UNLABELED OUT-OF-DOMAIN DATA IMPROVES GENERALIZATION

Amir Hossein Saberi Amir Najafi Alireza Heidari
Mohammad Hosein Movasaghinia Seyed Abolfazl Motahari Babak H. Khalaj
Sharif University of Technology

ICLR

MOTIVATION

The sample complexity for a linear binary classifier in a non-realizable setting in $\mathbb{R}^d$ is $O(d/\epsilon^2)$. It has been shown that out-of-domain unlabeled data can enhance generalization, but explicit bounds for the size of labeled and unlabeled data remain incomplete.

CONTRIBUTION

We propose a polynomial-time framework that leverages both labeled and slightly out-of-domain unlabeled data. Our framework guarantees improved generalization under the *cluster assumption* of the true data distribution. In the well-studied setting of the two-component Gaussian Mixture Model (GMM) for classification, with m labeled and n unlabeled data points, our theoretical findings demonstrate:

- **Non-asymptotic bounds** for both robust and non-robust learning.
- **Enhanced generalization** over ERM techniques when $n \geq \Omega(m^2/d)$.
- **Dimension-independent** sample complexity under well-defined conditions.
- **Improved sample complexity** from $O(d/\epsilon^2)$ to $O(d/\epsilon)$ when $n = O(d/\epsilon^6)$.

Preliminaries

Definition A.1 (Wasserstein Distance)**.** Given P and Q supported on $\mathcal{X}$, and a non-negative, lower semi-continuous cost function $c : \mathcal{X}^2 \to \mathbb{R}_+$ satisfying $c(\boldsymbol{X}, \boldsymbol{X}) = 0$ for all $\boldsymbol{X} \in \mathcal{X}$, is defined as:

$$\mathcal{W}_c(P, Q) = \inf_{\mu \in \Gamma(\mathcal{X}^2)} \mathbb{E}_{\boldsymbol{X}, \boldsymbol{X}' \sim \mu} \left[c(\boldsymbol{X}, \boldsymbol{X}') \right] \quad \text{s.t.} \quad \mu(\boldsymbol{X}, \cdot) = P \,,\ \mu(\cdot, \boldsymbol{X}') = Q.$$

Definition A.2 (ϵ-neighborhood of a Distribution P)**.**

$$\mathcal{B}_\epsilon^c(P) = \{Q : \mathcal{W}_c(P, Q) \leq \epsilon\}. \tag{1}$$

Definition 1.1 (Distributionally Robust Learning(DRL)). In DRL, the *learner* attempts to find a classifier with a small robust risk, denoted as $R^{\text{robust}}(\theta, P)$, where:

$$R_{\epsilon,c}^{\text{robust}}(\theta, P) = \sup_{P' \in \mathcal{B}_\epsilon^c(P)} R(\theta, P');\ \forall \theta \in \Theta \ \&\ \epsilon \geq 0. \tag{2}$$

Therefore, DRL solves the following minimax optimization problem:

$$\hat{\theta}_{\epsilon,c}^{\text{DRL}}(S) \triangleq_{\theta \in \Theta} R_{\epsilon,c}^{\text{robust}}\left(\theta, \hat{P}_S^m\right). \tag{3}$$

PROBLEM SETUP

Notations & Definitions:

- Let $\mathcal{X} \subseteq \mathbb{R}^d$, $y \in \{\pm 1\}$. The joint feature-label distribution of labeled data is denoted as $P_0 \sim \mathcal{X} \times \{\pm 1\}$.
- P_1 is a shifted version of P_0, where their marginal distributions on $\mathcal{X}$ are shifted by $\mathcal{W}_c(P_{0,X}, P_{1,X}) = \alpha$ with $\alpha > 0$, and no assumptions are made on $P_1(y \mid \boldsymbol{X})$.
- **Labeled Data**: $S_0 = \{(\boldsymbol{X}_i, y_i)\}_{i=1}^m \sim P_0^m$, where $P_0(y = 1) = \frac{1}{2}$.
- **Out-of-domain Unlabeled Data**: $S_1 = \{\boldsymbol{X}_i'\}_{i=1}^n \sim P_{1,X}^n$.

Isotropic GMM Setup:

- **Loss Function:** Without loss of generality, let $\ell(\boldsymbol{X}, y; \theta) = \mathbf{1}(y\langle \theta, \boldsymbol{X} \rangle \geq 0)$.
- $P_0(\boldsymbol{X} \mid y) = \mathcal{N}(y\boldsymbol{\mu}_0, \sigma_0^2)$ for some $\sigma_0 \geq 0$, $\boldsymbol{\mu}_0 \in \mathbb{R}^d$.
- $P_{1,X} = \frac{1}{2}\sum_{u=\pm 1} \mathcal{N}(u\boldsymbol{\mu}_1, \sigma_1^2)$,

where $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| \leq \mathcal{O}(\alpha)$ and $|\sigma_1 - \sigma_0| \leq \mathcal{O}(\alpha)$ since $\mathcal{W}_c(P_{0,X}, P_{1,X}) = \alpha$.

Framework: Robust Self-Supervised Training

Robust Self-Supervised (RSS) Training: Consider a cost function c and a parameter $\gamma \geq 0$. We define the *robust loss* $\phi_\gamma : \mathcal{X} \times \{\pm 1\} \times \Theta \to \mathbb{R}$ by:

$$\phi_\gamma(\boldsymbol{X}, y; \theta) \triangleq \sup_{\boldsymbol{Z} \in \mathcal{X}} \ell(\boldsymbol{Z}, y; \theta) - \gamma c(\boldsymbol{Z}, \boldsymbol{X}). \tag{4}$$

For a given set of parameters $\gamma, \gamma', \lambda \in \mathbb{R}_{\geq 0}$: $\hat{\theta}^{\text{RSS}}$ estimator is defined as:

$$\hat{\theta}^{\text{RSS}} \triangleq_{\theta \in \Theta} \left\{ \frac{1}{m}\sum_{i=1}^m \phi_\gamma(\boldsymbol{X}_i, y_i; \theta) + \frac{\lambda}{n}\sum_{j=1}^n \phi_{\gamma'}\left(\boldsymbol{X}_j', h_\theta(\boldsymbol{X}_j'); \theta\right) \right\}. \tag{5}$$

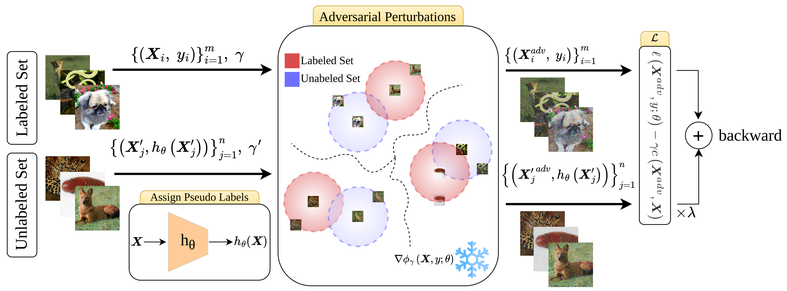

Figure 1: Overview of Robust Self-Supervised (RSS) Training

Generalization Bounds (GMM)

Theorem 4.1 (Non-asymptotic Bound for Robust Learning)**.**

$$\mathbb{E}_{P_0}\left[\phi_\gamma\left(\boldsymbol{X}, y; \hat{\theta}^{\text{RSS}}\right)\right] \leq \min_{\theta \in \Theta} \mathbb{E}_{P_0}\left[\phi_\gamma(\boldsymbol{X}, y; \theta)\right] + \mathcal{O}\left(\gamma\sqrt{\frac{2d}{m}\left(\alpha(\|\boldsymbol{\mu}_0\|_2^2 + \sigma_0^2) + \sqrt{\frac{2d}{2n+m}} + \sqrt{\frac{2\log(1/\delta)}{2n+m}}\right)} + \sqrt{\frac{2\log(1/\delta)}{m}}\right).$$

Theorem 4.2 (Non-asymptotic Bound for Non-robust Learning)**.**

$$R\left(\hat{\theta}^{\text{RSS}}, P\right) - \min_{\theta \in \Theta} R(\theta, P) \leq \mathcal{O}\left(\frac{e^{\frac{-\|\boldsymbol{\mu}_0\|_2^2}{4\sigma_0^2}}}{\sqrt{2\sigma_0\sqrt{2\pi}}}\left(\left(\|\boldsymbol{\mu}_1\|_2^2 + \sigma_1^2\right)\frac{2d\alpha}{m} + \frac{4d}{m}\sqrt{\frac{2d + 2\log\frac{1}{\delta}}{2n+m}}\right)^{1/4} + \sqrt{\frac{2\log\frac{1}{\delta}}{m}}\right).$$

Corollary (Following Theorem 4.2):

- The $\hat{\theta}^{\text{RSS}}$ estimator outperforms ERM when $\alpha \leq O\left(\frac{d}{m}\right)$, $n \geq \Omega\left(\frac{m^2}{d}\right)$.
- Sample complexity becomes independent of d if: $\alpha \leq O(d^{-1})$, $n \geq \Omega(d^3)$.
- When $\alpha = 0$ (no perturbation), $m = O(\frac{d}{\epsilon})$, and $n = O(\frac{d}{\epsilon^6})$, the generalization bound improves compared to having access to $m = O(\frac{d}{\epsilon^2})$ labeled data points.

Experiments & Results

Simulated Isotropic GMM Data (non-robust setup):

- **Data:** $d = 200$, $\boldsymbol{\mu_1} = \boldsymbol{\mu_0} + \alpha.\boldsymbol{v}$ where $\boldsymbol{v}$ and $\boldsymbol{\mu_0}$ are random normalized vectors.
- **Model:** $h(x; w) = w^T x$, with $l_\gamma(Y, X, f(x)) = \sum_{i=1}^{|X|} \min(1, \max(0, 1 - \gamma.Y_i.f(X_i)))$.

Same Distribution ($\alpha = 0$)				Different Distribution ($\alpha = 0.5 \cdot \|\|\boldsymbol{\mu}_0\|\|_2$)			
m	Acc	n	Acc	m	Acc	n	Acc
10	0.59	10 100 1,000 10,000	0.63 0.66 0.79 **0.82**	10	0.59	10 100 1,000 10,000	0.61 0.65 0.78 **0.81**
20	0.62	20 200 2,000 10,000	0.64 0.69 0.80 **0.82**	20	0.62	20 200 2,000 10,000	0.65 0.65 0.79 **0.80**
40	0.65	40 400 4,000 10,000	0.65 0.71 0.81 **0.82**	40	0.65	40 400 4,000 10,000	0.65 0.73 0.78 **0.80**
10,000	**0.83**	-	-	10,000	**0.83**	-	-

Histopathology Images (robust setup):

- **Data:** NCT-CRC-HE-100K (colorectal cancer) & PatchCamelyon (lymph node).

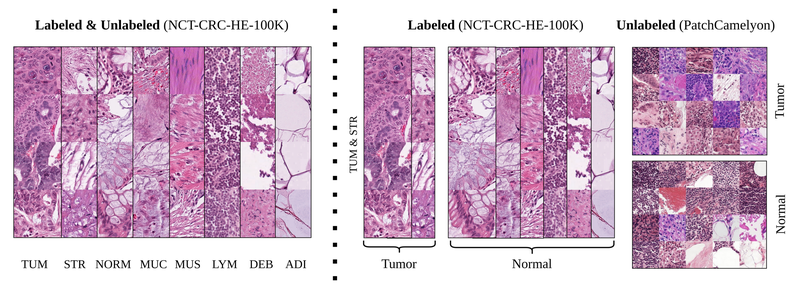

Figure 2: (**Left**) *Same Distribution Setting*. (**Right**) *Different Distribution Setting*.

Same distribution				Different distribution			
Labeled size	Acc	Unlabeled size	Acc	Labeled size	Acc	Unlabeled size	Acc
48	0.65	200 700 2,000	0.71 0.80 **0.82**	25	0.78	100 400 2,000	0.78 0.79 **0.81**
240	0.77	500 1,200 4,000	0.78 0.82 **0.83**	50	0.82	200 700 3,000	0.82 0.86 **0.87**
1040	0.83	3,000 10,000 20,000	0.87 0.89 **0.91**	300	0.87	600 2,000 8,000	0.88 0.89 **0.90**
50,000	**0.916**	-	-	32,000	**0.94**	-	-

Project Github:
Code & Dataset & Model