X

On Tilted Losses in Machine Learning : Theory and Application

AmirHossein Bagheri,Mahyar Jafari Nodeh, Radmehr Karimian

Sharif University of Technology

February 12, 2023

Table of Contents

1. Introduction
    - perspectives on exponential tilting
    - Motivation
2. TERM : Properties and Interpretations
    - Assumptions
    - General properties of TERM
    - Interpolation 1
    - Interpolation 2
3. Connection to other risk measures
    - TERM and Renyi cross entropy
    - TERM as Regularizer
4. Solving TERM
5. Hierarchical Multi-Objective Tilting
6. TERM in Practice: Use Cases
7. Related Approaches in Machine Learning
8. Discussion and Conclusion

Contents

1. Introduction
   - perspectives on exponential tilting
   - Motivation
2. TERM : Properties and Interpretations
   - Assumptions
   - General properties of TERM
   - Interpolation 1
   - Interpolation 2
3. Connection to other risk measures
   - TERM and Renyi cross entropy
   - TERM as Regularizer
4. Solving TERM
5. Hierarchical Multi-Objective Tilting
6. TERM in Practice: Use Cases
7. Related Approaches in Machine Learning
8. Discussion and Conclusion

Introduction

Definitions

for each $\theta \in \Theta \subseteq R^d$ and the Dataset $\{x_1, \cdots x_N\}$ we Define ERM as :

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

Introduction

Definitions

for each $\theta \in \Theta \subseteq R^d$ and the Dataset $\{x_1, \cdots x_N\}$ we Define ERM as :

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

Flaws :

Introduction

Definitions

for each $\theta \in \Theta \subseteq R^d$ and the Dataset $\{x_1, \cdots x_N\}$ we Define ERM as :

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

Flaws :

- average performance is not an appropriate surrogate for the problem of interest

Proposed solution = TERM

for a real-valued hyperparameter,$t \in R^{\backslash 0}$, TERM is given by :

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

- $\tilde{R}(+\infty;\theta) =$ max-loss
- $\tilde{R}(-\infty;\theta) =$ min-loss
- $\tilde{R}(0;\theta) =$ ERM

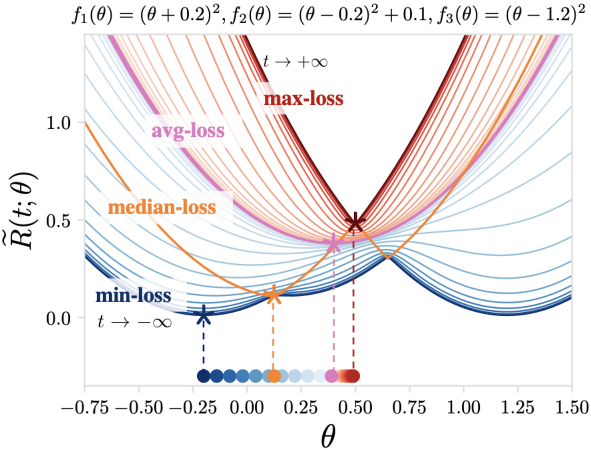


Figure: TERM for different values of t

Let $\mathcal{P} := \{p_\theta\}$.
Define information of x under θ as :

$$f(x:\theta) := -\log p_\theta(x)$$

Also Define : (Cumulant Generating Function)

$$\Lambda_X(t;\theta) := \log\left(\mathop{\mathbb{E}}_{x\sim p}\left[e^{tf(X;\theta)}\right]\right) = \log\sum_x p(x)p_\theta(x)^{-t}$$

and (Cumulant Generating Function) :

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

Let $\mathcal{P} := \{p_\theta\}$.
Define information of x under θ as :

$$f(x:\theta) := -\log p_\theta(x)$$

Also Define : (Cumulant Generating Function)

$$\Lambda_X(t;\theta) := \log\left(\underset{x\sim p}{\mathbb{E}}\left[e^{tf(X;\theta)}\right]\right) = \log\sum_x p(x)p_\theta(x)^{-t}$$

and (Cumulant Generating Function) :

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

proper scaling of $\tilde{R}(t;\theta)$

Example

- Statistics : Convergence properties of statistical estiimation

Example

- Statistics : Convergence properties of statistical estiimation
- Applied Probabilty : Concentration bounds in large deviation theory

Example

- Statistics : Convergence properties of statistical estiimation
- Applied Probabilty : Concentration bounds in large deviation theory
- Information theory : error exponents in channel coding - probability of error in list decoding - computational cost in sequential decoding

Example

- Statistics : Convergence properties of statistical estiimation
- Applied Probabilty : Concentration bounds in large deviation theory
- Information theory : error exponents in channel coding - probability of error in list decoding - computational cost in sequential decoding
- Machine Learning : robust regression - sequential decision making

Motivation Example

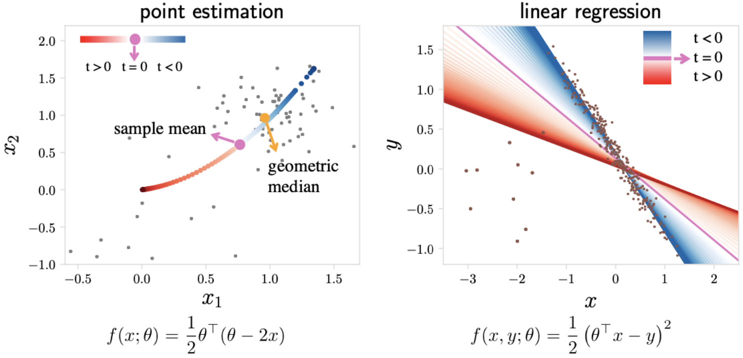


Figure: Motivation examples

Contents

1. Introduction
   - perspectives on exponential tilting
   - Motivation
2. TERM : Properties and Interpretations
   - Assumptions
   - General properties of TERM
   - Interpolation 1
   - Interpolation 2
3. Connection to other risk measures
   - TERM and Renyi cross entropy
   - TERM as Regularizer
4. Solving TERM
5. Hierarchical Multi-Objective Tilting
6. TERM in Practice: Use Cases
7. Related Approaches in Machine Learning
8. Discussion and Conclusion

Assumptions

Assumptions

1. (Continuous differentiability) : For $i \in [N]$ the loss function $f(x;\theta)$ belongs to the differentiabilty class C^1 with respect to $\theta \in \Theta \subseteq \mathbb{R}^d$

Assumptions

Assumptions

1. (Continuous differentiability) : For $i \in [N]$ the loss function $f(x;\theta)$ belongs to the differentiabilty class C^1 with respect to $\theta \in \Theta \subseteq \mathbb{R}^d$
2. (Smoothness and strong convexity) : for any $i \in [N], f(x_i;\theta)$ belongs to differentiabilty class C^2 with respect to θ, we further assume that

$$\beta_{min}\boldsymbol{I} \preceq \nabla^2_{\theta\theta^T} f(x_i;\theta) \preceq \beta_{max}\boldsymbol{I}$$

Assumptions

Assumptions

3. (Generalized linear model condition) Assume that

$$f(x;\theta) = A(\theta) - \theta^T T(x),$$

where A(.) is convex such :

$$\beta_{min}\boldsymbol{I} \preceq \nabla^2_{\theta\theta^T} A(\theta) \preceq \beta_{max}\boldsymbol{I}$$

and ,

$$\sum_{i\in[N]} T(x_i)T(x_i)^T \succeq 0$$

Assumptions

Define

$$\overset{\smile}{\theta}(t) \in \underset{\theta\in\Theta}{argmin}\tilde{R}(t;\theta)$$

and,

$$\tilde{F}(t) := \tilde{R}(t;\overset{\smile}{\theta}(t))$$

Then :

Assumptions

4. (Strict saddle property) for all t $\in \mathbb{R}$, $\tilde{R}(t;\theta)$ is strict saddle : $\nabla^2_{\theta\theta^T}\tilde{R}(t;\theta) \succ 0$, and for all stationary solutions, $\lambda_{min}(\nabla^2_{\theta\theta^T}\tilde{R}(t;\theta)) < 0$

General Properties

Lemma

Lemma 1: (Lipschitzness of $\tilde{R}(t;\theta)$) : for any t and θ, if for i $\in [N]$, $f(x_i;\theta)$ is L-Lip in θ, then $\tilde{R}(t;\theta)$ is L-Lip in θ

General Properties

Lemma

Lemma 1: (Lipschitzness of $\tilde{R}(t;\theta)$) : for any t and θ, if for i $\in [N]$, $f(x_i;\theta)$ is L-Lip in θ, then $\tilde{R}(t;\theta)$ is L-Lip in θ

Lemma

Lemma 2: Under Assumption 2, for any $t \in \mathbb{R}$,

$$\nabla^2_{\theta\theta^T}\tilde{R}(t;\theta) =$$

$$\frac{t}{N}\sum_{i\in[N]}\left(\nabla_\theta f(x_i;\theta) - \nabla_\theta \tilde{R}(t;\theta)\right)\left(\nabla_\theta f(x_i;\theta) - \nabla_\theta \tilde{R}(t;\theta)\right)^T e^{t(f(x_i;\theta)-\tilde{R}(t;\theta))}$$

$$+\frac{1}{N}\sum_{i\in[N]}\nabla^2_{\theta\theta^T} f(x_i;\theta) e^{t(f(x_i;\theta)-\tilde{R}(t;\theta))}$$

so if $t \in \mathbb{R}^{>0}$:

$$\nabla^2_{\theta\theta^T}\tilde{R}(t;\theta) \succ \beta_{min}\boldsymbol{I}$$

General properties

Lemma

Lemma 3 For any $t \in \mathbb{R}$, let $\beta(t)$ be smoothness parameter of $\tilde{R}$:

$$\beta(t) := \lambda_{max}\left(\nabla^2_{\theta\theta^T}\tilde{R}(t;\theta)\right)$$

Further, for $t \in \mathbb{R}^{\leq 0}$,

$$\beta(t) < \beta_{max}$$

and for $t \in \mathbb{R}^{>0}$,

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

General properties

Theorem

- *Theorem 1: Under Assumption 3:*

$$\frac{\partial}{\partial t}\tilde{R}(t;\theta) \geq 0$$

General properties

Theorem

- *Theorem 1: Under Assumption 3:*

$$\frac{\partial}{\partial t}\tilde{R}(t;\theta) \geq 0$$

- *Theorem 2: Under Assumption 3:*

$$\frac{\partial}{\partial t}\tilde{F}(t) = \frac{\partial}{\partial t}\tilde{R}(t;\breve{\theta}(t)) \geq 0$$

Re-Weighting Samples to Magnify/Suppress Outliers

Lemma

Lemma 5 :

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

where ,

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

Re-Weighting Samples to Magnify/Suppress Outliers

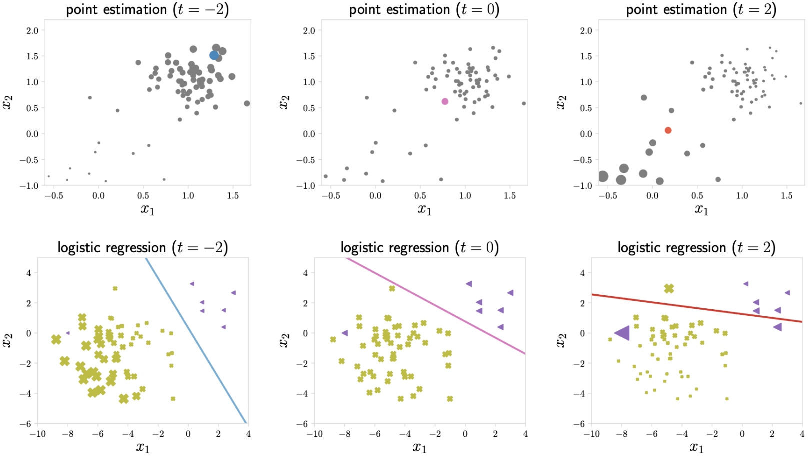


Figure: Interpolation 1

Definition

For $\boldsymbol{u} \in \mathbb{R}^N$, weighted empirical mean with weights $\boldsymbol{w} \in \Delta^N$ be :

$$\hat{\mathbb{E}}_{\boldsymbol{w}}(\boldsymbol{u}) := \sum_{i \in [N]} w_i u_i$$

Tilted empirical mean: just above, but weights are tilted weights.

$$\hat{\mathbb{E}}_t := \hat{\mathbb{E}}_{\boldsymbol{w}(t;\breve{\theta}(t))}(\boldsymbol{u})$$

and variance as :

$$v\hat{a}r_t(\boldsymbol{u}) := \hat{\mathbb{E}}_t \left(u_i - \hat{\mathbb{E}}_t(\boldsymbol{u})\right)^2$$

Empirical Bias/Variance Trade-Off

Theorem

Theorem 3 (Variance Reduction), Let

$$\boldsymbol{f}(\theta) := (f(x_1;\theta),\cdots,f(x_N;\theta))$$

, Under assumption 3 and 4 :

$$\frac{\partial}{\partial t}\left\{v\hat{a}r_\tau(\boldsymbol{f}(\breve{\theta}(t)))\right\}\|_{t=\tau} < 0$$

Empirical Bias/Variance Trade-Off

Theorem

Theorem 3 (Variance Reduction), Let

$$\boldsymbol{f}(\theta) := (f(x_1;\theta), \cdots, f(x_N;\theta))$$

, Under assumption 3 and 4 :

$$\frac{\partial}{\partial t}\left\{v\hat{a}r_{\tau}(\boldsymbol{f}(\breve{\theta}(t)))\right\}\|_{t=\tau} < 0$$

Theorem

Theorem 4: Under assumption 3, 4 for any $t \in \mathbb{R}^{>0}$:

$$\frac{\partial}{\partial t}H(\boldsymbol{w}(\tau;\breve{\theta}(t)))\|_{\tau=t} > 0$$

Contents

1 Introduction
- perspectives on exponential tilting
- Motivation

2 TERM : Properties and Interpretations
- Assumptions
- General properties of TERM
- Interpolation 1
- Interpolation 2

3 Connection to other risk measures
- TERM and Renyi cross entropy
- TERM as Regularizer

4 Solving TERM

5 Hierarchical Multi-Objective Tilting

6 TERM in Practice: Use Cases

7 Related Approaches in Machine Learning

8 Discussion and Conclusion

Coming up with a distributional version of TERM :

$$R_X(t;\theta) := \frac{1}{t}\Lambda_X(t;\theta) = \frac{1}{t}\log\left(\mathbb{E}\left[e^{tf(X;\theta)}\right]\right)$$

$$R(t,\theta) = \frac{1}{t}\log\sum_x p(x)p_\theta^{-t}(x)$$

TERM and Renyi cross entropy

Remember :

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

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

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

So it's straightforward that : $R_X(t;\theta) = H_{1-t}(p||p_\theta)$

And also : $\tilde{R}(t;\theta) = H_{1-t}(\boldsymbol{u}||\boldsymbol{w}(1;\theta))$ where $\boldsymbol{u}$ is uniform N-vector

TERM as a Regularizer to Empirical Risk

The entropic risk of order t can be stated as:

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

Where T is mismatched tilted distribution :

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

also the TERM objective can be written as following :

$$\tilde{R}(t;\theta) = \bar{R}(\theta) + \frac{1}{t} D(\boldsymbol{u}||\boldsymbol{w(t;\theta)})$$

Contents

1 Introduction
- perspectives on exponential tilting
- Motivation

2 TERM : Properties and Interpretations
- Assumptions
- General properties of TERM
- Interpolation 1
- Interpolation 2

3 Connection to other risk measures
- TERM and Renyi cross entropy
- TERM as Regularizer

4 Solving TERM

5 Hierarchical Multi-Objective Tilting

6 TERM in Practice: Use Cases

7 Related Approaches in Machine Learning

8 Discussion and Conclusion

Solving TERM

t-tilted loss remains strongly convex for $t > 0$, so long as the original loss function is strongly convex. On the other hand, for sufficiently large negative t, the t-tilted loss becomes non-convex. Hence, while the t-tilted solutions for positive t are unique, the objective may have multiple (spurious) local minima for negative t even if the original loss function is strongly convex. For negative t, we seek the solution for which the parametric set of t-tilted solutions obtained by sweeping t.()

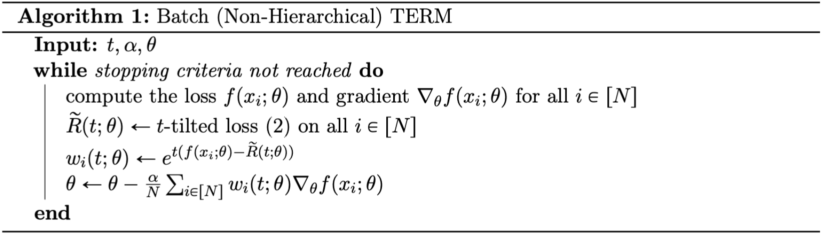
First order Batch

Theorem 9(Convergence of Algorithm 1 for strongly-convex problems)

under Assumption 2, there exist,$\beta_{max} \leqslant C_1 < \infty \, and \, C_2 < \infty$ that do not depend on t such that for any $t \in \mathbb{R}^{>0}$,setting the step size $\alpha = \frac{1}{C1+C2t}$ after k iteration:

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

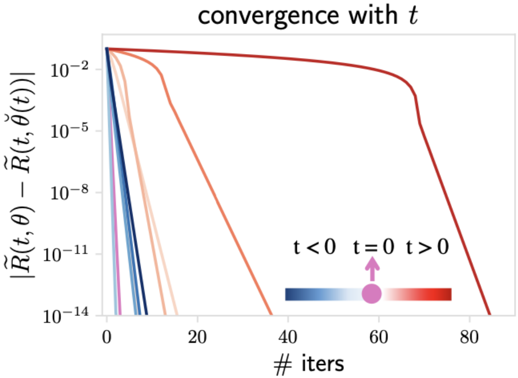
First order Batch



Algorithm 1: Batch (Non-Hierarchical) TERM

```
Input: $t, \alpha, \theta$
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
```

First order Batch



First order Batch

Theorem 10(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 $t \in^{>0}$ with setting step $\alpha = \frac{1}{C1+C2t}$ after k iteration:

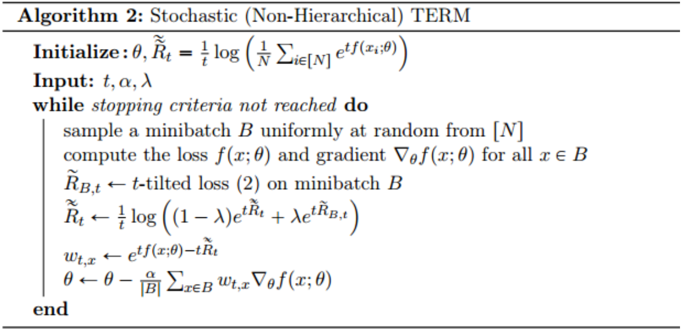
$$\tilde{R}(t; \theta_k) - \tilde{R}(t; \theta(\check{t})) \leqslant (1 - \frac{\mu}{C_1 + C_2 t})^k (\tilde{R}(t; \theta_0) - \tilde{R}(t; \theta(\check{t})))$$

Theorem 10 applies to both convex and non-convex smooth functions satisfying PL conditions.

First order Stochastic Methods

To obtain unbiased stochastic gradients, we need to have access to the normalization weights for each sample.which is often intractable to compute for large-scale problems.Hence, we use $\tilde{\tilde{R}}_t$,a term that incorporates stochastic dynamics, to estimate the tilted objective .For the purpose of analysis, we sample two independent mini-batches to obtain the gradient of the original loss functions.

First order Stochastic Methods



Algorithm 2: Stochastic (Non-Hierarchical) TERM

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

Contents

1. Introduction
   - perspectives on exponential tilting
   - Motivation
2. TERM : Properties and Interpretations
   - Assumptions
   - General properties of TERM
   - Interpolation 1
   - Interpolation 2
3. Connection to other risk measures
   - TERM and Renyi cross entropy
   - TERM as Regularizer
4. Solving TERM
5. Hierarchical Multi-Objective Tilting
6. TERM in Practice: Use Cases
7. Related Approaches in Machine Learning
8. Discussion and Conclusion

Solving Hierarchical TERM

Definition

$$\tilde{J}(t,\tau,\theta) := \frac{1}{t}\log\frac{1}{N}\sum_{g\in[G]}|g|e^{t\tilde{R}_g(\tau,\theta)}$$

lemma

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

where:

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

Solving Hierarchical TERM

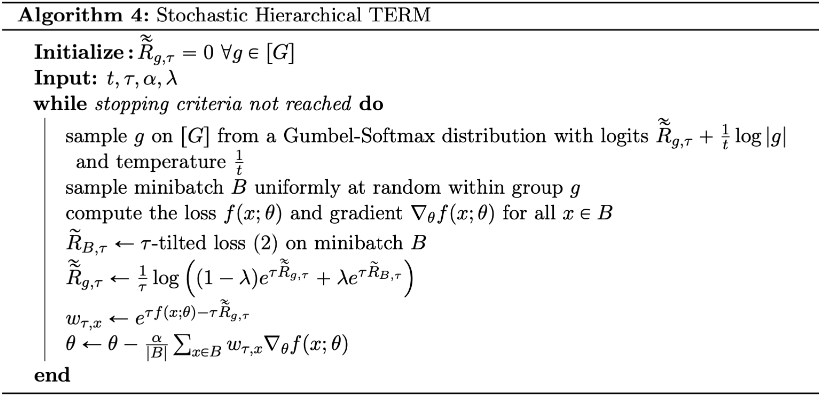
To solve hierarchical TERM in the batch setting, we can directly use gradient-based methods with tilted gradients defined for the hierarchical objective in Lemma.

We next discuss stochastic solvers for hierarchical multi-objective tilting. We extend Algorithm 2 to the multi-objective setting, presented in Algorithm 4. At a high level, at each iteration, group-level tilting is addressed by choosing a group based on the tilted weight vector.

Solving Hierarchical TERM

```
Algorithm 3: Batch Hierarchical TERM
Input: $t, \tau, \alpha$
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
```

Solving Hierarchical TERM



Algorithm 4: Stochastic Hierarchical TERM

```
Initialize: $\tilde{\tilde{R}}_{g,\tau} = 0 \ \forall g \in [G]$
Input: $t, \tau, \alpha, \lambda$
while stopping criteria not reached do
    sample $g$ on $[G]$ from a Gumbel-Softmax distribution with logits $\tilde{\tilde{R}}_{g,\tau} + \frac{1}{t}\log|g|$
      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 $x \in B$
    $\tilde{R}_{B,\tau} \leftarrow \tau$-tilted loss (2) on minibatch $B$
    $\tilde{\tilde{R}}_{g,\tau} \leftarrow \frac{1}{\tau}\log\left((1-\lambda)e^{\tau\tilde{\tilde{R}}_{g,\tau}} + \lambda e^{\tau\tilde{R}_{B,\tau}}\right)$
    $w_{\tau,x} \leftarrow e^{\tau f(x;\theta) - \tau\tilde{\tilde{R}}_{g,\tau}}$
    $\theta \leftarrow \theta - \frac{\alpha}{|B|}\sum_{x\in B} w_{\tau,x}\nabla_\theta f(x;\theta)$
end
```

Contents

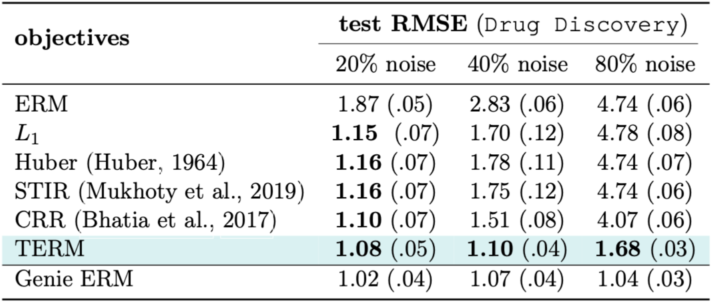
1. Introduction
   - perspectives on exponential tilting
   - Motivation
2. TERM : Properties and Interpretations
   - Assumptions
   - General properties of TERM
   - Interpolation 1
   - Interpolation 2
3. Connection to other risk measures
   - TERM and Renyi cross entropy
   - TERM as Regularizer
4. Solving TERM
5. Hierarchical Multi-Objective Tilting
6. TERM in Practice: Use Cases
7. Related Approaches in Machine Learning
8. Discussion and Conclusion

Mitigating Noisy Outliers $t < 0$

- Robust regression
- Robust classification
- Low-quality annotators

Robust regression

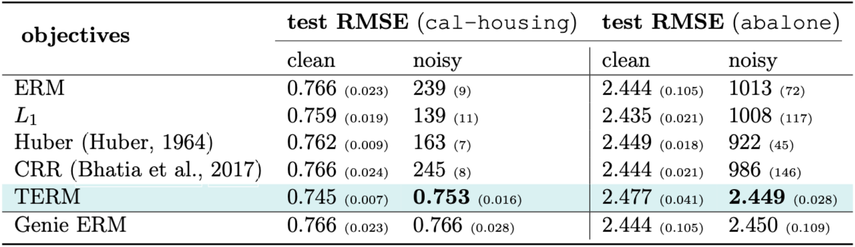
- Label noise



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)

Robust regression

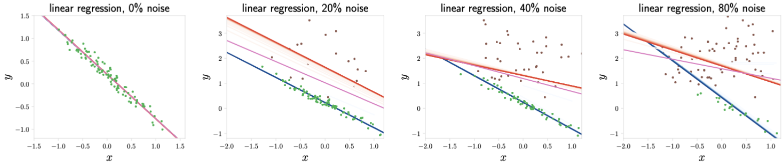
- Label and feature noise



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)

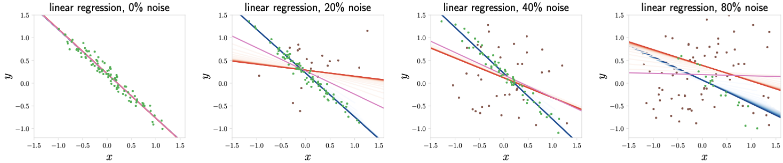
Robust regression

- Unstructured random v.s. adversarial noise.



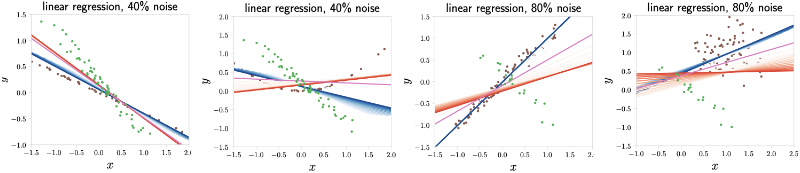
Robust regression

- Unstructured random v.s. adversarial noise.

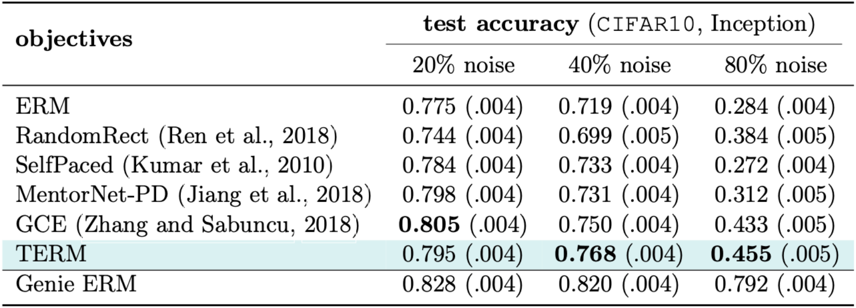


Robust regression

- Unstructured random v.s. adversarial noise.



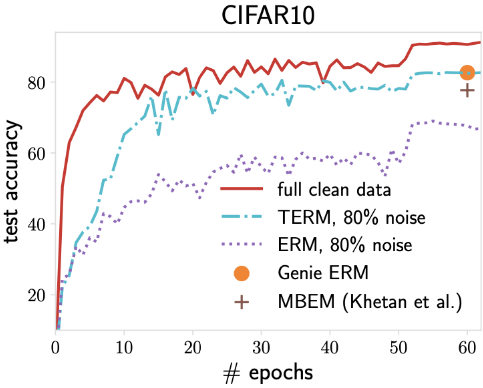
Robust Classification



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)

Robust regression

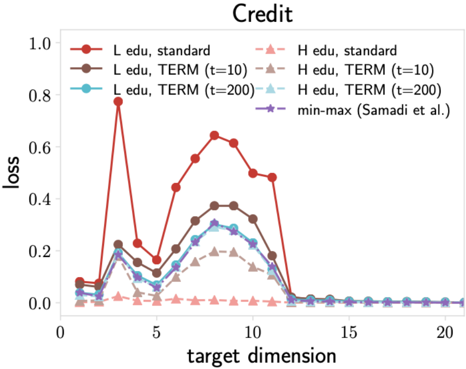
- Low-Quality Annotators



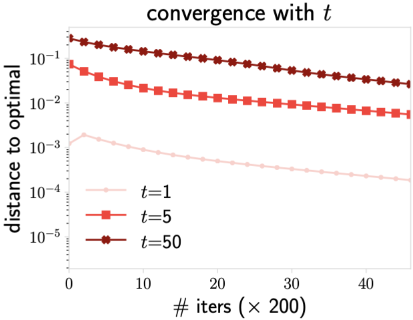
Solving Hierarchical TERM $t > 0$

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

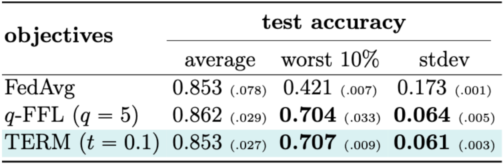
Fair Principal Component Analysis (PCA)



Fair Principal Component Analysis (PCA)

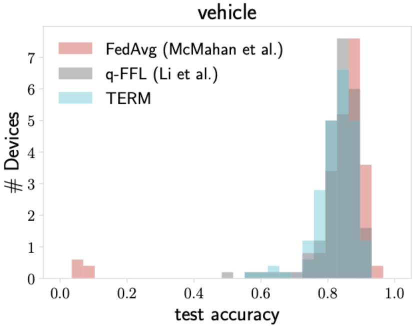


Fair Federated Learning

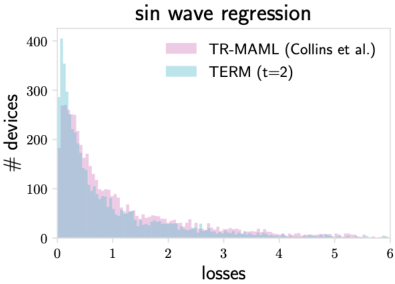


objectives	test accuracy		
	average	worst 10%	stdev
FedAvg	0.853 (.078)	0.421 (.007)	0.173 (.001)
q-FFL ($q = 5$)	0.862 (.029)	**0.704** (.033)	**0.064** (.005)
TERM ($t = 0.1$)	0.853 (.027)	**0.707** (.009)	**0.061** (.003)

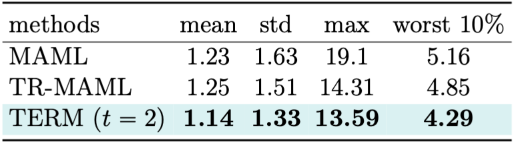
Fair Federated Learning



Fair Meta Learning

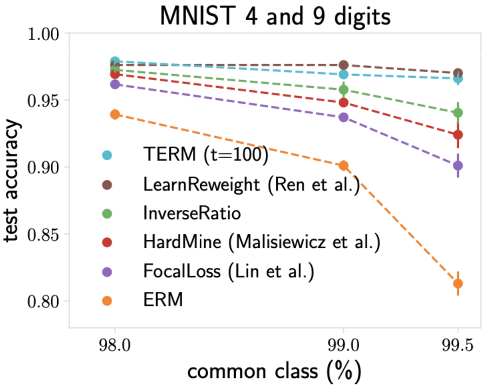


Fair Meta Learning

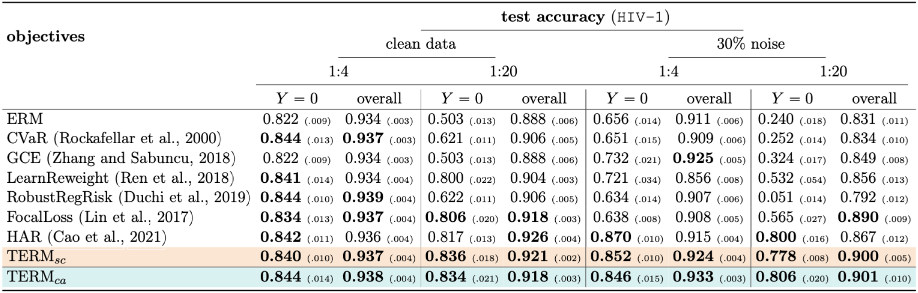


methods	mean	std	max	worst 10%
MAML	1.23	1.63	19.1	5.16
TR-MAML	1.25	1.51	14.31	4.85
TERM ($t = 2$)	**1.14**	**1.33**	**13.59**	**4.29**

Handling Class Imbalance



Class Imbalance and Random Noise



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

Contents

1. Introduction
   - perspectives on exponential tilting
   - Motivation
2. TERM : Properties and Interpretations
   - Assumptions
   - General properties of TERM
   - Interpolation 1
   - Interpolation 2
3. Connection to other risk measures
   - TERM and Renyi cross entropy
   - TERM as Regularizer
4. Solving TERM
5. Hierarchical Multi-Objective Tilting
6. TERM in Practice: Use Cases
7. **Related Approaches in Machine Learning**
8. Discussion and Conclusion

Related Approaches

- Alternate aggregation schemes
- Alternate loss functions
- Sample re-weighting schemes

Contents

1. Introduction
   - perspectives on exponential tilting
   - Motivation
2. TERM : Properties and Interpretations
   - Assumptions
   - General properties of TERM
   - Interpolation 1
   - Interpolation 2
3. Connection to other risk measures
   - TERM and Renyi cross entropy
   - TERM as Regularizer
4. Solving TERM
5. Hierarchical Multi-Objective Tilting
6. TERM in Practice: Use Cases
7. Related Approaches in Machine Learning
8. Discussion and Conclusion

Future Works

- Use TERM in semi-supervised learning
- Define experiment with pseudo label and TERM
- Read more about other loss functions

Refrences

- On Tilted Losses in Machine Learning: Theory and Applications: Tian Li Ahmad Beirami Maziar Sanjabi Virginia Smith