Tilted Empirical Risk Minimization

Tian Li [*1] **Ahmad Beirami** [*2] **Maziar Sanjabi** [2] **Virginia Smith** [1]

Abstract

Empirical risk minimization (ERM) is typically designed to perform well on the average loss, which can result in estimators that are sensitive to outliers, generalize poorly, or treat subgroups unfairly. While many methods aim to address these problems individually, in this work, we explore them through a unified framework—tilted empirical risk minimization (TERM). In particular, we show that it is possible to flexibly tune the impact of individual losses through a straightforward extension to ERM using a hyperparameter called the tilt. We provide several interpretations of the resulting framework: 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 a superquantile method. We develop batch and stochastic first-order optimization methods for solving TERM, and show that the problem can be efficiently solved relative to common alternatives. Finally, we demonstrate that TERM can be used for a multitude of applications, such as enforcing fairness between subgroups, mitigating the effect of outliers, and handling class imbalance. TERM is not only competitive with existing solutions tailored to these individual problems, but can also enable entirely new applications, such as simultaneously addressing outliers and promoting fairness.[3]

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:

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

While ERM is widely used and offers nice statistical properties, it can also perform poorly in practical situations where average performance is not an appropriate surrogate for the objective 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 or outliers (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 (Samadi et al., 2018; Li et al., 2020b; Mohri et al., 2019), or developing solutions with favorable out-of-sample performance (Namkoong and Duchi, 2017).

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:[4]

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

TERM generalizes ERM as the 0-tilted loss recovers the average loss, i.e., $\widetilde{R}(0, \theta) = \overline{R}(\theta)$. It also recovers other common alternatives, e.g., $t \to +\infty$ recovers the max-loss, and $t \to -\infty$ the min-loss (Lemma 2). For $t > 0$, the objective is a common form of exponential smoothing, used to approximate the max (Kort and Bertsekas, 1972; Pee and Royset, 2011). A more general notion of "tilting" has also been studied in statistics, though for very different purposes, such as importance sampling and large deviations theory (Dembo and Zeitouni, 2009; Beirami et al., 2018; Wainwright et al., 2005) (Appendix B in (Li et al., 2020a)).

To highlight how the TERM objective can help with issues such as outliers or imbalanced classes, we discuss three motivating 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

[*]Equal contribution [1]Carnegie Mellon University, Pittsburgh, PA [2]Facebook AI, Menlo Park, CA. Correspondence to: Tian Li <tianli@cmu.edu>, Ahmad Beirami <beirami@fb.com>.

Presented at the ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning.

[3]See Li et al. (2020a) for an extended version of this paper that contains full statements and proofs of the results, the full set of experiments, and the appendices referred to in this paper.

[4]$\widetilde{R}(0; \theta)$ is defined via continuous extension of $R(t; \theta)$.

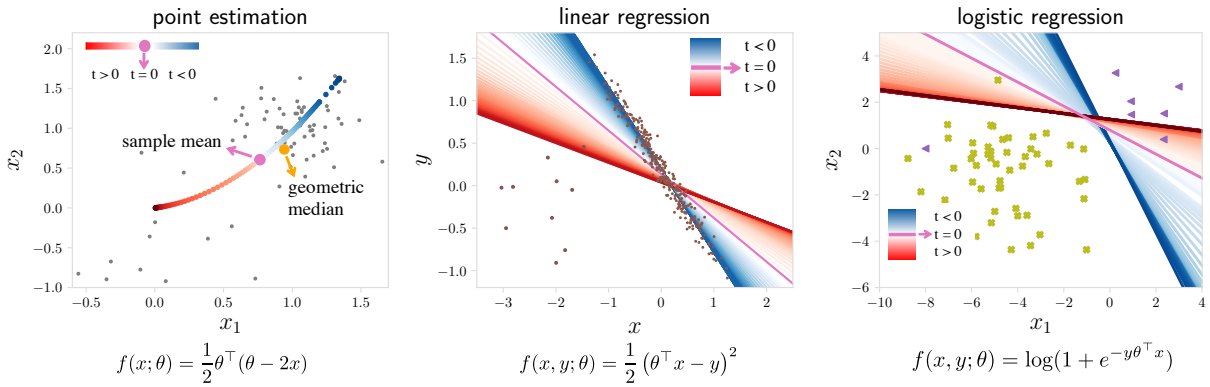

Figure 1: Toy examples illustrating TERM as a function of t: (a) finding a point estimate from a set of 2D samples, (b) linear regression with outliers, and (c) logistic regression with imbalanced classes. While positive values of t magnify outliers, negative values suppress them. Setting t=0 recovers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

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 more rigorously in Section 2, Theorem 4.

(b) Linear regression: A similar interpretation holds for the case of linear regression (Figure 2b). As $t \to -\infty$, TERM is able to find a solution that captures the underlying data while ignoring outliers. However, this solution may not be preferred if we have reason to believe that the outlier values should not be ignored. As $t \to +\infty$, TERM recovers the minimax 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 (Samadi et al., 2018; Mohri et al., 2019). We explore several use-cases involving robust regression and fairness in more detail in Section 4.

(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 4, 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.

Contributions. We propose TERM as a simple, unified framework to flexibly address various challenges with empirical risk minimization. We analyze the objective to understand its behavior with varying t, and develop efficient methods for solving TERM. We also extend TERM to handle compound issues, such as the simultaneous existence of noisy samples and imbalanced classes. Empirically, we show via multiple case studies that TERM is competitive with existing, problem-specific state-of-the-art solutions.

2. Tilted Empirical Risk Minimization: Properties & Interpretations

To better understand the performance of the t-tilted losses in (2), we provide several interpretations of the TERM solutions. See (Li et al., 2020a) for the full statements of theorems and proofs. 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 (Appendix A in (Li et al., 2020a)). However, we also obtain favorable empirical results using TERM with other objectives such as deep neural networks and PCA in Section 4, motivating the extension of our theory beyond GLMs in future work.

General properties. We begin by noting several general properties of the TERM objective (2). Given a smooth $f(x;\theta)$, the t-tilted loss is smooth for all finite t (Lemma 4). If $f(x;\theta)$ is strongly convex, the t-tilted loss is strongly convex for $t > 0$ (Lemma 3). We visualize the solutions to TERM for a toy problem in Figure 2, which allows us to illustrate several special cases of the general framework. As discussed in Section 1, TERM can recover traditional ERM (t=0), the max-loss ($t \to +\infty$), and the min-loss ($t \to -\infty$).

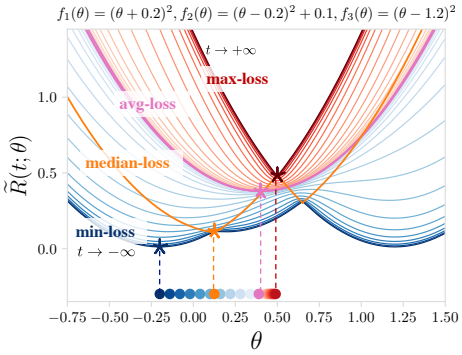

Figure 2: TERM objectives for a squared loss problem with $N = 3$. As t moves from $-\infty$ to $+\infty$, t-tilted losses recover min-loss, avg-loss, and max-loss, and approximate median-loss. TERM is smooth for all finite t and convex for positive t.

As we demonstrate in Section 4, providing a smooth tradeoff 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. Interestingly, we additionally show that the TERM objective can be viewed as a smooth approximation to a *superquantile* method, which aims to minimize quantiles of losses such as the median loss. 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.

Interpretations. We make these rough connections more explicit via four interpretations of the TERM objective. We defer more detailed interpretations and the corresponding proofs to the full version of the paper Li et al. (2020a).

1. TERM can be viewed as an implicit re-weighting of samples (using t) to magnify/suppress outliers (Lemma 1).

2. TERM allows for tradeoffs between average-loss and min/max-loss (Figure 8 in (Li et al., 2020a)). For positive t's, TERM enables a smooth tradeoff between the average-loss and max-loss, thus promoting uniformity/fairness (Theorem 2). For negative values of t, the solution trades average-loss for min-loss, which has the benefit of mitigating outliers (Theorem 3).

3. The variance of the loss across samples will decrease as t increases (Theorem 4), allowing to achieve potentially better bias-variance tradeoff for better generalization.

4. TERM approximates Value-at-Risk (VaR) or the superquantile method, which aims to minimize the the specific quantiles of the individual losses (Rockafellar et al., 2000).

3. Solving TERM

While the main focus of this work is in understanding properties of the TERM objective and its minimizers, we also develop first-order batch and stochastic methods for solving TERM, and explore the effect of t on the convergence (Section 4 in Li et al. (2020a)). We find that these methods perform well empirically on a variety of tasks (Section 5 in Li et al. (2020a)).

4. Case Studies

We showcase the flexibility, wide applicability, and competitive performance of TERM through empirical results on a variety of real-world problems such as handling outliers, ensuring fairness, and improving generalization, and addressing compound issues. 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 baselines on a wide range of applications. Due to space constraints, we show only two examples below (one with positive t, one negative t). See complete use-cases in Section 5 of the full version. [5]

Robust classification ($t < 0$). It is well-known that 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. It has two variants: MentorNet-PD which does not require clean validation data and MentorNet-DD which does. Following the setup in (Jiang et al., 2018), we explore classification on CIFAR-10 (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 1, TERM performs competitively with 20% noise, and outperforms all baselines without additional clean data in the high noise regimes. In particular, in contrast to the other methods, MentorNet-DD uses 5,000 clean validation images. TERM is competitive with can even exceed the performance of MentorNet-DD, even though it does not have access to this clean data.

Table 1: TERM is competitive with robust classification baselines, and is superior in high noise regimes.

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)

[5] Code available at: github.com/litian96/TERM

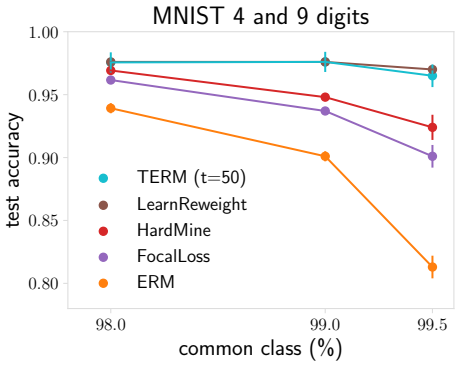

Figure 3: TERM ($t = 50$) is competitive with state-of-the-art methods for classification with imbalanced classes.

Handling class imbalance ($t > 0$)**.** 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 focal loss (Lin et al., 2017), HardMine (Malisiewicz et al., 2011), and LearnReweight (Ren et al., 2018). Following Ren et al. (2018), the datasets are composed of imbalanced 4 and 9 digits from MNIST (LeCun et al., 1998). From Figure 3, we see that TERM obtains similar (or higher) final accuracy on the clean test data as the state-of-the-art methods. We also 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.

5. Related Work

Alternate aggregation schemes: exponential smoothing/superquantile methods. A common alternative to the standard average loss in empirical risk minimization is to consider a minimax objective, which aims to minimize the max-loss. Minimax objectives are commonplace in machine learning, and have been used for a wide range of applications, such as ensuring fairness across subgroups (Mohri et al., 2019; Stelmakh et al., 2019; Samadi et al., 2018; Tantipongpipat et al., 2019; Hashimoto et al., 2018), 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., 2020), directly trading off between robustness/fairness and utility for positive and negative values of t (see Appendix B in (Li et al., 2020a) 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 4). Interestingly, Cohen et al. (Cohen and Shashua, 2014; Cohen et al., 2016) introduce Simnets, with a similar exponential smoothing operator, though for a differing purpose of flexibly achieving layer operations *between* sum and max in deep neural networks.

Alternate loss functions. Rather than modifying the way the losses are aggregated, as in (smoothed) minimax 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 as a way to mitigate the effect of outliers (Bhatia et al., 2015). 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). On the other extreme, the largest losses can be magnified in order to encourage focus on hard samples (Lin et al., 2017; Wang et al., 2016; Li et al., 2020b), which is a popular approach for curriculum learning. Ignoring the log portion of the objective in (2), TERM can in fact 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 5.3).

Sample re-weighting schemes. Finally, 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" examples (Shrivastava et al., 2016; Jiang et al., 2019; Katharopoulos and Fleuret, 2017). 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 of crowd-worker qualities (Khetan et al., 2018), which can make the objective more robust to rare instances. Such an explicit re-weighting has also been explored for other applications (e.g., Lin et al., 2017; Jiang et al., 2018; Shu et al., 2019; 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 1), which could be viewed as a convexified version of loss clipping. We compare to several sample re-weighting schemes empirically in Section 5 in Li et al. (2020a).

Acknowledgements

We are grateful to Arun Sai Suggala and Adarsh Prasad (CMU) for their helpful comments on robust regression; to Zhiguang Wang, Dario Garcia Garcia, Alborz Geramifard, and other members of Facebook AI for productive discussions and feedback and pointers to prior work (Cohen and Shashua, 2014; Cohen et al., 2016; Wang et al., 2016; Rockafellar et al., 2000); and to Meisam Razaviyayn (USC) for helpful discussions and pointers to exponential smoothing (Kort and Bertsekas, 1972; Pee and Royset, 2011), Value-at-Risk (Rockafellar and Uryasev, 2002; Nouiehed et al., 2019), and general properties of gradient-based methods in non-convex optimization problems (Jin et al., 2017; 2019; Ge et al., 2015; Ostrovskii et al., 2020). The work of TL and VS was supported in part by the National Science Foundation grant IIS1838017, a Google Faculty Award, a Carnegie Bosch Institute Research Award, and the CONIX Research Center. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the National Science Foundation or any other funding agency.

References

S. Abdelkarim, P. Achlioptas, J. Huang, B. Li, K. Church, and M. Elhoseiny. Long-tail visual relationship recognition with a visiolinguistic hubless loss. *arXiv preprint arXiv:2004.00436*, 2020.

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

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

H.-S. Chang, E. Learned-Miller, and A. McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples. In *Advances in Neural Information Processing Systems*, 2017.

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

N. Cohen, O. Sharir, and A. Shashua. Deep simnets. In *Conference on Computer Vision and Pattern Recognition*, 2016.

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

J. Gao, H. Jagadish, and B. C. Ooi. Active sampler: Light-weight accelerator for complex data analytics at scale. *arXiv preprint arXiv:1512.03880*, 2015.

R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, 2015.

T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, 2018.

A. H. Jiang, D. L.-K. Wong, G. Zhou, D. G. Andersen, J. Dean, G. R. Ganger, G. Joshi, M. Kaminksy, M. Kozuch, Z. C. Lipton, et al. Accelerating deep learning by focusing on the biggest losers. *arXiv preprint arXiv:1910.00762*, 2019.

L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, 2018.

C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, 2017.

C. Jin, P. Netrapalli, and M. I. Jordan. Minmax optimization: Stable limit points of gradient descent ascent are locally optimal. *arXiv preprint arXiv:1902.00618*, 2019.

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

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

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

A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

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

Y. Laguel, K. Pillutla, J. Malick, and Z. Harchaoui. Device heterogeneity in federated learning: A superquantile approach. *arXiv preprint arXiv:2002.11223*, 2020.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

T. Li, A. Beirami, M. Sanjabi, and V. Smith. Tilted empirical risk minimization. *arXiv preprint arXiv:2007.01162*, 2020a.

T. Li, M. Sanjabi, A. Beirami, and V. Smith. Fair resource allocation in federated learning. In *International Conference on Learning Representations*, 2020b.

T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *International Conference on Computer Vision*, 2017.

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

A. K. Menon, A. S. Rawat, S. J. Reddi, and S. Kumar. Can gradient clipping mitigate label noise? In *International Conference on Learning Representations*, 2020.

M. Mohri, G. Sivek, and A. T. Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, 2019.

H. Namkoong and J. C. Duchi. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems*, 2017.

M. Nouiehed, J.-S. Pang, and M. Razaviyayn. On the pervasiveness of difference-convexity in optimization and statistics. *Mathematical Programming*, 2019.

D. M. Ostrovskii, A. Lowy, and M. Razaviyayn. Efficient search of first-order nash equilibria in nonconvex-concave smooth min-max problems. *arXiv preprint arXiv:2002.07919*, 2020.

E. Pee and J. O. Royset. On solving large-scale finite minimax problems using exponential smoothing. *Journal of Optimization Theory and Applications*, 2011.

M. Ren, W. Zeng, B. Yang, and R. Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, 2018.

R. T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 2002.

R. T. Rockafellar, S. Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2000.

S. Samadi, U. Tantipongpipat, J. H. Morgenstern, M. Singh, and S. Vempala. The price of fair PCA: One extra dimension. In *Advances in Neural Information Processing Systems*, 2018.

A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. In *Conference on Computer Vision and Pattern Recognition*, 2016.

J. Shu, Q. Xie, L. Yi, Q. Zhao, S. Zhou, Z. Xu, and D. Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *Advances in Neural Information Processing Systems*, 2019.

A. Sinha, H. Namkoong, and J. Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

I. Stelmakh, N. B. Shah, and A. Singh. Peerreview4all: Fair and accurate reviewer assignment in peer review. In *Algorithmic Learning Theory*, 2019.

U. Tantipongpipat, S. Samadi, M. Singh, J. H. Morgenstern, and S. Vempala. Multi-criteria dimensionality reduction with applications to fairness. In *Advances in Neural Information Processing Systems*, 2019.

R. Volpi, H. Namkoong, O. Sener, J. C. Duchi, V. Murino, and S. Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems*, 2018.

M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. A new class of upper bounds on the log partition function. *IEEE Transactions on Information Theory*, 2005.

Z. Wang, T. Oates, and J. Lo. Adaptive normalized risk-averting training for deep neural networks. In *AAAI Conference on Artificial Intelligence*, 2016.

M. Yang, L. Xu, M. White, D. Schuurmans, and Y.-l. Yu. Relaxed clipping: A global training method for robust regression and classification. In *Advances in Neural Information Processing Systems*, 2010.

Y.-l. Yu, Ö. Aslan, and D. Schuurmans. A polynomial-time form of robust regression. In *Advances in Neural Information Processing Systems*, 2012.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Z. Zhang and M. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems*, 2018.