Robustness in Semi-Supervised Learning

AmirHossein Bagheri
Mahyar Jafari Nodeh
Radmehr Karimian

April 25, 2023

Abstract

This report delves into various techniques for classifier learning in the presence of label noise in training data. The aim is to establish sufficient conditions for a loss function that would make risk minimization tolerant to label noise for multi-class classification problems [GKS17]. Another approach discussed is a semi-supervised learning method for deep neural networks, where both labeled and unlabeled data are utilized with the help of pseudo-labels, similar to minimum entropy regularization [GB04] [Lee13]. This method can leverage unlabeled data to outperform mixture models and be superior to manifold learning.

The report focuses on Tilted Empirical Risk Minimization (TERM) [Li+21], a machine learning technique that uses exponential tilting to adjust the impact of individual losses. TERM has several desirable properties such as increasing or decreasing the influence of outliers for fairness or robustness, reducing variance for better generalization, and being a smooth approximation to the tail probability of losses. The report highlights connections between TERM and other objectives such as Value-at-Risk, Conditional Value-at-Risk, and distributional robust optimization (DRO). The report also provides efficient optimization methods for solving TERM and shows that it outperforms traditional Empirical Risk Minimization (ERM) and delivers competitive performance compared to state-of-the-art, problem-specific approaches. Exponential tilting, empirical risk minimization, superquantile optimization, fairness, robustness.

1 Pseudo labeling and Entropy regularization

1.1 Explanation

The entropy regularization minimization approach suggests the use of a loss function on labeled data, referred to as the supervised loss, and an unsupervised loss applied to unlabeled data to increase the model's confidence in predictions on unlabeled data. This approach is a common assumption in semi-supervised learning methods that place decision boundaries in low-density regions, as referred to as the cluster assumption [CZ05].

The semi-supervised learning problem is incorporated into standard supervised learning by using the maximum (conditional) likelihood estimation principle. The learning set, denoted $\mathcal{L}_n = \mathbf{x}i, \mathbf{z}i\ i = 1^n$, is comprised of $\mathbf{x}$ and $\mathbf{z}$, with $\mathbf{z} \in 0, 1^K$ representing the available labels (while y represents the complete class information). If $\mathbf{x}i$ is labeled ω_k, then $\mathbf{z}ik = 1$ and $\mathbf{z}i\ell = 0$ for $\ell \neq k$, and if $\mathbf{x}i$ is unlabeled, then $\mathbf{z}i\ell = 1$ for $\ell = 1, \ldots, K$.

Assuming that labeling is missing at random, meaning for all unlabeled examples, $P(\mathbf{z} \mid \mathbf{x}, \omega_k) = P(\mathbf{z} \mid \mathbf{x}, \omega_\ell)$ for any (ω_k, ω_ℓ) pair, we can calculate

$$P(\omega_k \mid \mathbf{x}, \mathbf{z}) = \frac{z_k P(\omega_k \mid \mathbf{x})}{\sum_{\ell=1}^{K} z_\ell P(\omega_\ell \mid \mathbf{x})}. \tag{1}$$

Given independent examples, the conditional log-likelihood of $(Z \mid X)$ on the observed sample is

$$L(\boldsymbol{\theta}; \mathcal{L}_n) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} z_{ik} f_k(\mathbf{x}_i; \boldsymbol{\theta}) \right) + h(\mathbf{z}_i) \tag{2}$$

where h(z), which does not depend on P (X, Y), is only affected by the messiness mechanism, and $f_k(x;\theta)$ is the concave model of $P(k|x)$ parameterized by θ.
now regularization implies that Empirical Entropy is defined as

$$H_{\text{emp}}\left(Y \mid X, Z; \mathcal{L}_n\right) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} P\left(\omega_k \mid \mathbf{x}_i, \mathbf{z}_i\right) \log P\left(\omega_k \mid \mathbf{x}_i, \mathbf{z}_i\right) \tag{3}$$

Entropy Regularization Recalling that $f_k(\mathbf{x};\boldsymbol{\theta})$ denotes the model of $P\left(\omega_k \mid \mathbf{x}\right)$, the model of $P\left(\omega_k \mid \mathbf{x}, \mathbf{z}\right)$is defined as follows:

$$g_k(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = \frac{z_k f_k(\mathbf{x};\boldsymbol{\theta})}{\sum_{\ell=1}^{K} z_\ell f_\ell(\mathbf{x};\boldsymbol{\theta})}. \tag{4}$$

For labeled data, $g_k(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = z_k$, and for unlabeled data, $g_k(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = f_k(\mathbf{x};\boldsymbol{\theta})$.
MAP estimate is the maximizer of the posterior distribution, that is, the maximizer of

$$\begin{aligned} C\left(\boldsymbol{\theta}, \lambda; \mathcal{L}_n\right) &= L\left(\boldsymbol{\theta}; \mathcal{L}_n\right) - \lambda H_{\text{emp}}\left(Y \mid X, Z; \mathcal{L}_n\right) \\ &= \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} z_{ik} f_k\left(\mathbf{x}_i\right)\right) + \lambda \sum_{i=1}^{n}\sum_{k=1}^{K} g_k\left(\mathbf{x}_i, \mathbf{z}_i\right) \log g_k\left(\mathbf{x}_i, \mathbf{z}_i\right) \end{aligned} \tag{5}$$

where the constant terms in the log-likelihood and log-prior have been dropped. While $L\left(\boldsymbol{\theta}; \mathcal{L}n\right)$ is only sensitive to labeled data, $H\text{emp}\left(Y \mid X, Z; \mathcal{L}_n\right)$ is only affected by the value of $f_k(\mathbf{x})$ on unlabeled data.
Setting λ to a proper value as a hyperparameter leads the model to become more confident in the perturbations in the input.
Another work [Lee13] claimed that their simple approach of assigning labels with higher confidence, as determined by the model, to the unlabeled data and then training the model on those labels, is equivalent to entropy minimization methods.

$$y_i' = \begin{cases} 1 & i = \text{argmax}_{i'}\, f_{i'}(x) \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Uinge Pseudo-Label in a fine-tuning phase with Dropout. The pre-trained network is trained in a supervised fashion with labeled and unlabeled data simultaneously. For unlabeled data, Pseudo-Labels recalculated every weights update are used for the same loss function of supervised learning task.
Because the total number of labeled data and unlabeled data is quite different and the training balance between them is quite important for the network performance, the overall loss function is

$$L = \frac{1}{n}\sum_{m=1}^{n}\sum_{i=1}^{C} L\left(y_i^m, f_i^m\right) + \alpha(t)\frac{1}{n'}\sum_{m=1}^{n'}\sum_{i=1}^{C} L\left(y_i'^m, f_i'^m\right), \tag{7}$$

where n is the number of mini-batch in labeled data for SGD, n' for unlabeled data, f_i^m is the output units of m's sample in labeled data, y_i^m is the label of that, $f_i'^m$ for unlabeled data, $y_i'^m$ is the pseudo-label of that for unlabeled data, $\alpha(t)$ is a coefficient balancing them.

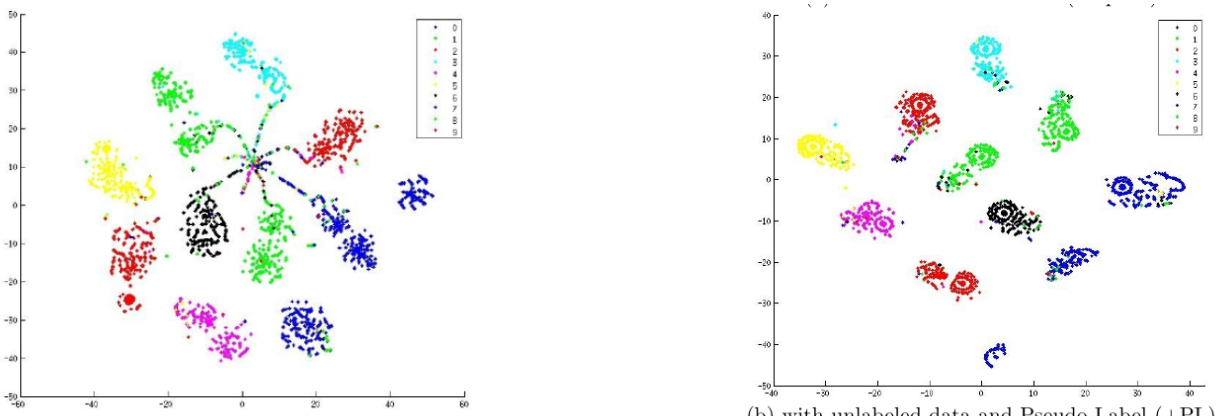
1.2 Experiments

Pseudo labeling have remarkable results on MNIST dataset their method involves two initialzation phase first they use auto encoder and for second one they use demonising auto encoder.

METHOD	100	600	1000	3000
NN	25.81	11.44	10.7	6.04
SVM	23.44	8.85	7.77	4.21
CNN	22.98	7.68	6.45	3.35
TSVM	16.81	6.16	5.38	3.45
DBN-RNCA	–	8.7	–	3.3
EMBEDNN	16.86	5.97	5.73	3.59
CAE	13.47	6.3	4.77	3.22
MTC	12.03	5.13	3.64	**2.57**
DROPNN	21.89	8.57	6.59	3.72
+PL	16.15	5.03	4.30	2.80
+PL+DAE	**10.49**	**4.01**	**3.46**	2.69

Table 1: Comparing simple pseudo-labeling method with previous works

We also can see the effect of training on unlabeled data with pseudo label method. when we use unsupervised data the boundaries lay on less dense regions.



(a) without unlabeled data (dropNN)

(b) with unlabeled data and Pseudo-Label (+PL)

Figure 1: T-SNE embedding of the network output of MNIST test data

Entropy Minimization on the other hand have drown sample from two Gaussian distribution. In the first series of experiments, the researchers examine two-class problems in a 50-dimensional input space. The classes are generated with equal probability from normal distributions. Class ω_1 has a mean of $(aa\dots a)$ and a unit covariance matrix, while class ω_2 has a mean of $-(aa\dots a)$ and a unit covariance matrix. The parameter a adjusts the Bayes error, which ranges from 1 to 20 $(1, 2.5, 5, 10, 20)$. The learning sets consist of n_l labeled examples, $(n_l = 50, 100, 200)$, and n_u unlabeled examples, $(n_u = n_l \times (1, 3, 10, 30, 100))$. In total, 75 different setups are evaluated, and for each setup, 10 different training samples are generated. Generalization performance is estimated on a test set of size 10,000.
This benchmark provides a way to compare the algorithms in a scenario where unlabeled data is known to convey information. In addition to the EM algorithm's favorable initialization to the optimal parameters, it also benefits from the accuracy of the model; the data was generated based on the model of two Gaussian subpopulations with identical covariances. The logistic regression model, on the other hand, is only capable of accommodating the joint distribution, which is a weaker requirement than accuracy.

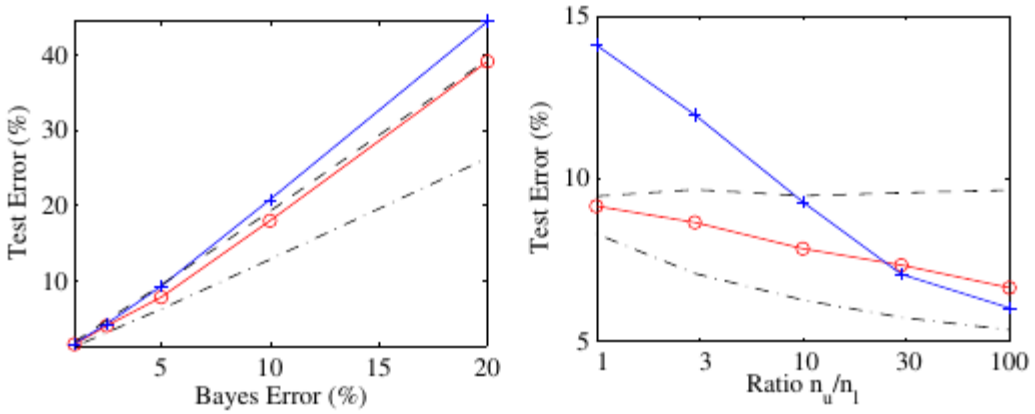


Figure 2: Left: test error vs Bayes error rate for $n_u/n_l = 10$; right: test error $vs. n_u/n_l$ ratio for 5% Bayes error ($a = 0.23$). Test errors of minimum entropy logistic regression (o) and mixture models (+). The errors of logistic regression (dashed), and logistic regression with all labels known (dash-dotted) are shown for reference

In the second series of experiments, the setup was modified to include outliers in the class-conditional densities. The examples were generated from a mixture of two Gaussians centered on the same mean, with 98% of examples generated from a unit variance component and the remaining 2% generated from a large variance component with a standard deviation of 10. The EM algorithm used a simple Gaussian mixture model, which was slightly misspecified. The results showed that the generative model performed worse than logistic regression for all sample sizes due to the misspecification. The unlabeled examples initially had a beneficial effect on the test error but eventually had a detrimental effect when they outnumbered the labeled examples. However, the diagnosis models performed smoothly and the minimum entropy criterion performance improved.

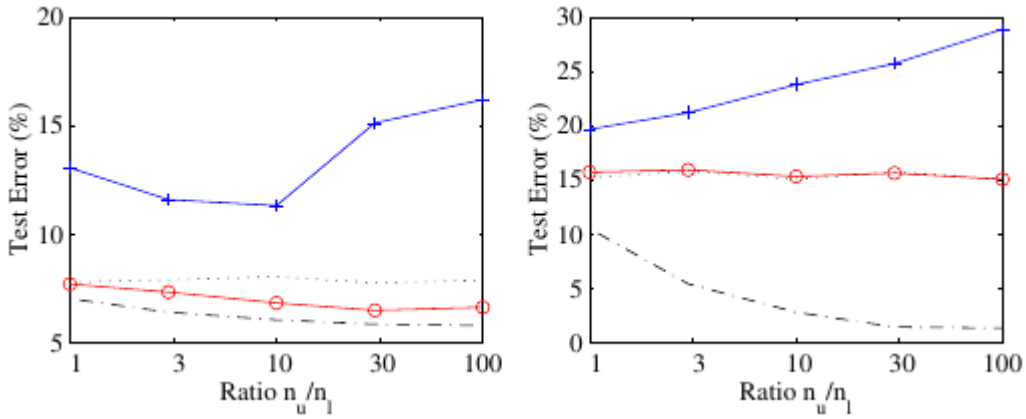


Figure 3: Left: Test error vs. n_u/n_l ratio for a = 0.23. Average test errors for minimum entropy logistic regression o and mixture models (+). The test error rates of logistic regression (dotted), and logistic regression with all labels known (dash-dotted) are shown for reference

1.3 applied applications

Some applied use of models and frameworks above worth mentioning are works which have been done in segmentation and domain generalizations works [Luo+20] [Yu+19] [OHT20] [Ver+22].

2 Robust Loss Functions under Label Noise for Deep Neural Networks

2.1 explanation

This work Consider unlabeled data as noise in data set and try to provide robust function for noisy label. A loss function is a map $L : \mathcal{C} \times \mathcal{Y} \to \mathbb{R}^+$. Given any loss function, L, and a classifier, f, define the L-risk of f by

$$R_L(f) = \mathbb{E}_{\mathcal{D}}\left[L\left(f(\mathbf{x}), y_{\mathbf{x}}\right)\right] = \mathbb{E}_{\mathbf{x},y_{\mathbf{x}}}\left[L\left(f(\mathbf{x}), y_{\mathbf{x}}\right)\right] \tag{8}$$

In the presence of label noise, the clean training data, represented by S, is not accessible to the learner. Instead, the learner has access to the noisy training data $S_\eta = \left(\mathbf{x}n, \hat{y}\mathbf{x}n\right), n = 1, \cdots, N$, where $\hat{y}\mathbf{x}_n$ is the noisy label for instance $\mathbf{x}_n$. This noisy label is given by:

$$\hat{y}_{\mathbf{x}_n} = \begin{cases} y_{\mathbf{x}_n} & \text{with probability } \left(1 - \eta_{\mathbf{x}_n}\right) \\ j, \quad j \in [k], j \neq y_{\mathbf{x}_n} & \text{with probability } \bar{\eta}_{\mathbf{x}_n j} \end{cases} \tag{9}$$

where $\eta_{\mathbf{x}n}$ is the noise rate for instance $\mathbf{x}n$ and $\bar{\eta}\mathbf{x}nj$ is the probability that instance $\mathbf{x}n$ is corrupted to label j, given that its true label is $y\mathbf{x}n$. It is important to note that for all $\mathbf{x}$, conditioned on $y\mathbf{x} = i$, we have $\sum_{j\neq i} \bar{\eta}\mathbf{x}_j = \eta_{\mathbf{x}}$.
The true label (under distribution $\mathcal{D}$) of an instance $\mathbf{x}$ is denoted by the random variable $y_{\mathbf{x}}$, while its corrupted label is denoted by $\hat{y}\mathbf{x}$. The joint probability distribution of $\mathbf{x}$ and $\hat{y}\mathbf{x}$ is denoted by $\mathcal{D}_\eta$.
Noise can be symmetric or uniform, if $\eta_{\mathbf{x}} = \eta$ and $\bar{\eta}\mathbf{x}j = \frac{\eta}{k-1}$ for all $\mathbf{x}$ and $j \neq y\mathbf{x}$. In this case, η is a constant.
Noise can also be class-conditional or asymmetric, if the dependence of $\eta_{\mathbf{x}}$ and $\bar{\eta}\mathbf{x}j$ is only through $y\mathbf{x}$. In this case, we write $\eta_{\mathbf{x}} = \eta_{y_{\mathbf{x}}}$ and $\bar{\eta}\mathbf{x}j = \bar{\eta}y_{\mathbf{x}}j$. For example, $\bar{\eta}_{ij}$ is the probability that a pattern of class i is corrupted to label j when the label is noisy.
The L-risk, Let f^* be the global minimizer (over the chosen function class) of $R_L(f)$. When there is label noise, the data is drawn according to distribution $\mathcal{D}_\eta$. Then L-risk of a classifier f under noisy data is

$$R_L^\eta(f) = \mathbb{E}_{\mathcal{D}_\eta}\left[L\left(f(\mathbf{x}), \hat{y}_{\mathbf{x}}\right)\right] = \mathbb{E}_{\mathbf{x},\hat{y}_{\mathbf{x}}}\left[L\left(f(\mathbf{x}), \hat{y}_{\mathbf{x}}\right)\right] \tag{10}$$

They suggest three theorem which we will not go throw proof for them you can find proof in [GKS17].
Loss function is symmetric if satisfy condition below and remember for all three theorem this is required as loss function feature.

$$\sum_{i=1}^{i=k} L(f(x), i) = C, \forall x \in X \tag{11}$$

- Theorem 1: In a multi-class classification problem, let loss function L. Then L is noise tolerant under symmetric or uniform label noise if $\eta < \frac{k-1}{k}$
- Theorem 2: Suppose loss L if $R_L\left(f^*\right) = 0$, then L is also noise tolerant under simple non uniform noise when $\eta_{\mathbf{x}} < \frac{k-1}{k}, \forall \mathbf{x}$. If $R_L\left(f^*\right) = \rho > 0$ then, under simple non-uniform noise, $R_L\left(f_\eta^*\right)$ is upper bounded by $\rho / \left(1 - \frac{k\eta_{\max}}{k-1}\right)$, where $\eta_{\max}$ is maximum noise rate over $\mathbf{x} \in \mathcal{X}$. (Recall that f^* is minimizer of R_L and f_η^* is minimizer of R_L^η).
- Theorem 3: Suppose L and $0 \leq L(f(\mathbf{x}), i) \leq C/(k-1), \forall i \in [k]$. If $R_L\left(f^*\right) = 0$, then, L is noise tolerant under class conditional noise when $\bar{\eta}_{ij} < \left(1 - \eta_i\right), \forall j \neq i, \forall i, j \in [k]$.

Now they investigate different loss functions and conclusively the MAE is the symmetric loss

$$|e_i - u| = 2 - 2u_i \tag{12}$$

$$\sum L(f(x), i) = 2k - 2 \tag{13}$$

2.2 Empirical results

The authors performed experiments on various image and text data sets and reported the results figure 4 and table 2. The table shows the size of the training and test sets, the number of classes, and the input dimension. Different network architectures were used for each data set based on the nature of the data (image or text) and feature space dimensions. All networks used ReLU in the hidden layers and a softmax layer at the output with the size of the layer being the number of classes. The networks were trained through backpropagation with momentum term, weight decay, and dropout regularization. The results are the average of six runs and label noise was added to the training set by changing the labels randomly. The noise was either symmetric or class-conditional with a diagonal dominant noise probability matrix.

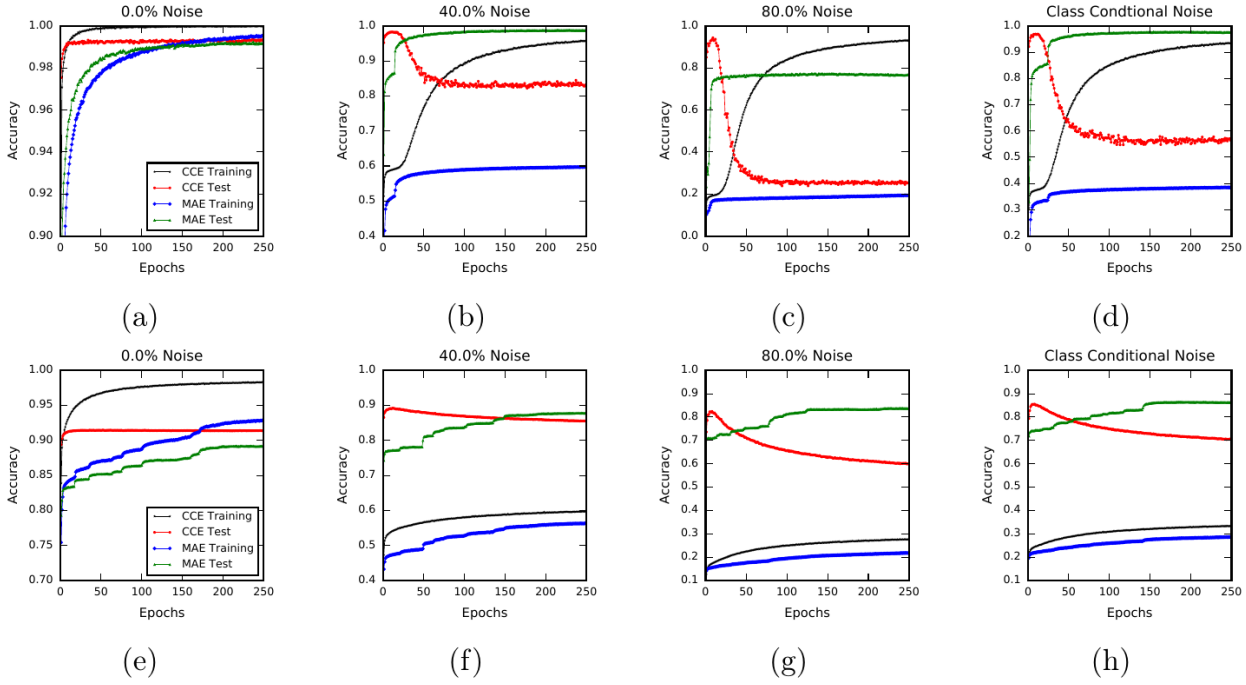


Figure 4: Train-Test Accuracies for log loss and MAE over epochs, for MNIST Datasets under (a) 0% noise (b) 40% noise (c) 80% noise (d) CC noise; and RCV1 Datasets under (e) 0% noise (f) 40% noise (g) 80% noise (h) CC noise. Legends shown in (a) and (e).

3 On Tilted Losses in Machine Learning: Theory and Applications

3.1 Brief introduction

This work introduces the use of exponential tilting in the field of machine learning. The authors propose Tilted Empirical Risk Minimization (TERM), an extension to traditional Empirical Risk Minimization (ERM), which adjusts the impact of individual losses through tilting. TERM offers several advantages, such as the ability to handle outliers, improve generalization, and be viewed as a smooth approximation of the

Data	loss	$\eta = 0\%$	$\eta = 30\%$	$\eta = 60\%$	CC
MNIST	CCE	0.9936	0.9138	0.5888	0.5775($\pm$0.0291)
	MAE	0.9916	0.9886	0.9799	0.9713
	MSE	0.9921	0.9868	0.9766	0.8505($\pm$0.0473)
RCV1	CCE	0.9126	0.8738	0.7905	0.7418($\pm$0.025)
	MAE	0.8732($\pm$0.0107)	0.8688	0.8637($\pm$0.0201)	0.8587
	MSE	0.9014	0.8943	0.8682($\pm$0.0120)	0.8315
Cifar	CCE	0.7812	0.5598($\pm$0.0170)	0.3083	0.4896
10	MAE	0.7810($\pm$0.0190)	0.7011($\pm$0.0264)	0.5328($\pm$0.0251)	0.61425($\pm$0.0320)
	MSE	0.8074	0.7027	0.5257($\pm$0.0146)	0.6249($\pm$0.0359)
Imdb	CCE	0.8808	0.7729	0.6466	0.7858($\pm$0.0135)
	MAE	0.8813	0.8500	0.7352($\pm$0.0145)	0.8382($\pm$0.0127)
	MSE	0.8816	0.7725($\pm$0.0105)	0.6506($\pm$0.0103)	0.7874
News	CCE	0.7842	0.6905	0.4670	0.4973($\pm$0.0148)
wire	MAE	0.8081	0.7553	0.6357($\pm$0.0106)	0.6535
	MSE	0.7916	0.6626	0.4078($\pm$0.0172)	0.4377($\pm$0.0140)
News	CCE	0.8006	0.7571	0.6435	0.5629
group	MAE	0.7890	0.7749	0.7319	0.6772
	MSE	0.7999	0.7553	0.6347	0.5519

Table 2: Accuracies under different noise rates (η) for all datasets (for Imdb, η 's are halved). The last column gives accuracies under class conditional noise. In all the cases, standard deviation is shown only when it is more than 0.01

tail probability of losses. The authors provide optimization methods for solving TERM, demonstrate its efficient solution relative to alternatives, and show its applications in various areas of machine learning, such as fairness, outlier mitigation, and class imbalance. The results indicate that TERM outperforms ERM and is competitive with state-of-the-art problem-specific approaches.
This method suggest new loss function with one more hyper-parameter by definition it's defined as

$$\overline{R(\theta)} = \frac{1}{N}\sum f(x_i;\theta) \tag{14}$$

$$R(\tilde{t};\theta) = \frac{1}{t} = \log\frac{1}{N}\sum e^{tf(x_i;\theta)} \tag{15}$$

As you can see we can derive below results from definition

$$\lim_{t\to 0} R(t;\theta) = \overline{R(\theta)} \tag{16}$$

$$\lim_{t\to\infty} R(t;\theta) = \text{max-loss} \tag{17}$$

$$\lim_{t\to-\infty} R(t;\theta) = \text{min-loss} \tag{18}$$

Exponential tilt of the information density which derive from the matter of perspective of how we see f as function is defined as we define f [DZ98] [CT06], Information of x under $f(x_i;\theta) = -\log p_\theta(x_i)$ so we have Expectation $\log E[e^{tf(x;\theta)}] = \log\sum p(x)p_\theta^{-t}(x_i)$. It is noteworthy that if P is an exponential family of distributions parameterized by θ TERM can be viewed as an appropriately scaled variant of the empirical cumulant generating function $\log\frac{1}{N}e^{td(x;\theta)}$.

3.1.1 motivation

we can see in figure 5 effect of TERM in three different scenarios like point estimator, linear regression and logistic regression.

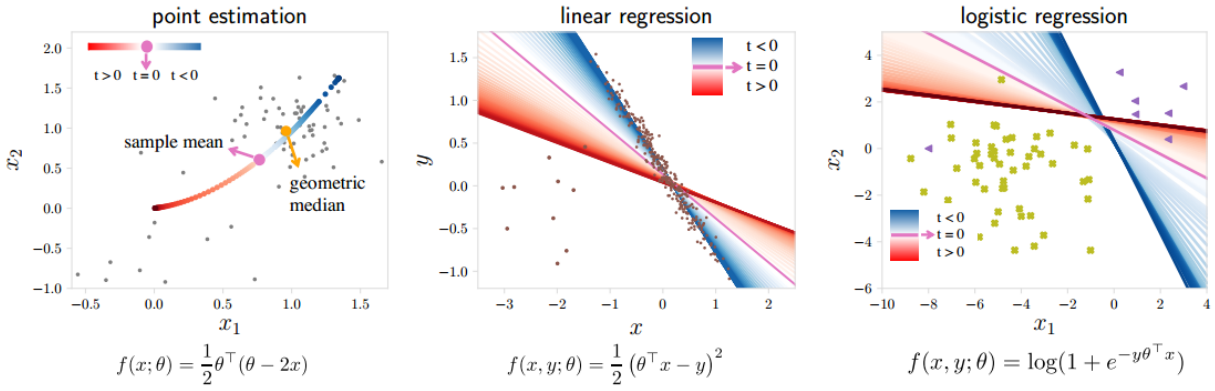


Figure 5: Toy examples illustrating TERM as a function of t: (a) finding a point estimate from a set of 2D samples, (b) linear regression with outliers, and (c) logistic regression with imbalanced classes. While positive values of t magnify outliers, negative values suppress them. Setting t=0 recovers the original ERM objective

3.2 TERM

3.2.1 Assumptions

- (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 $\theta \in \Theta \subseteq \mathbb{R}^d$
- (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 $\theta \in \Theta \subseteq \mathbb{R}^d$,

$$\beta_{min} I \leq \nabla^2_{\theta\theta^T} f(x_i;\theta) \leq \beta_{max} I \tag{19}$$

- (Generalized linear model condition [WJ08]). Assume that Assumption 2 is satisfied. Further, assume that the loss function $f(x;\theta)$ is given by

$$f(x;\theta) = A(\theta) - \theta^T T(x) \tag{20}$$

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

$$\beta_{min} I \leq \nabla^2_{\theta\theta^T} A(\theta) \leq \beta_{max} I, \tag{21}$$

and

$$\sum_{i\in[N]} T(x_i)\, T(x_i)^T > 0. \tag{22}$$

- (Strict saddle property [Ge+15]): 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, $\nabla^2_{\theta\theta^T} \widetilde{R}(t;\theta) > 0$, and for all other stationary solutions, $\lambda_{\min}\left(\nabla^2_{\theta\theta^T} \widetilde{R}(t;\theta)\right) < 0$, where $\lambda_{\min}(\cdot)$ is the minimum eigenvalue of the matrix.

3.2.2 Interpretation

As discussed via the toy examples, 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.

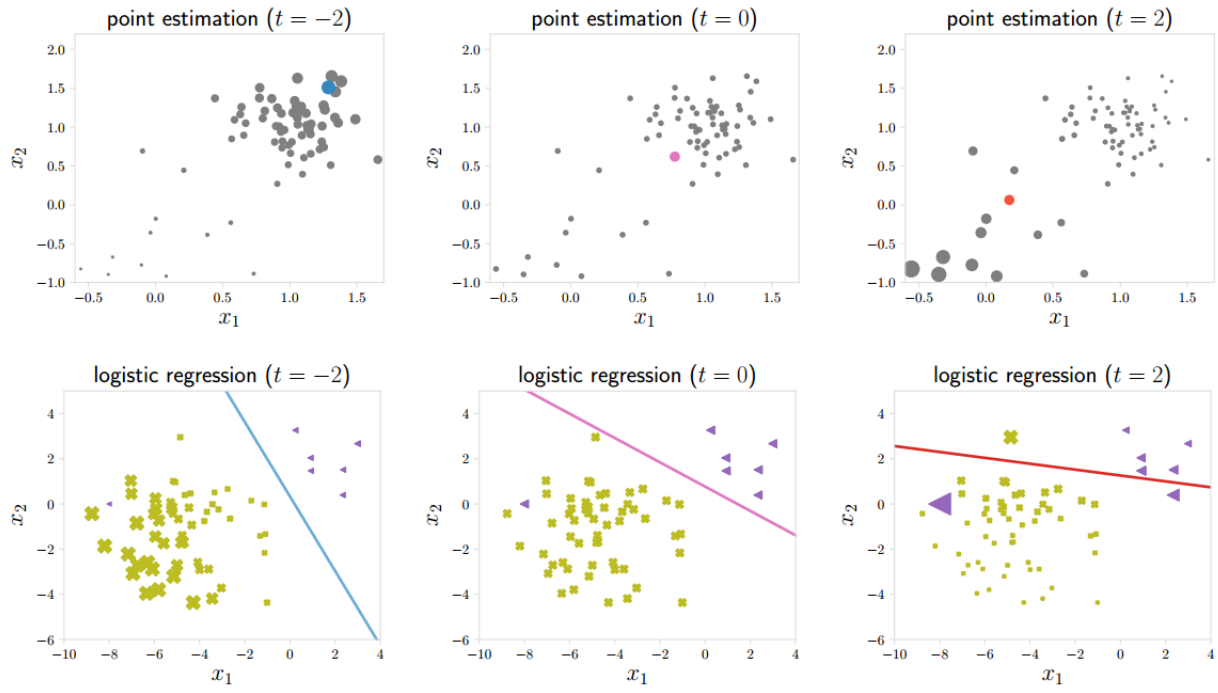


Figure 6: visualize the size of the samples using their gradient weights. Negative t's (t = -2 on the left) focus on the inliner samples (suppressing outliers), while positive t's (t = 2 on the right) magnify the outlier samples.

3.2.3 Tilted gradient & Bias/Variance trade-off

Lemma:
For a smooth loss function $f(x;\theta)$

$$\nabla_\theta \widetilde{R}(t;\theta) = \sum_{i\in[N]} w_i(t;\theta)\nabla_\theta f(x_i;\theta) \tag{23}$$

where tilted weight are given by (calculating radiant reach us here)

$$w_i(t,\theta) := \frac{e^{tf(x_i;\theta}}{\sum_{j\in[N]} e^{tf(x_j;\theta}} = \frac{1}{N} e^{t(f(x_i;\theta)-\widetilde{R}(t;\theta))} \tag{24}$$

The gradient of the tilted objective has been previously studied in the context of exponential smoothing (as referenced in [PR11]). The tilted gradient is a weighted average of the gradients of the original individual losses, with each data point weighted exponentially according to the value of its loss. The parameter t controls the weighting of the losses, with t = 0 corresponding to uniform weighting as in ERM.
This property allows for a trade-off between optimizing the average loss and reducing variance, potentially leading to a better bias-variance trade-off for generalization (as referenced in [MP09] and [Ben62]). This property is consistent with and extends the approximation of TERM by Liu and Theodorou (as referenced in [LT19]), which approximates TERM as empirical risk regularized with the variance of the loss at t = 0.
In addition to empirical variance across all losses, other measures of distribution uniformity exist.

3.3 Connections to Other Risk Measures

The paper suggest three relation which we will discuss here briefly, we will not go through mathematical proof and mention results only

- TERM can be expressed as:
$$\widetilde{R}(t;\theta) = H_{1-t}(\mathbf{u}||\mathbf{w}(1;\theta)), \tag{25}$$
where $\mathbf{u}$ denotes the uniform N-vector and $\mathbf{w}(1;\theta) := (w_1(1;\theta), \ldots, w_n(1;\theta))$ with $w_i(1;\theta)$ and for any two N-vectors $\mathbf{p}$ and $\mathbf{q}$,
$$H_\rho(\mathbf{p}||\mathbf{q}) := \frac{1}{1-\rho} \log\left(\sum_{i\in N} p_i q_i^{\rho-1}\right). \tag{26}$$
In simpler terms, if we view the loss $f(x_i;\theta)$ as the log-likelihood of sample x_i under p_θ, then TERM is the Rényi entropy of order $(1-t)$ between the uniform vector and the normalized likelihood vector of all samples, $\mathbf{w}(1;\theta)$. Therefore, minimizing over θ promotes the uniformity of $\mathbf{w}(1;\theta)$ as measured by the Rényi cross entropy with the uniform vector.

- TERM can be seen as a form of regularization in traditional ERM. This can be approximately demonstrated by a Taylor series expansion at t = 0, which decomposes TERM into empirical risk regularized by t times the empirical variance of the loss for small t (as referenced in [LT19]). In this paper, we present an exact interpretation of TERM as a regularized ERM for all values of t. To do this, we first consider the distributional case and relate $R_X(t;\theta)$ to cross entropy. The entropic risk of order t can be stated as:
$$R_X(t;\theta) = H(p||p_\theta) + \frac{1}{t} D(p||T(p, p_\theta, -t)), \tag{27}$$
where D denotes KL divergence between two distributions and $T(p, p_\theta, -t)$ is a mismatched tilted distribution defined as:
$$T(p, p_\theta, -t)(x) := \frac{p(x)p_\theta(x)^{-t}}{\sum_u p(u)p_\theta(u)^{-t}} \tag{28}$$

- Additionally, TERM has close ties to distributionally robust optimization (DRO) objectives. Specifically, TERM with a large value of t (i.e., $t >> 0$) is equivalent to a form of DRO with a max-entropy regularizer [ND17] [DN19] [CP20].

  The max-entropy regularizer promotes uniformity in the distribution of the samples, similar to TERM with a large t value. By encouraging uniformity in the distribution, the max-entropy regularizer helps to make the optimization problem more robust to distributional uncertainty.

  Therefore, TERM with a large t can be seen as a special case of DRO with a max-entropy regularizer, providing a way to incorporate distributional robustness into traditional ERM. This connection between TERM and DRO highlights the importance of considering the distribution of the data when training machine learning models and the potential benefits of incorporating distributional robustness into the training process.

4 Solve TERM & Hierarchical TERM

Here paper introduce some algorithm for solving the problem we first introduce algorithm and then explain some theorem and their proof.
Developing optimization methods for solving TERM for both batch and stochastic cases and analyze the effects of the parameter "t" on the convergence of these methods. They found that when t is positive, the t-tilted loss is strongly convex and has a unique solution. However, for negative t, the loss may become non-convex and have multiple local minima. To address this issue, they solve TERM by smoothly decreasing t from 0 to ensure that the solutions form a continuous set in high-dimensional space. Despite the non-convexity of TERM for negative t, the authors found that this approach produces effective solutions for real-world problems.

4.1 First-Order Batch Methods

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

Figure 7: Batch Algorithm

Theorem 1: proof in appendix 6
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 $\alpha = \frac{1}{C1+C2t}$ after k iteration:

$$\tilde{R}(t;\theta_k) - \tilde{R}(t;\theta(\check{t})) \leqslant (1 - \frac{\beta}{C_1 + C_2 t})^k (\tilde{R}(t;\theta_0) - \tilde{R}(t;\theta(\check{t}))) \tag{29}$$

Theorem 2: proof in appendix 6
Convergence of Algorithm 1 for smooth problems satisfying PL conditions
Assume $f(x,\theta)$ is β_{max} smooth and Possibly non-convex. Further assume $\sum_{i\in[N]} p_i f(x,\theta)$ is $\mu/2$ -PL for any $P \in \Delta_N$ where P is $P := (p_1, ..., p_n)$. There exists $\beta_{max} \leqslant C_1 < \infty and C_2 < \infty$ that do not depend on t such that for any $t \in \mathbb{R}^{>0}$ with setting step $\alpha = \frac{1}{C1+C2t}$ after k iteration:

$$\tilde{R}(t;\theta_k) - \tilde{R}(t;\theta(\check{t})) \leqslant (1 - \frac{\mu}{C_1 + C_2 t})^k (\tilde{R}(t;\theta_0) - \tilde{R}(t;\theta(\check{t}))) \tag{30}$$

Theorem 2 applies to both convex and non-convex smooth functions satisfying PL conditions.

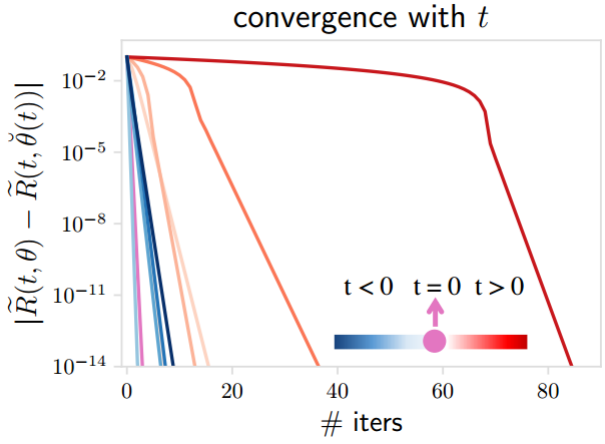


Figure 8: Effect of t on convergence

4.2 First-Order Stochastic Methods

In order to obtain unbiased stochastic gradients for optimizing the TERM objective, a term called $\widetilde{\widetilde{R}}$ that incorporates stochastic dynamics is used to estimate the tilted objective $\widetilde{R}$. The TERM objective can be optimized using previous stochastic compositional optimization techniques and is optimized by maintaining two sequences: the model parameters and the objective estimate $\widetilde{\widetilde{R}}$. The optimization process is done by sampling two independent mini-batches to obtain the gradient of the original loss functions and to update $\widetilde{\widetilde{R}}$.

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

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 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

$\widetilde{\widetilde{R}}_t \leftarrow \frac{1}{t}\log\left((1-\lambda)e^{t\widetilde{\widetilde{R}}_t} + \lambda e^{t\widetilde{R}_{B,t}}\right)$

$w_{t,x} \leftarrow e^{tf(x;\theta) - t\widetilde{\widetilde{R}}_t}$

$\theta \leftarrow \theta - \frac{\alpha}{|B_1|}\sum_{x\in B_1} w_{t,x}\nabla_\theta f(x;\theta)$

end

Figure 9: Stochastic Non-Hierarchical Algorithm

Theorem 3: proof in appendix 6
Convergence of Algorithm 5 for strongly-convex problems
Assume $f : \chi \times \Theta \to [\tilde{F_{min}}, \tilde{F_{max}}]$ is L-Lipschitz in θ.Assume $\tilde{R}(t,\theta)$ is μ strongly convex with uniformly bounded stochastic gradient. Denote $k_t := argmax_k(k < \frac{2e}{\mu} + \frac{etLB\exp t(\tilde{F_{max}} - \tilde{F_{min}})}{\mu k})$. Assume batch size is 1 and for $k \geqslant k_t$:

$$E[\| \theta_{k+1} - \theta^* \|^2] \leqslant \frac{V_t}{k+1} \tag{31}$$

where:

$$\theta^* := \check{\theta}(t) \; and \; V_t = max\{k_t E[\| \theta_k - \theta^* \|^2], \frac{4B^2 e^{2+2t(\tilde{F_{max}} - \tilde{F_{min}})}}{\mu^2}\} \tag{32}$$

and

$$E[\| \theta_k - \theta^* \|^2] \leqslant max\{E[\| \theta_1 - \theta^* \|^2], \frac{B^2 e^{2t(\tilde{F_{max}} - \tilde{F_{min}})+1}}{\mu(1 + tLBe^{t(\tilde{F_{max}} - \tilde{F_{min}})})}\} \tag{33}$$

Theorem 4: proof in appendix 6
Convergence of Algorithm 5 for non-convex smooth problems
Assume $f : \chi \times \Theta \to [\tilde{F_{min}}, \tilde{F_{max}}]$ is L-Lipschitz in θ.Assume $\tilde{R}(t,\theta)$ is β-smooth with uniformly bounded stochastic gradient. Denote $k_t := \lceil \frac{2t^2L^2(\tilde{F_{max}} - \tilde{F_{min}})}{\beta e^2} \rceil$. Assume batch size is 1 and for $k \geqslant k_t$:

$$\frac{1}{K}\sum_{k=k_t}^{K} E[\| \nabla \tilde{R}(t;\theta_k) \|^2] \leqslant \sqrt{8}Be^{t(\tilde{F_{max}} - \tilde{F_{min}})+1}\sqrt{\frac{\beta(\tilde{F_{max}} - \tilde{F_{min}})}{K}} \tag{34}$$

We have convergence for non-convex smooth problems with PL conditions,too.

4.3 Hierarchical TERM

We evaluate the gradient of the hierarchical multi-objective tilt objective in order to optimize the TERM algorithm for multiple objectives. The gradient is calculated as follows:

$$\frac{\partial \widetilde{J}(t,\tau;\theta)}{\partial \theta} = \frac{\sum_{g\in[G]} |g| e^{t\widetilde{R}_g(\tau;\theta)} \frac{\partial \widetilde{R}g(\tau;\theta)}{\partial \theta}}{\sum g \in [G]|g| e^{t\widetilde{R}_g(\tau;\theta)}} \tag{35}$$

where

$$\frac{\partial \widetilde{R}g(\tau;\theta)}{\partial \theta} = \frac{\sum_{x\in g} e^{\tau f(x;\theta)} \frac{\partial f(x;\theta)}{\partial \theta}}{\sum_{x\in g} e^{\tau f(x;\theta)}} \tag{36}$$

We arrive at

$$\widetilde{J}(t,\tau;\theta) := \frac{1}{t}\log\left(\frac{1}{N}\sum_{g\in[G]} |g| e^{t\tilde{R}_g(\tau;\theta)}\right), \text{ where } \widetilde{R}_g(\tau;\theta) := \frac{1}{\tau}\log\left(\frac{1}{|g|}\sum_{x\in g} e^{\tau f(x;\theta)}\right) \tag{37}$$

$$\nabla_\theta \widetilde{J}(t,\tau;\theta) = \sum_{g\in[G]}\sum_{x\in g} w_{g,x}(t,\tau;\theta)\nabla_\theta f(x;\theta) \tag{38}$$

where

$$w_{g,x}(t,\tau;\theta) := \frac{\left(\frac{1}{|g|}\sum_{y\in g} e^{\tau f(y;\theta)}\right)^{\left(\frac{t}{\tau}-1\right)}}{\sum_{g'\in[G]} |g'| \left(\frac{1}{|g'|}\sum_{y\in g'} e^{\tau f(y;\theta)}\right)^{\frac{t}{\tau}}} e^{\tau f(x;\theta)}. \tag{39}$$

In summary, the TERM (Tilted Empirical Risk Minimization) method is used to solve hierarchical optimization problems in both batch and stochastic settings. In the batch setting, gradient-based methods are used with tilted gradients defined for the hierarchical objective. In the stochastic setting, Algorithm 4 extends Algorithm 2 to the multi-objective setting by addressing group-level tilting by choosing a group based on the tilted weight vector and sample-level tilting by re-weighting the samples in a mini-batch. The tilted objective for each group is estimated via a tilted average of the current estimate and the history. The method allows for the sampling of a group from which the mini batch is drawn. Group-level tilting can be recovered by setting the inner-level tilt parameter to 0. TERM has been applied to a variety of machine learning problems. The TERM method is a powerful tool for solving hierarchical optimization problems in machine learning. The basic idea behind TERM is to use tilted gradients to address both group-level and sample-level tilting, which allows for a more effective optimization process.

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

  $\widetilde{J}_{t,\tau} \leftarrow \frac{1}{t} \log \left(\frac{1}{N} \sum_{g \in [G]} |g| e^{t \widetilde{R}_g(\tau;\theta)} \right)$

  $w_{t,\tau,g} \leftarrow |g| e^{t \widetilde{R}_{\tau,g} - t \widetilde{J}_{t,\tau}}$

  $\theta \leftarrow \theta - \frac{\alpha}{N} \sum_{g \in [G]} w_{t,\tau,g} \nabla_\theta \widetilde{R}_{g,\tau}$

end

Figure 10: Batch Hierarchical Algorithm

In the batch setting, gradient-based methods are used with tilted gradients defined for the hierarchical objective. This means that the optimization process takes into account the hierarchical structure of the problem, which results in a more accurate solution.

In the stochastic setting, Algorithm 4 extends Algorithm 2 to incorporate multi-objective tilting. At each iteration, the group-level tilting is addressed by choosing a group based on the tilted weight vector. The sample-level tilting is then incorporated by re-weighting the samples in a uniformly drawn mini-batch. The tilted objective for each group is estimated via a tilted average of the current estimate and the history.

In practice, it is possible to sample a group from which the mini-batch is drawn, which can lead to improved optimization performance. Additionally, for small numbers of groups, one might want to draw one mini-batch per each group and weight the resulting gradients accordingly. Group-level tilting can be recovered from Algorithms 3 and 4 by setting the inner-level tilt parameter to 0. This means that the optimization process will only consider the sample-level tilting, rather than taking into account the hierarchical structure of the problem.

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

  $\widetilde{\widetilde{R}}_{g,\tau} \leftarrow \frac{1}{\tau} \log \left((1-\lambda) e^{\tau \widetilde{\widetilde{R}}_{g,\tau}} + \lambda e^{\tau \widetilde{R}_{B,\tau}} \right)$

  $w_{\tau,x} \leftarrow e^{\tau f(x;\theta) - \tau \widetilde{\widetilde{R}}_{g,\tau}}$

  $\theta \leftarrow \theta - \frac{\alpha}{|B|} \sum_{x \in B} w_{\tau,x} \nabla_\theta f(x;\theta)$

end

Figure 11: Stochastic Hierarchical Algorithm

5 Use Cases

5.1 Mitigating Noisy Outlier ($t < 0$)

The authors are investigating the ability of TERM to find solutions that are robust to noisy outliers. They focus on the setting of random additive noise and acknowledge that its applicability to other forms of robustness would be a future area of exploration. They do not compare with approaches that require additional clean validation data as obtaining such data can be difficult in practice.

5.1.1 Robust Regression

The authors are evaluating TERM in a regression task with noisy targets, using the Drug Discovery dataset. They compare TERM to standard ERM with L2 loss (also known as mean squared error (MSE) [Oli+18] [Dia+19]), L1 loss, Huber loss [Hub64], consistent robust regression (CRR) [Bha+17], and STIR [Muk+20]. TERM is equivalent to exponential squared loss. The results show that TERM is competitive with the baselines at 20% noise level and performs better with moderate-to-extreme noise. The performance of the oracle method (Genie ERM) is also presented as the expected performance limit in the noisy setting. The results show similar trends in scenarios with noisy features and targets. CRR tends to be slow as it scales cubicly with the number of dimensions [Bha+17], while TERM is roughly as efficient as ERM.

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)

Figure 12: TERM is competitive with robust regression baselines, particularly in high noise regimes.

Label and feature noise:
he authors present results of TERM's performance with both label and feature noise on two datasets (cal-housing and abalone). The noisy samples are generated by corrupting 5% of 100 training samples by multiplying the features by 100 and the targets by 10,000. The results in Table 3 show that TERM significantly outperforms the baseline methods in the noisy regime on both datasets.

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)

Figure 13: An alternative noise setup involving both feature noise and label noise. Similarly, TERM with t = -2 significantly outperforms several baseline objectives for noisy outlier mitigation.

Unstructured random v.s. adversarial noise:
In the experiments, the focus has been on random noise, allowing the methods to find the structure of clean data even when the majority of the samples are noisy outliers. TERM performs well for linear regression

under different noise levels but could potentially overfit to outliers if they are constructed in an adversarial way, as seen in examples where TERM had a high error measured on clean data under 40% and 80% noise.

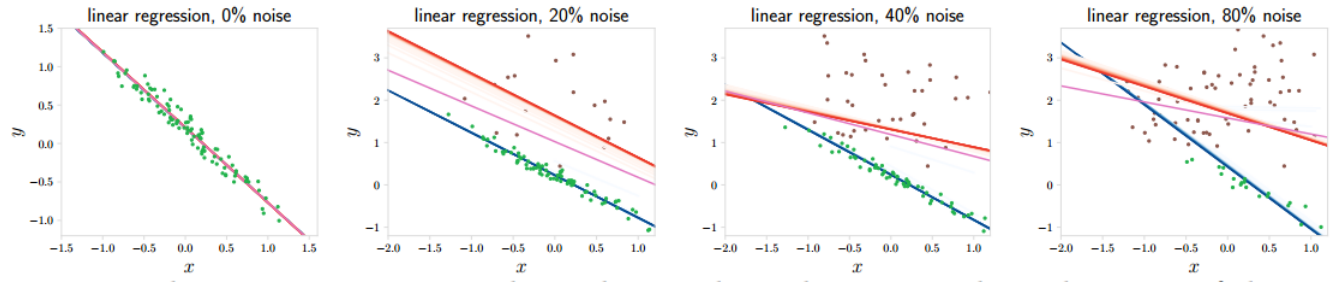


Figure 14: Robust regression on synthetic data with random noise where the mean of the noisy samples is different from that of clean ones. TERM with negative t's (blue, t = - 2) can fit structured clean data at all noise levels, while ERM (purple) and TERM with positive t's (red) overfit to corrupted data. We color inliers in green and outliers in brown for visualization.

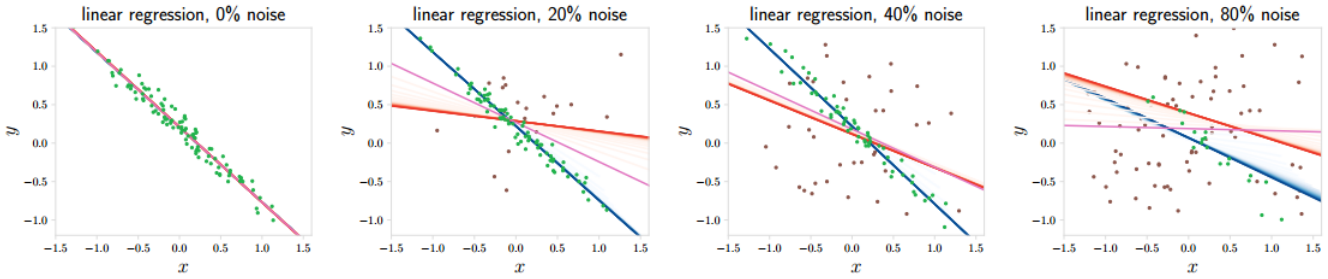


Figure 15: In the presence of random noise with the same mean as that of clean data, TERM with negative t's (blue) can still surpass outliers in all cases, while ERM (purple) and TERM with positive t's (red) over-fit to corrupted data. While the performance drops for 80% noise, TERM can still learn useful information, and achieves much lower error than ERM.

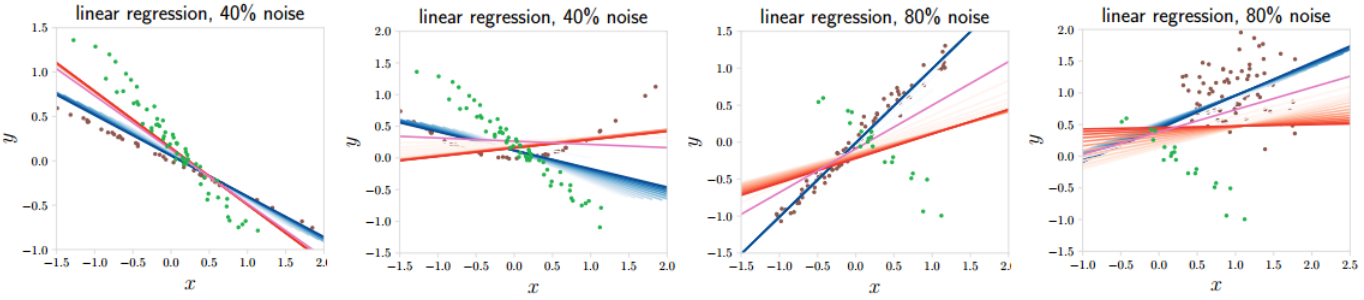


Figure 16: TERM with negative t's (blue) cannot fit clean data if the noisy samples (brown) are adversarial or structured in a manner that differs substantially from the underlying true distribution.

5.1.2 Robust Classification

DNNs are known to easily over-fit to corrupted labels [YAS12]. Although the theoretical properties of TERM do not directly cover objectives with DNNs, the authors show that TERM can be applied to DNNs to achieve robustness to noisy training labels. The authors compare TERM with ERM and several state-of-the-art approaches ([KPK10] [Ren+18] [YAS12] [Kri12]) with a fraction of the training labels corrupted with

uniform noise. The results, shown in Figure 17, indicate that TERM performs competitively with 20% noise and outperforms all baselines in the high noise regimes. The authors also study a simpler, two-dimensional logistic regression problem and find that TERM with t = -2 is similarly robust across the considered 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)

Figure 17: TERM is competitive with robust classification baselines, and is superior in high noise regimes.

6 Conclusion & Future works

In this paper, the authors present a framework called Tilted Empirical Risk Minimization (TERM) as an extension to the traditional Empirical Risk Minimization (ERM) approach in machine learning. The TERM framework is established to address various issues faced by ERM, including robustness to noise, class imbalance, fairness, and generalization, among others. The authors showed, both theoretically and empirically, that TERM outperforms ERM and delivers competitive performance with state-of-the-art, problem-specific methods in a wide range of applications. The authors hope that the TERM framework will allow machine learning practitioners to easily modify the ERM objective to handle practical concerns such as fairness, robustness to outliers, and robust performance on new, unseen data. The authors also note that the TERM framework has the potential to magnify the impact of biased or corrupted data if not used correctly and therefore require a thorough understanding of the implications of the modified objective. The authors used benchmark datasets for their experiments but acknowledge that some of these datasets contain sensitive information and require further examination. The authors believe that the TERM framework has rich implications and wide applicability beyond what is studied in this work and could be applied to a wide range of real-world machine learning applications in the future.

Semi-Supervised setup has not been applied with this loss function so one of interesting area of working in this field could be working on semi-supervised setup with this loss function or similar to this loss function loss entropic risk. It is worth mentioning that the TERM framework is not limited to just supervised learning. In fact, the semi-supervised setup is an interesting area for future work with this loss function. To date, the TERM framework has only been applied to supervised learning, but there is potential for it to be adapted for use in semi-supervised learning. This could potentially lead to improved performance in situations where labeled data is limited. Additionally, exploring the use of similar loss functions, such as entropic risk, in the semi-supervised setup could also be a valuable direction for future research. This highlights the versatility and potential of the TERM framework, as it can be adapted and modified to address different machine learning challenges and real-world applications.

Appendices

Theorem 1 proof:
The article discusses a result in optimization that states that solving a specific optimization problem called TERM using gradient-based methods is efficient for small to moderate values of a parameter t. This is supported by previous works such as [KNS16] and [LR21]. The results are confirmed through experiments on real-world datasets. The authors note that solving for the min-max solution would be similar to solving TERM as t approaches infinity, but this approach would be less efficient.
Theorem 2 proof:
If the function $\sum_{i\in[N]} p_i f\left(x_i;\theta\right)$ is μ-PL for any $\mathbf{p}\in\Delta_N$, then the function $\widetilde{R}(t;\theta)$ is μ-PL. The theorem applies to both convex and non-convex smooth functions and holds under the assumption of a PL condition. If the function $f(x;\theta)$ is Lipschitz, the explicit smoothness parameter can be derived from Lowy and Razaviyayn (2021, Lemma 5.3). The theorem states results without assuming the PL condition for completeness.
If $\sum_{i\in[N]} p_i f\left(x_i;\theta\right)$ is μ-PL for any $\mathbf{p}\in\Delta_N$, then $\widetilde{R}(t;\theta)$ is μ-PL [Qi+20]. $\widetilde{R}(t;\theta)$ is $\beta_{\max}$ smooth for $t<0$ and its smoothness parameter scales linearly with t for $t>0$.
Theorem 2 applies to both convex and non-convex smooth functions satisfying PL conditions. Again, here we can plug in explicit smoothness parameter [LR21] if $f(x;\theta)$ is Lipschitz. We next state results without the PL condition assumption for completeness.
Theorem 3 proof:
To prove our convergence results in Theorem 3, we first prove a lemma below. Lemma. Denote $k_t := \arg\max_k \left(k < \frac{2e}{\mu} + \frac{etLBe^{t\left(\tilde{F}_{\max}-\tilde{F}_{\min}\right)}}{\mu k}\right)$. Let $\lambda = 1-\frac{1}{2e}$, and

$$\alpha_k = \begin{cases} \frac{1}{tLBe^{t\left(\tilde{F}_{\max}-\tilde{F}_{\min}\right)}+1}, & \text{if } k\leqslant k_t \\ \frac{2e}{\mu k}, & \text{otherwise} \end{cases} \tag{40}$$

then for any k,

$$\mathbb{E}\left[e^{t\left(\frac{\tilde{R}_k}{R_k}-\tilde{R}_k\right)} \mid \theta_1,\ldots,\theta_k\right] \leqslant 2e, \tag{41}$$

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

$$e^{t\tilde{\tilde{R}}_{k+1}} = \lambda e^{tf(\xi_k,\theta_k)} + (1-\lambda)e^{t\tilde{R}_k} \tag{42}$$

Taking conditional expectation $\mathbb{E}\left[\cdot \mid \theta_1,\ldots,\theta_{k+1}\right]$ on both sides gives

$$\mathbb{E}\left[e^{t\left(\tilde{R}_{k+1}-\tilde{R}_k\right)} \mid \theta_1,\ldots,\theta_{k+1}\right] = \tag{43}$$

$$\lambda\mathbb{E}\left[e^{t\left(f(\xi_k;\theta_k)-\tilde{R}_k\right)} \mid \theta_1,\ldots,\theta_{k+1}\right] + (1-\lambda)\mathbb{E}\left[e^{t\left(\tilde{R}_k-\tilde{R}_k\right)} \mid \theta_1,\ldots,\theta_{k+1}\right] = \tag{44}$$

$$\lambda + (1-\lambda)\mathbb{E}\left[e^{t\left(\tilde{R}_k-\tilde{R}_k\right)} \mid \theta_1,\ldots,\theta_k\right]. \tag{45}$$

For any k, we have

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

Therefore,

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

and

$$e^{-t\alpha_k LBe^{t\left(\tilde{F}_{\max}-\tilde{F}_{\min}\right)}} \leqslant e^{t\left(\tilde{R}_k-\tilde{R}_{k+1}\right)} = \frac{\sum_{i\in[N]} e^{tf(x_i;\theta_k)}}{\sum_{i\in[N]} e^{tf(x_i;\theta_{k+1})}} \leqslant e^{t\alpha_k LBe^{t\left(\tilde{F}_{\max}-\tilde{F}_{\min}\right)}}, \tag{46}$$

$$e^{-t\alpha_k LBe^{t\left(\tilde{F}_{\max}-\tilde{F}_{\min}\right)}}\mathbb{E}\left[e^{t\left(\tilde{R}_{k+1}-\tilde{R}_{k+1}\right)} \mid \theta_1,\ldots,\theta_{k+1}\right] \leqslant \mathbb{E}\left[e^{t\left(\tilde{R}_{k+1}-\tilde{R}_k\right)} \mid \theta_1,\ldots,\theta_{k+1}\right] \tag{47}$$

$$\leqslant e^{t\alpha_k LBe^{t\left(\tilde{F}_{\max}-\tilde{F}_{\min}\right)}}\mathbb{E}\left[e^{t\left(\tilde{R}_{k+1}-\tilde{R}_{k+1}\right)} \mid \theta_1,\ldots,\theta_{k+1}\right]. \tag{48}$$

Hence,

$$e^{-t\alpha_k LBe^{t\left(\tilde{F}_{\max}-\tilde{F}_{\min}\right)}}\mathbb{E}\left[e^{t\left(\tilde{R}_{k+1}-\widetilde{R}_{k+1}\right)} \mid \theta_1,\ldots,\theta_{k+1}\right] \tag{49}$$

$$\leqslant \lambda + (1-\lambda)\mathbb{E}\left[e^{t\left(\tilde{\varkappa}_k-\widetilde{R}_k\right)} \mid \theta_1,\ldots,\theta_k\right] \tag{50}$$

$$\leqslant e^{t\alpha_k LBe^{t\left(\tilde{F}_{\max}-\tilde{F}_{\min}\right)}}\mathbb{E}\left[e^{t\left(\tilde{R}_{k+1}-\tilde{R}_{k+1}\right)} \mid \theta_1,\ldots,\theta_{k+1}\right] \tag{51}$$

When $k \leqslant k_t$, under the learning rate α_k, we have

$$\alpha_k LBe^{t\left(\tilde{F}_{\max}-\tilde{F}_{\min}\right)} < 1. \tag{52}$$

Hence,

$$\mathbb{E}\left[e^{t\left(\tilde{R}_k-\tilde{R}_k\right)} \mid \theta_1,\ldots,\theta_k\right] \leqslant e\left(\lambda + (1-\lambda)\mathbb{E}\left[e^{t\left(\left(\tilde{R}_{k-1}-\widetilde{R}_{k-1}\right)\right.} \mid \theta_1,\ldots,\theta_{k-1}\right]\right) \tag{53}$$

$$\leqslant e + \frac{1}{2}\mathbb{E}\left[e^{t\left(\tilde{R}_{k-1}-\widetilde{R}_{k-1}\right)} \mid \theta_1,\ldots,\theta_{k-1}\right] \tag{54}$$

$$\leqslant \cdots \leqslant e\left(2-\frac{1}{2^{k-2}}\right) + \frac{1}{2^{k-1}}\mathbb{E}\left[e^{t\left(\tilde{R}_1-\widetilde{R}_1\right)} \mid \theta_1\right] \leqslant 2e. \tag{55}$$

When $k > k_t$,

$$\alpha_k = \frac{2e}{\mu k} < \frac{k}{k + tLBe^{t\left(\tilde{F}_{\max}-\tilde{F}_{\min}\right)}} \tag{56}$$

Similarly, we have

$$\mathbb{E}\left[e^{t\left(\tilde{R}_k-\tilde{R}_k\right)} \mid \theta_1,\ldots,\theta_k\right] \leqslant e^{t\alpha_k LBe^{t\left(\tilde{F}_{\max}-\tilde{F}_{\min}\right)}}\left(\lambda + (1-\lambda)\mathbb{E}\left[e^{t\left(\tilde{\varepsilon}_{R_{k-1}}-\tilde{R}_{k-1}\right)} \mid \theta_1,\ldots,\theta_{k-1}\right]\right) \tag{57}$$

$$\leqslant \cdots \leqslant 2e \tag{58}$$

which completes the proof.

Proof of Theorem 3. Denote the empirical optimal solution $\breve{\theta}(t)$ as θ^*. Denote the tilted stochastic gradient on data ζ_k as g_k, where

$$g_k = \frac{e^{tf(\zeta_k;\theta_k)}}{e^{t\tilde{R}_k}}\nabla f\left(\zeta_k;\theta_k\right) = \frac{e^{t\tilde{R}_k}}{e^{t\tilde{R}_k}}\frac{e^{tf(\zeta_k;\theta_k)}}{e^{t\tilde{R}_k}}\nabla f\left(\zeta_k;\theta_k\right) = \frac{e^{t\tilde{R}_k}}{e^{t\tilde{\tau}_k}}\nabla\widetilde{R}_k\left(\zeta_k\right). \tag{59}$$

Therefore, for any $k \geqslant 1$,

$$\mathbb{E}\left[\langle\theta_k-\theta^*, g_k\rangle\right] = \mathbb{E}\left[\mathbb{E}\left[\langle\theta_k-\theta^*, g_k\rangle \mid \theta_1,\ldots,\theta_k\right]\right] \tag{60}$$

$$= \mathbb{E}\left[\langle\theta_k-\theta^*, \mathbb{E}\left[g_k \mid \theta_1,\ldots,\theta_k\right]\rangle\right] \tag{61}$$

$$= \mathbb{E}\left[\left\langle\theta_k-\theta^*, \mathbb{E}\left[e^{t\left(\tilde{R}_k-\tilde{R}_k\right)} \mid \theta_1,\ldots,\theta_k\right]\mathbb{E}\left[\nabla\widetilde{R}_k\left(\zeta_k\right) \mid \theta_1,\ldots,\theta_k\right]\right\rangle\right] \tag{62}$$

$$\geqslant \frac{1}{2e}\mathbb{E}\left[\left\langle\theta_k-\theta^*, \nabla\widetilde{R}\left(\theta_k\right)\right\rangle\right] \quad \left(\mathbb{E}\left[e^{t\left(\tilde{R}_k-\tilde{R}_k\right)} \mid \theta_1,\ldots,\theta_k\right] \geqslant 1/\mathbb{E}\left[e^{t\left(\tilde{R}_k-\tilde{R}_k\right)} \mid \theta_1,\ldots,\theta_k\right]\right) \tag{63}$$

$$\geqslant \frac{\mu}{2e}\mathbb{E}\left[\left|\theta_k-\theta^*\right|^2\right] \quad (\mu\text{-strong convexity of } \widetilde{R}) \tag{64}$$

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

$$\mathbb{E}\left[\left|\theta_{k+1}-\theta^*\right|^2\right]=\mathbb{E}\left[\left|\theta_k-\alpha_k g_k-\theta^*\right|^2\right] \tag{65}$$

$$=\mathbb{E}\left[\left|\theta_k-\theta^*\right|^2\right]-2\alpha_k\mathbb{E}\left[\langle\theta_k-\theta^*,g_k\rangle\right]+\alpha_k^2\mathbb{E}\left[\left|g_k\right|^2\right] \tag{66}$$

$$\leqslant\left(1-\frac{\alpha_k\mu}{e}\right)\mathbb{E}\left[\left|\theta_k-\theta^*\right|^2\right]+\alpha_k^2\mathbb{E}\left[\left|e^{t\left(\tilde{R}_k-\tilde{R}_k\right)}\nabla\widetilde{R}_k\left(\zeta_k\right)\right|^2\right] \tag{67}$$

$$\leqslant\left(1-\frac{2}{k}\right)\mathbb{E}\left[\left|\theta_k-\theta^*\right|^2\right]+\frac{4e^2B^2e^{2t\left(\tilde{F}_{\max}-\tilde{F}_{\min}\right)}}{\mu^2k^2}. \tag{68}$$

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

$$\mathbb{E}\left[\left|\theta_k-\theta^*\right|^2\right]\leqslant\left(1-\frac{\mu}{e\left(tLBe^{t\left(\tilde{F}_{\max}-\tilde{F}_{\min}\right)}+1\right)}\right)\mathbb{E}\left[\left|\theta_{k-1}-\theta^*\right|^2\right]+\frac{B^2e^{2t\left(\tilde{F}_{\max}-\tilde{F}_{\min}\right)}}{\left(1+tLBe^{t\left(\tilde{F}_{\max}-\tilde{F}_{\min}\right)}\right)^2} \tag{69}$$

We can thus prove

$$\mathbb{E}\left[\left|\theta_{k_t}-\theta^*\right|^2\right]\leqslant\max\left\{\mathbb{E}\left[\left|\theta_1-\theta^*\right|^2\right],\frac{B^2e^{2t\left(\tilde{F}_{\max}-\tilde{F}_{\min}\right)+1}}{\mu\left(1+tLBe^{t\left(\tilde{F}_{\max}-\tilde{F}_{\min}\right)}\right)}\right\} \tag{70}$$

Let

$$V_t=\max\left\{k_t\mathbb{E}\left[\left|\theta_{k_t}-\theta^*\right|^2\right],\frac{4B^2e^{2+2t\left(\tilde{F}_{\max}-\tilde{F}_{\min}\right)}}{\mu^2}\right\}. \tag{71}$$

We next prove for $k\geqslant k_t$,

$$\mathbb{E}\left[\left|\theta_k-\theta^*\right|^2\right]\leqslant\frac{V_t}{k} \tag{72}$$

Suppose $\mathbb{E}\left[\left|\theta_k-\theta^*\right|^2\right]\leqslant\frac{V_t}{k}$. we have

$$\mathbb{E}\left[\left|\theta_{k+1}-\theta^*\right|^2\right]\leqslant\left(1-\frac{2}{k}\right)\mathbb{E}\left[\left|\theta_k-\theta^*\right|^2\right]+\frac{4e^2B^2e^{2t\left(\tilde{F}_{\max}-\tilde{F}_{\min}\right)}}{k^2\mu^2} \tag{73}$$

$$\leqslant\left(1-\frac{2}{k}\right)\frac{V_t}{k}+\frac{V_t^2}{k^2} \tag{74}$$

$$\leqslant\frac{V_t}{k+1} \tag{75}$$

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

Theorem 4 proof:
Proof of Theorem 4. Assume $\widetilde{R}(t;\theta)$ is non-convex and β-smooth, we have

$$\widetilde{R}_{k+1}-\widetilde{R}_k-\left\langle\nabla\widetilde{R}_k,\theta_{k+1}-\theta_k\right\rangle\leqslant\frac{\beta}{2}\left|\theta_{k+1}-\theta_k\right|^2 \tag{76}$$

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

$$\theta_{k+1} - \theta_k = -\alpha_k \frac{e^{t(\zeta_k;\theta_k)}}{e^{t\tilde{\tilde{R}}_k}} \nabla f\left(\zeta_k;\theta_k\right) = -\alpha_k e^{t\left(\tilde{R}_k - \tilde{R}_k\right)} \nabla \widetilde{R}_k\left(\zeta_k\right) \tag{77}$$

gives

$$\widetilde{R}_{k+1} - \widetilde{R}_k + \alpha_k \left\langle \nabla \widetilde{R}_k, e^{t\left(\tilde{R}_k - \tilde{R}_k\right)} \nabla \widetilde{R}_k\left(\zeta_k\right) \right\rangle \leqslant \frac{\beta}{2} \left| \alpha_k e^{t\left(\tilde{R}_k - \tilde{R}_k\right)} \nabla \widetilde{R}_k\left(\zeta_k\right) \right|^2 \tag{78}$$

First, we note

$$\left| \alpha_k^2 e^{t\left(\tilde{R}_k - \tilde{R}_k\right)} \nabla \tilde{R}_k\left(\zeta_k\right) \right|^2 \leqslant \alpha_k^2 e^{2t\left(\tilde{F}_{\max} - \tilde{F}_{\min}\right)} \left| \nabla \widetilde{R}_k\left(\zeta_k\right) \right|^2 \tag{79}$$

Take expectation on both sides,

$$\mathbb{E}\left[\widetilde{R}_{k+1}\right] - \mathbb{E}\left[\widetilde{R}_k\right] + \alpha_k \mathbb{E}\left[\left\langle \nabla \widetilde{R}_k, e^{t\left(\tilde{R}_k - \tilde{R}_k\right)} \nabla \widetilde{R}_k\left(\zeta_k\right) \right\rangle\right] \leqslant \frac{\beta \alpha_k^2 e^{2t\left(\tilde{F}_{\max} - \tilde{F}_{\min}\right)} B^2}{2} \tag{80}$$

Let

$$k_t := \left\lceil \frac{2\left(\tilde{F}_{\max} - \tilde{F}_{\min}\right) t^2 L^2}{\beta e^2} \right\rceil \tag{81}$$

For any $k \geqslant k_t$, let

$$\alpha_k = \frac{\sqrt{2\left(\widetilde{F}_{\max} - \widetilde{F}_{\min}\right)}}{e^{t\left(\tilde{F}_{\max} - \tilde{F}_{\min}\right)} \sqrt{\beta B^2 K}}. \tag{82}$$

For $k < k_t$, let

$$\alpha_k = \frac{1}{tLBe^{t\left(\tilde{F}_{\max} - \tilde{F}_{\min}\right)} + 1}. \tag{83}$$

We have for any $k \geqslant 1$,

$$\alpha_k t L B e^{t\left(\tilde{F}_{\max} - \tilde{F}_{\min}\right)} \leqslant 1. \tag{84}$$

Therefore, for any $k \geqslant 1$,

$$\mathbb{E}\left[e^{t\left(\tilde{\tilde{R}}_k - \tilde{R}_k\right)} \mid \theta_1, \ldots, \theta_k\right] \leqslant 2e. \tag{85}$$

Thus, for any $k \geqslant 1$,

$$\mathbb{E}\left[\left|\nabla \tilde{R}_k\right|^2\right] + \frac{2e}{\alpha_k}\left(\mathbb{E}\left[\widetilde{R}_{k+1}\right] - \mathbb{E}\left[\widetilde{R}_k\right]\right) \leqslant \beta \alpha_k e^{2t\left(\tilde{F}_{\max} - \tilde{F}_{\min}\right)} e B^2. \tag{86}$$

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

$$\frac{1}{K}\sum_{k=k_t}^{K} \mathbb{E}\left[\left|\nabla \widetilde{R}_k\right|^2\right] + \frac{2e\left(\mathbb{E}\left[\tilde{R}_{K+1}\right] - \mathbb{E}\left[\widetilde{R}_{k_t}\right]\right)}{\alpha_k K} \leqslant \beta \alpha_k e^{2t\left(\tilde{F}_{\max} - \tilde{F}_{\min}\right)} e B^2. \tag{87}$$

$$\frac{1}{K}\sum_{k=k_t}^{K} \mathbb{E}\left[\left|\nabla \widetilde{R}_k\right|^2\right] \leqslant \beta \alpha_k e^{2t\left(\tilde{F}_{\max} - \tilde{F}_{\min}\right)} e B^2 + \frac{2e\left(\mathbb{E}\left[\widetilde{R}_{k_t} - \widetilde{R}_{K+1}\right]\right)}{\alpha_k K} \tag{88}$$

$$\leqslant \beta \alpha_k e^{2t\left(\tilde{F}_{\max} - \tilde{F}_{\min}\right)} e B^2 + \frac{2e\left(\tilde{F}_{\max} - \tilde{F}_{\min}\right)}{\alpha_k K} \tag{89}$$

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

$$\frac{1}{K}\sum_{k=k_t}^{K}\mathbb{E}\left[\left|\nabla\widetilde{R}_k\right|^2\right] \leqslant \sqrt{8}Be^{t\left(\tilde{F}_{\max}-\tilde{F}_{\min}\right)+1}\sqrt{\frac{\beta\left(\widetilde{F}_{\max}-\widetilde{F}_{\min}\right)}{K}}, \tag{90}$$

completing the proof.

References

[Ben62] George Bennett. "Probability Inequalities for the Sum of Independent Random Variables". In: *Journal of the American Statistical Association* 57.297 (1962), pp. 33–45. ISSN: 01621459. URL: http://www.jstor.org/stable/2282438 (visited on 02/12/2023).

[Hub64] Peter J. Huber. "Robust Estimation of a Location Parameter". In: *The Annals of Mathematical Statistics* 35.1 (1964), pp. 73–101. DOI: 10.1214/aoms/1177703732. URL: https://doi.org/10.1214/aoms/1177703732.

[DZ98] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Springer, 1998. ISBN: 978-1-4612-5320-4. DOI: 10.1007/978-1-4612-5320-4. URL: https://doi.org/10.1007/978-1-4612-5320-4.

[GB04] Yves Grandvalet and Yoshua Bengio. "Semi-Supervised Learning by Entropy Minimization". In: NIPS'04 (2004), pp. 529–536.

[CZ05] Olivier Chapelle and Alexander Zien. "Semi-Supervised Classification by Low Density Separation". In: *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*. Ed. by Robert G. Cowell and Zoubin Ghahramani. Vol. R5. Proceedings of Machine Learning Research. Reissued by PMLR on 30 March 2021. PMLR, June 2005, pp. 57–64. URL: https://proceedings.mlr.press/r5/chapelle05b.html.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, July 2006. ISBN: 0471241954.

[WJ08] Martin J. Wainwright and Michael I. Jordan. "Graphical Models, Exponential Families, and Variational Inference". In: *Foundations and Trends® in Machine Learning* 1.1–2 (2008), pp. 1–305. ISSN: 1935-8237. DOI: 10.1561/2200000001. URL: http://dx.doi.org/10.1561/2200000001.

[MP09] Andreas Maurer and Massimiliano Pontil. *Empirical Bernstein Bounds and Sample Variance Penalization*. 2009. DOI: 10.48550/ARXIV.0907.3740. URL: https://arxiv.org/abs/0907.3740.

[KPK10] M. Kumar, Benjamin Packer, and Daphne Koller. "Self-Paced Learning for Latent Variable Models". In: *Advances in Neural Information Processing Systems*. Ed. by J. Lafferty et al. Vol. 23. Curran Associates, Inc., 2010. URL: https://proceedings.neurips.cc/paper/2010/file/e57c6b956a6521b28495f2886ca0977a-Paper.pdf.

[PR11] E. Y. Pee and Johannes O. Royset. "On Solving Large-Scale Finite Minimax Problems Using Exponential Smoothing". In: *Journal of Optimization Theory and Applications* 148 (2011), pp. 390–421.

[Kri12] Alex Krizhevsky. "Learning Multiple Layers of Features from Tiny Images". In: *University of Toronto* (May 2012).

[YAS12] Yao-liang Yu, Özlem Aslan, and Dale Schuurmans. "A Polynomial-time Form of Robust Regression". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper/2012/file/ae5e3ce40e0404a45ecacaaf05e5f735-Paper.pdf.

[Lee13] Dong-Hyun Lee. "Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks". In: 2013.

[Ge+15] Rong Ge et al. "Escaping From Saddle Points — Online Stochastic Gradient for Tensor Decomposition". In: *Proceedings of The 28th Conference on Learning Theory*. Ed. by Peter Grünwald, Elad Hazan, and Satyen Kale. Vol. 40. Proceedings of Machine Learning Research. Paris, France: PMLR, Mar. 2015, pp. 797–842. URL: https://proceedings.mlr.press/v40/Ge15.html.

[KNS16] Hamed Karimi, Julie Nutini, and Mark Schmidt. *Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition*. 2016. DOI: 10.48550/ARXIV.1608.04636. URL: https://arxiv.org/abs/1608.04636.

[Bha+17] Kush Bhatia et al. "Consistent Robust Regression". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/e702e51da2c0f5be4dd354bb3e295d37-Paper.pdf.

[GKS17] Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. *Robust Loss Functions under Label Noise for Deep Neural Networks*. 2017. DOI: 10.48550/ARXIV.1712.09482. URL: https://arxiv.org/abs/1712.09482.

[ND17] Hongseok Namkoong and John C Duchi. "Variance-based Regularization with Convex Objectives". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/5a142a55461d5fef016acfb927fee0bd-Paper.pdf.

[Oli+18] Ivan Olier et al. "Meta-QSAR: a large-scale application of meta-learning to drug design and discovery". In: *Machine Learning* 107 (Jan. 2018). DOI: 10.1007/s10994-017-5685-x.

[Ren+18] Mengye Ren et al. "Learning to Reweight Examples for Robust Deep Learning". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, Oct. 2018, pp. 4334–4343. URL: https://proceedings.mlr.press/v80/ren18a.html.

[Dia+19] Ilias Diakonikolas et al. "Sever: A Robust Meta-Algorithm for Stochastic Optimization". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, Sept. 2019, pp. 1596–1606. URL: https://proceedings.mlr.press/v97/diakonikolas19a.html.

[DN19] John Duchi and Hongseok Namkoong. "Variance-based Regularization with Convex Objectives". In: *Journal of Machine Learning Research* 20.68 (2019), pp. 1–55. URL: http://jmlr.org/papers/v20/17-750.html.

[LT19] Guan-Horng Liu and Evangelos A. Theodorou. *Deep Learning Theory Review: An Optimal Control and Dynamical Systems Perspective*. 2019. DOI: 10.48550/ARXIV.1908.10920. URL: https://arxiv.org/abs/1908.10920.

[Yu+19] Lequan Yu et al. *Uncertainty-aware Self-ensembling Model for Semi-supervised 3D Left Atrium Segmentation*. 2019. DOI: 10.48550/ARXIV.1907.07034. URL: https://arxiv.org/abs/1907.07034.

[CP20] Ruidi Chen and Ioannis Ch. Paschalidis. "Distributionally Robust Learning". In: *Found. Trends Optim.* 4.1–2 (Dec. 2020), pp. 1–243. ISSN: 2167-3888. DOI: 10.1561/2400000026. URL: https://doi.org/10.1561/2400000026.

[Luo+20] Xiangde Luo et al. *Efficient Semi-Supervised Gross Target Volume of Nasopharyngeal Carcinoma Segmentation via Uncertainty Rectified Pyramid Consistency*. 2020. DOI: 10.48550/ARXIV.2012.07042. URL: https://arxiv.org/abs/2012.07042.

[Muk+20] Bhaskar Mukhoty et al. "Globally-convergent Iteratively Reweighted Least Squares for Robust Regression Problems". In: (2020). DOI: 10.48550/ARXIV.2006.14211. URL: https://arxiv.org/abs/2006.14211.

[OHT20] Yassine Ouali, Céline Hudelot, and Myriam Tami. *Semi-Supervised Semantic Segmentation with Cross-Consistency Training*. 2020. DOI: 10.48550/ARXIV.2003.09005. URL: https://arxiv.org/abs/2003.09005.

[Qi+20] Qi Qi et al. *An Online Method for A Class of Distributionally Robust Optimization with Non-Convex Objectives*. 2020. DOI: 10.48550/ARXIV.2006.10138. URL: https://arxiv.org/abs/2006.10138.

[Li+21] Tian Li et al. *On Tilted Losses in Machine Learning: Theory and Applications*. 2021. DOI: 10.48550/ARXIV.2109.06141. URL: https://arxiv.org/abs/2109.06141.

[LR21] Andrew Lowy and Meisam Razaviyayn. *Output Perturbation for Differentially Private Convex Optimization with Improved Population Loss Bounds, Runtimes and Applications to Private Adversarial Training*. 2021. DOI: 10.48550/ARXIV.2102.04704. URL: https://arxiv.org/abs/2102.04704.

[Ver+22] Vikas Verma et al. "Interpolation consistency training for semi-supervised learning". In: *Neural Networks* 145 (Jan. 2022), pp. 90–106. DOI: 10.1016/j.neunet.2021.10.008. URL: https://doi.org/10.1016%2Fj.neunet.2021.10.008.