Elliptic Loss Regularization

Ali Hasan [1,2] Haoming Yang [1] Yuting Ng [1] Vahid Tarokh [1]

[1]Duke University [2]Morgan Stanley

Motivating Questions

1. Can we anticipate the behavior of the loss within regions of low training data density?
2. Can we enforce this using a scalable computational solution that applies to general function approximators?

Empirical Risk Minimization (ERM) does not explicitly regularize model behavior in regions of low training data density. A common way to transformed risk minimization (TRM) with $\mathcal{T}$ being randomly sampled transformation parameterized by ϕ

$$\min_\theta \mathbb{E}_{\delta_{\mathcal{X}\times\mathcal{Y}}}[\ell(f_\theta(X), y)] \tag{ERM}$$

$$\min_\theta \mathbb{E}_\phi \mathbb{E}_{\delta_{\mathcal{X}\times\mathcal{Y}}}[\ell(f_\theta(\mathcal{T}_X^\phi X), \mathcal{T}_y^\phi y)] \tag{TRM}$$

We propose a particular class of transformations to enforce the loss land scape to solve a partial differential equation (PDE) with useful properties.

Our Contributions

1. We describe a new regularization scheme that enforces the model loss as a function of inputs to satisfy an elliptic PDE;
2. We theoretically characterize the practical properties of this regularization through PDE theory;
3. We introduce an efficient computational approach that endows the properties of the elliptic regularization.

Why Elliptic PDEs?

PDEs characterize the joint rates in change of different variables. We apply PDE to study how the loss changes as a function of perturbations in the input space, this:

- provides insights into the **robustness** of the model for regions outside of the training data;
- provides a correspondence between the model loss as a function of inputs and qualitative properties of elliptic PDEs to **bound the loss**.

We define the *loss landscape* as a function $u(X, y) : \mathcal{D} \to \mathbb{R}_+$. Our goal is to prescribe the function $u(X, y)$ with a specified level of regularity over the data space in a way that also imbues $\ell(f_\theta(X), y)$ with regularity and thereby obtain desirable properties.

$$\min_\theta u(X, y), \quad (X, y) \in \mathcal{D}$$

$$0 = \sigma\nabla^2 u(X, y), \quad (X, y) \in \mathcal{D} \tag{1}$$

$$u(X, y) = \ell(f_\theta(X), y), \quad (X, y) \in \partial\mathcal{D} \tag{2}$$

A Scalable Solution

PDE in (1) can be solved by the following expectation with a by applying the Feynman-Kac formula, which requires sampling a stopping time τ:

$$u(x, y) = \min_\theta \mathbb{E}[\ell(f_\theta(x_\tau), y_\tau) | x_0 = x, y_0 = y] \tag{3}$$

To avoid infinite stopping times, we instead take expectations over Brownian bridges connecting points on the boundary:

$$\min_\theta \mathbb{E}_{(X_\rightharpoonup, y_\rightharpoonup)\times(X_\leftharpoondown, y_\leftharpoondown)\sim P(\Pi)} \mathbb{E}_{X_s, y_s \sim \mathrm{BB}^{X_\leftharpoondown, y_\leftharpoondown}_{X_\rightharpoonup, y_\rightharpoonup}} \left[\int_0^1 \ell(f_\theta(X_s), y_s) \mathrm{d}s \right] \tag{4}$$

for all s, where we denote $\mathrm{BB}^{X_\leftharpoondown, y_\leftharpoondown}_{X_\rightharpoonup, y_\rightharpoonup}$ as a Brownian bridge sample path where $\rightharpoonup$ denotes starting points and $\leftharpoondown$ denotes end points and Π is the set of points in the support of $\delta_{(\mathcal{X}\times\mathcal{Y})} \times \delta_{(\mathcal{X}\times\mathcal{Y})}$.

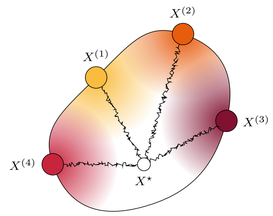


Figure 1. Illustration of the loss values over a domain with 4 points on the boundary. The expected loss at point $X^\star$ is composed of losses at $\varepsilon-$balls around $X^{(i)}, i = 1 \ldots 4$. Black paths represent sample paths starting at $X^\star$.

The Maximum Principle

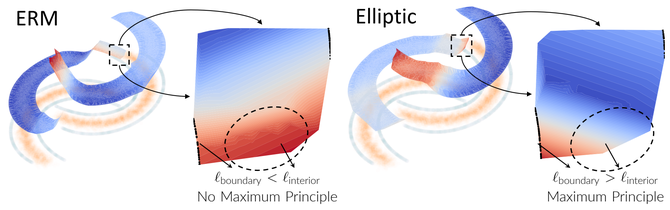
Once the optimization is solved, the loss landscape $u(x, y)$ should follows:

The Maximum Principle Consider any point $X, y \in \mathcal{D}$, suppose the function pairs u, f_θ solves (1). Then, the expected loss u at X, y satisfies the inequality:

$$\min_{X,y\in\mathcal{X}\times\mathcal{Y}} \ell(f_\theta(X), y) \leq u(X, y) \leq \max_{X,y\in\mathcal{X}\times\mathcal{Y}} \ell(f_\theta(X), y).$$

To illustrate this (Figure 2) consider two-moon classifiers, one trained with ERM and another with elliptic regularization, using only the blue scattered boundary data and evaluate on the red interior data to plot the loss surface.

Zooming in on the interior of the ERM loss surface, the training loss exceeds the loss of boundary (circled area) whereas elliptic regularization bounds the loss, as the theory predicts.



Empirical Evaluation

We apply elliptic regularization on classic group-imbalance, data-shift, and spurious correlation datasets and compare the results for classification:

Algorithm	WaterBirds		CelebA		Camelyon17
	Avg(%)	Worst(%)	Avg(%)	Worst(%)	Avg(%)
CVaR-DRO [Levy et al., 2020]	$90.3_{\pm 1.2}$	$77.2_{\pm 2.2}$	$86.8_{\pm 0.7}$	$76.9_{\pm 3.1}$	$70.5_{\pm 5.1}$
CVaR-DORO [Zhai et al., 2021]	$91.5_{\pm 0.7}$	$77.0_{\pm 2.8}$	$89.6_{\pm 0.4}$	$75.6_{\pm 4.2}$	$67.3_{\pm 7.2}$
χ^2-DRO [Levy et al., 2020]	$88.3_{\pm 1.5}$	$74.0_{\pm 1.8}$	$87.7_{\pm 0.3}$	$\underline{78.4}_{\pm 3.4}$	$68.0_{\pm 6.7}$
χ^2-DORO [Zhai et al., 2021]	$89.5_{\pm 1.0}$	$76.0_{\pm 3.1}$	$87.0_{\pm 0.6}$	$75.6_{\pm 3.4}$	$68.0_{\pm 7.5}$
Elliptic + IW	$\underline{92.0}_{\pm 0.3}$	$\underline{84.1}_{\pm 1.1}$	$\underline{\mathbf{91.3}}_{\pm 0.3}$	$77.4_{\pm 4.5}$	$\underline{\mathbf{77.9}}_{\pm 3.0}$
Two-stage: JTT [Liu et al., 2021]	$\mathbf{93.6}_{\pm \text{NA}}$	$86.0_{\pm \text{NA}}$	$88.0_{\pm \text{NA}}$	$81.1_{\pm \text{NA}}$	$69.1_{\pm 6.4}$
Two-stage: UMIX [Han et al., 2022]	$93.0_{\pm 0.5}$	$\mathbf{90.0}_{\pm 1.1}$	$90.1_{\pm 0.4}$	$\mathbf{85.3}_{\pm 4.1}$	$75.1_{\pm 5.9}$

for robust regression under (sub)domain shift and spurious correlation:

Algorithm	SkillCraft		Crime		RCF-MNIST	
	Avg	Worst	Avg	Worst	Avg	Worst
C-mixup [Yao et al., 2022]	$6.27_{\pm 0.537}$	$\mathbf{8.83}_{\pm 1.010}$	$0.132_{\pm 0.003}$	$0.167_{\pm 0.010}$	$0.165_{\pm 0.001}$	$0.180_{\pm 0.001}$
Elliptic	$\mathbf{5.97}_{\pm 0.283}$	$9.17_{\pm 1.150}$	$0.132_{\pm 0.003}$	$\mathbf{0.164}_{\pm 0.01}$	$\mathbf{0.162}_{\pm 0.002}$	$\mathbf{0.178}_{\pm 0.002}$

and with scenarios with additional 50% label corruption:

	Breast		Blood		Path		OrganC	
Method	Avg(%)	Worst(%)	Avg(%)	Worst(%)	Avg(%)	Worst(%)	Avg(%)	Worst(%)
ERM	$80.0_{\pm 3.4}$	$32.4_{\pm 13.8}$	$81.4_{\pm 2.7}$	$56.3_{\pm 17.3}$	$55.7_{\pm 3.8}$	$10.7_{\pm 8.6}$	$84.1_{\pm 1.9}$	$65.1_{\pm 8.2}$
mixup	$83.1_{\pm 2.4}$	$47.1_{\pm 10.9}$	$78.8_{\pm 3.1}$	$41.7_{\pm 19.3}$	$54.3_{\pm 2.1}$	$1.2_{\pm 2.1}$	$85.6_{\pm 2.4}$	$56.8_{\pm 11.0}$
mixupE	$76.5_{\pm 2.8}$	$16.7_{\pm 11.3}$	$74.2_{\pm 3.6}$	$31.9_{\pm 16.9}$	$52.5_{\pm 2.0}$	$2.7_{\pm 7.6}$	$70.4_{\pm 7.6}$	$35.6_{\pm 16.8}$
Elliptic	$\mathbf{87.6}_{\pm 1.8}$	$\underline{67.6}_{\pm 4.4}$	$\mathbf{85.5}_{\pm 1.6}$	$\underline{60.2}_{\pm 14.9}$	$\underline{62.3}_{\pm 2.5}$	$\mathbf{26.4}_{\pm 10.2}$	$\underline{87.7}_{\pm 0.9}$	$\underline{65.9}_{\pm 3.4}$
+ IW	$\underline{87.3}_{\pm 0.9}$	$\mathbf{67.6}_{\pm 3.0}$	$\underline{84.1}_{\pm 4.0}$	$\mathbf{68.0}_{\pm 13.7}$	$\mathbf{62.7}_{\pm 1.8}$	$\underline{20.9}_{\pm 6.8}$	$\mathbf{88.0}_{\pm 0.7}$	$\mathbf{66.2}_{\pm 4.8}$

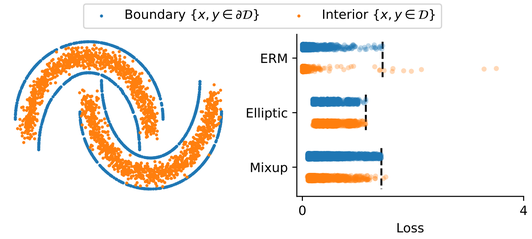


Figure 2. Maximum principle satisfied for the elliptic regularization when trained on boundary data.

References

Z. Han, Z. Liang, F. Yang, L. Liu, L. Li, Y. Bian, P. Zhao, B. Wu, C. Zhang, and J. Yao. Umix: Improving importance weighting for subpopulation shift via uncertainty-aware mixup. *Advances in Neural Information Processing Systems*, 35:37704–37718, 2022.

D. Levy, Y. Carmon, J. C. Duchi, and A. Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33:8847–8860, 2020.

E. Z. Liu, B. Haghgoo, A. S. Chen, A. Raghunathan, P. W. Koh, S. Sagawa, P. Liang, and C. Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021.

H. Yao, Y. Wang, L. Zhang, J. Y. Zou, and C. Finn. C-mixup: Improving generalization in regression. *Advances in Neural Information Processing Systems*, 35: 3361–3376, 2022.

R. Zhai, C. Dan, Z. Kolter, and P. Ravikumar. Doro: Distributional and outlier robust optimization. In *International Conference on Machine Learning*, pages 12345–12355. PMLR, 2021.

ICLR 2025

This work was supported in part by the Air Force Office of Scientific Research under award number FA9550-22-1-0315.

ali.hasan@duke.edu