Conference on Learning Representations*, 2020.

Ahmad Beirami, Robert Calderbank, Mark M Christiansen, Ken R Duffy, and Muriel Médard. A characterization of guesswork on swiftly tilting curves. *IEEE Transactions on Information Theory*, 2018.

George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 1962.

Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 1997.

Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, 2015.

Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar. Consistent robust regression. In *Advances in Neural Information Processing Systems*, 2017.

Vivek S Borkar. Q-learning for risk-sensitive control. *Mathematics of Operations Research*, 2002.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. 2013.

Ronald W Butler. *Saddlepoint approximations with applications*. Cambridge University Press, 2007.

Giuseppe C Calafiore and Laurent El Ghaoui. *Optimization Models*. Cambridge University Press, 2014.

Kaidi Cao, Yining Chen, Junwei Lu, Nikos Arechiga, Adrien Gaidon, and Tengyu Ma. Heteroskedastic and imbalanced deep learning with adaptive regularization. In *International Conference on Learning Representations*, 2021.

Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples. In *Advances in Neural Information Processing Systems*, 2017.

Ruidi Chen and Ioannis Ch Paschalidis. Distributionally robust learning. *Foundations and Trends® in Optimization*, 2020.

Nadav Cohen and Amnon Shashua. Simnets: A generalization of convolutional networks. *arXiv preprint arXiv:1410.0781*, 2014.

Nadav Cohen, Or Sharir, and Amnon Shashua. Deep simnets. In *Conference on Computer Vision and Pattern Recognition*, 2016.

Liam Collins, Aryan Mokhtari, and Sanjay Shakkottai. Task-robust model-agnostic meta-learning. In *Advances in Neural Information Processing Systems*, 2020.

Andrew Cotter, Heinrich Jiang, Maya R Gupta, Serena Wang, Taman Narayan, Seungil You, and Karthik Sridharan. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research*, 2019.

Thomas M Cover and Joy A Thomas. Information theory and statistics. *Elements of Information Theory*, 1991.

Sebastian Curi, Kfir Y Levy, Stefanie Jegelka, and Andreas Krause. Adaptive sampling for stochastic risk-averse learning. In *Advances in Neural Information Processing Systems*, 2020.

A. Dembo and O. Zeitouni. *Large deviations techniques and applications*. Springer Science & Business Media, 2009.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *International Conference on Machine Learning*, 2019.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, 2018.

D Dua and C Graff. UCI machine learning repository [http://archive. ics. uci. edu/ml]. https://archive. ics. uci. edu/ml/datasets. 2019.

Marco F Duarte and Yu Hen Hu. Vehicle classification in distributed sensor networks. *Journal of Parallel and Distributed Computing*, 2004.

John Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

John Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 2019.

Paul Dupuis and Richard S Ellis. *A Weak Convergence Approach to the Theory of Large Deviations*. 1997.

Yanbo Fan, Siwei Lyu, Yiming Ying, and Baogang Hu. Learning with average top-k loss. In *Advances in Neural Information Processing Systems*, 2017.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 2017.

Hans Föllmer and Thomas Knispel. Entropic risk measures: Coherence vs. convexity, model ambiguity and robust large deviations. *Stochastics and Dynamics*, 2011.

Hans Föllmer and Alexander Schied. Stochastic finance: an introduction in discrete time. 2004.

Robert G Gallager. *Information theory and reliable communication*. Springer, 1968.

Jinyang Gao, HV Jagadish, and Beng Chin Ooi. Active sampler: Light-weight accelerator for complex data analytics at scale. *arXiv preprint arXiv:1512.03880*, 2015.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, 2015.

Saeed Ghadimi, Andrzej Ruszczynski, and Mengdi Wang. A single timescale stochastic approximation method for nested stochastic optimization. *SIAM Journal on Optimization*, 2020.

Mert Gürbüzbalaban, Andrzej Ruszczyński, and Landi Zhu. A stochastic subgradient method for distributionally robust non-convex and non-smooth learning. *Journal of Optimization Theory and Applications*, 2022.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 2016.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, 2018.

Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In *Advances in Neural Information Processing Systems*, 2018.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*. 1994.

Matthew Holland and Kazushi Ikeda. Better generalization with less data using robust gradient descent. In *International Conference on Machine Learning*, 2019.

Ronald A Howard and James E Matheson. Risk-sensitive markov decision processes. *Management science*, 1972.

Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 1964.

Angela H Jiang, Daniel L-K Wong, Giulio Zhou, David G Andersen, Jeffrey Dean, Gregory R Ganger, Gauri Joshi, Michael Kaminksy, Michael Kozuch, Zachary C Lipton, et al. Accelerating deep learning by focusing on the biggest losers. *arXiv preprint arXiv:1910.00762*, 2019.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, 2018.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, 2017.

Chi Jin, Praneeth Netrapalli, and Michael Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *International Conference on Machine Learning*, 2020.

Ishan Jindal, Matthew Nokleby, and Xuewen Chen. Learning deep networks from noisy labels with dropout regularization. In *International Conference on Data Mining*, 2016.

Ishan Jindal, Daniel Pressel, Brian Lester, and Matthew Nokleby. An effective label noise model for dnn text classification. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

Philippe Jorion. Value at risk: a new benchmark for measuring derivatives risk. *Irwin Professional Pub*, 1996.

Dionysios S Kalogerias and Warren B Powell. Recursive optimization of convex risk measures: Mean-semideviation models. *arXiv preprint arXiv:1804.00636*, 2018.

Mohammad Mahdi Kamani, Farzin Haddadpour, Rana Forsati, and Mehrdad Mahdavi. Efficient fair principal component analysis. *arXiv preprint arXiv:1911.04931*, 2019.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2016.

Angelos Katharopoulos and François Fleuret. Biased importance sampling for deep neural network training. *arXiv preprint arXiv:1706.00043*, 2017.

Ashish Khetan, Zachary C Lipton, and Anima Anandkumar. Learning from noisy singly-labeled data. In *International Conference on Learning Representations*, 2018.

Barry W Kort and Dimitri P Bertsekas. A new penalty function method for constrained minimization. In *IEEE Conference on Decision and Control and 11th Symposium on Adaptive Processes*, 1972.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

M Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, 2010.

Yassine Laguel, Krishna Pillutla, Jérôme Malick, and Zaid Harchaoui. A superquantile approach for federated learning with heterogeneous devices. In *Annual Conference on Information Sciences and Systems*, 2021.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in Neural Information Processing Systems*, 2020.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

Liu Leqi, Adarsh Prasad, and Pradeep K Ravikumar. On human-aligned risk minimization. In *Advances in Neural Information Processing Systems*, 2019.

Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. In *Advances in Neural Information Processing Systems*, 2020.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 2020a.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *International Conference on Learning Representations*, 2020b.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. In *International Conference on Learning Representations*, 2021.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *International Conference on Computer Vision*, 2017.

Guan-Horng Liu and Evangelos A Theodorou. Deep learning theory review: An optimal control and dynamical systems perspective. *arXiv preprint arXiv:1908.10920*, 2019.

Andrew Lowy and Meisam Razaviyayn. Output perturbation for differentially private convex optimization with improved population loss bounds, runtimes and applications to private adversarial training. *arXiv preprint arXiv:2102.04704*, 2021.

Andrew Lowy, Sina Baharlouei, Rakesh Pavan, Meisam Razaviyayn, and Ahmad Beirami. A stochastic optimization framework for fair risk minimization. *arXiv preprint arXiv:2102.12586*, 2021.

Tomasz Malisiewicz, Abhinav Gupta, and Alexei A Efros. Ensemble of exemplar-SVMs for object detection and beyond. In *International Conference on Computer Vision*, 2011.

Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean. Boosting algorithms as gradient descent. In *Advances in Neural Information Processing Systems*, 1999.

James L Massey. Guessing and entropy. In *IEEE International Symposium on Information Theory*, 1994.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*, 2017.

Aditya Krishna Menon, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Can gradient clipping mitigate label noise? In *International Conference on Learning Representations*, 2020.

Neri Merhav. List decoding—Random coding exponents and expurgated exponents. *IEEE Transactions on Information Theory*, 2014.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, 2019.

Bhaskar Mukhoty, Govind Gopakumar, Prateek Jain, and Purushottam Kar. Globally-convergent iteratively reweighted least squares for robust regression problems. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Hongseok Namkoong and John C Duchi. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems*, 2017.

Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems*, 2019.

Ivan Olier, Noureddin Sadawi, G Richard Bickerton, Joaquin Vanschoren, Crina Grosan, Larisa Soldatova, and Ross D King. Meta-qsar: a large-scale application of meta-learning to drug design and discovery. *Machine Learning*, 2018.

Dmitrii M Ostrovskii, Andrew Lowy, and Meisam Razaviyayn. Efficient search of first-order nash equilibria in nonconvex-concave smooth min-max problems. *arXiv preprint arXiv:2002.07919*, 2020.

R Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 1997.

EY Pee and Johannes O Royset. On solving large-scale finite minimax problems using exponential smoothing. *Journal of Optimization Theory and Applications*, 2011.

Flavien Prost, Hai Qian, Qiuwen Chen, Ed H Chi, Jilin Chen, and Alex Beutel. Toward a better trade-off between performance and fairness with kernel-based distribution matching. *arXiv preprint arXiv:1910.11779*, 2019.

Qi Qi, Zhishuai Guo, Yi Xu, Rong Jin, and Tianbao Yang. A practical online method for distributionally deep robust optimization. *arXiv preprint arXiv:2006.10138*, 2020a.

Qi Qi, Yi Xu, Rong Jin, Wotao Yin, and Tianbao Yang. Attentional biased stochastic gradient for imbalanced classification. *arXiv preprint arXiv:2012.06951*, 2020b.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, 2018.

Ashkan Rezaei, Rizal Fathony, Omid Memarrast, and Brian D Ziebart. Fairness for robust log loss classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

R Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 2002.

R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2000.

Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Fr-train: A mutual information-based approach to fair and robust training. In *International Conference on Machine Learning*, 2020.

Thorsteinn Rögnvaldsson. UCI repository of machine learning databases. *https://archive.ics.uci.edu/ml/datasets/HIV-1+protease+cleavage*, 2013.

Salman Salamatian, Litian Liu, Ahmad Beirami, and Muriel Médard. Mismatched guesswork. *arXiv preprint arXiv:1907.00531*, 2019.

Samira Samadi, Uthaipon Tantipongpipat, Jamie H Morgenstern, Mohit Singh, and Santosh Vempala. The price of fair PCA: One extra dimension. In *Advances in Neural Information Processing Systems*, 2018.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. 2014.

Chunhua Shen and Hanxi Li. On the dual formulation of boosting algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Conference on Computer Vision and Pattern Recognition*, 2016.

Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *Advances in Neural Information Processing Systems*, 2019a.

Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *Advances in Neural Information Processing Systems*, 2019b.

David Siegmund. Importance sampling in the monte carlo study of sequential tests. *The Annals of Statistics*, 1976.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

Agnieszka Słowik and Léon Bottou. On distributionally robust optimization and data rebalancing. In *International Conference on Artificial Intelligence and Statistics*, 2022.

Tasuku Soma and Yuichi Yoshida. Statistical learning with conditional value at risk. *arXiv preprint arXiv:2002.05826*, 2020.

Ivan Stelmakh, Nihar B Shah, and Aarti Singh. PeerReview4All: Fair and accurate reviewer assignment in peer review. In *Algorithmic Learning Theory*, 2019.

Tyler Sypherd, Mario Diaz, John Kevin Cava, Gautam Dasarathy, Peter Kairouz, and Lalitha Sankar. A tunable loss function for robust classification: Calibration, landscape, and generalization. *arXiv preprint arXiv:1906.02314*, 2019.

Attila Szabo, Hadi Jamali-Rad, and Siva-Datta Mannava. Tilted cross entropy (TCE): Promoting fairness in semantic segmentation. *arXiv preprint arXiv:2103.14051*, 2021.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Conference on Computer Vision and Pattern Recognition*, 2016.

Uthaipon Tantipongpipat, Samira Samadi, Mohit Singh, Jamie H Morgenstern, and Santosh Vempala. Multi-criteria dimensionality reduction with applications to fairness. In *Advances in Neural Information Processing Systems*, 2019.

Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *Conference on Computer Vision and Pattern Recognition*, 2017.

Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems*, 2018.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 2008.

Martin J Wainwright, Tommi S Jaakkola, and Alan S Willsky. A new class of upper bounds on the log partition function. *IEEE Transactions on Information Theory*, 2005.

Mengdi Wang, Ji Liu, and Ethan Fang. Accelerating stochastic composition optimization. In *Advances in Neural Information Processing Systems*, 2016a.

Mengdi Wang, Ethan X Fang, and Han Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 2017.

Xueqin Wang, Yunlu Jiang, Mian Huang, and Heping Zhang. Robust variable selection with exponential squared loss. *Journal of the American Statistical Association*, 2013.

Zhiguang Wang, Tim Oates, and James Lo. Adaptive normalized risk-averting training for deep neural networks. In *AAAI Conference on Artificial Intelligence*, 2016b.

Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 1912.

Jun Yang and Jeffrey S Rosenthal. Complexity results for mcmc derived from quantitative bounds. *arXiv preprint arXiv:1708.00829*, 2017.

Min Yang, Linli Xu, Martha White, Dale Schuurmans, and Yao-liang Yu. Relaxed clipping: A global training method for robust regression and classification. In *Advances in Neural Information Processing Systems*, 2010.

I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 2009.

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

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Conference on World Wide Web*, 2017.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems*, 2018.

Tianyi Zhou, Shengjie Wang, and Jeff Bilmes. Robust curriculum learning: from clean label detection to noisy label self-correction. In *International Conference on Learning Representations*, 2021.

Landi Zhu, Mert Gürbüzbalaban, and Andrzej Ruszczyński. Distributionally robust learning with weakly convex losses: Convergence rates and finite-sample guarantees. *arXiv preprint arXiv:2301.06619*, 2023.