Out-Of-Domain Unlabeled Data Improves Generalization

Amir Hossein Saberi [‡*]
Amir Najafi [†]
Alireza Heidari [†]
Mohammad Hosein Movasaghinia [†]
Seyed Abolfazl Motahari [†1]
Babak H. Khalaj [‡*^]

[*] Department of Electrical Engineering,
[†] Department of Computer Engineering,
[‡]Sharif Center for Information Systems and Data Science,
^ Institute for Research in Fundamental Sciences, Tehran, Iran
Sharif University of Technology, Tehran, Iran

Abstract

We propose a novel framework for incorporating unlabeled data into semi-supervised classification problems, where scenarios involving the minimization of either i) adversarially robust or ii) non-robust loss functions have been considered. Notably, we allow the unlabeled samples to deviate slightly (in total variation sense) from the in-domain distribution. The core idea behind our framework is to combine Distributionally Robust Optimization (DRO) with self-supervised training. As a result, we also leverage efficient polynomial-time algorithms for the training stage. From a theoretical standpoint, we apply our framework on the classification problem of a mixture of two Gaussians in $\mathbb{R}^d$, where in addition to the m independent and labeled samples from the true distribution, a set of n (usually with $n >> m$) out of domain and unlabeled samples are given as well. Using only the labeled data, it is known that the generalization error can be bounded by $\propto (d/m)^{1/2}$. However, using our method on both isotropic and non-isotropic Gaussian mixture models, one can derive a new set of analytically explicit and non-asymptotic bounds which show substantial improvement on the generalization error compared to ERM. Our results underscore two significant insights: 1) out-of-domain samples, even when unlabeled, can be harnessed to narrow the generalization gap, provided that the true data distribution adheres to a form of the "cluster assumption", and 2) the semi-supervised learning paradigm can be regarded as a special case of our framework when there are no distributional shifts. We validate our claims through experiments conducted on a variety of synthetic and real-world datasets.

Keywords: Learning Theory, Robust Learning, Rademacher Complexity, Out of Domain Generalization, Semi-Supervised Learning

[1] Corresponding author: motahari@sharif.edu.

1 Introduction

Semi-supervised learning has long been a focal point in the machine learning literature, primarily due to the cost-effectiveness of utilizing unlabeled data compared to labeled counterparts. However, unlabeled data in various domains, such as medicine, genetics, imaging, and audio processing, often originates from diverse sources and technologies, leading to distributional differences between labeled and unlabeled samples. Concurrently, the development of robust classifiers against adversarial attacks has emerged as a vibrant research area, driven by the rise of large-scale neural networks [1,2]. While the primary objective of these methods is to reduce model sensitivity to minor adversarial perturbations, recent observations suggest that enhancing adversarial robustness may also improve the utilization of unlabeled samples [3,4].

This paper aims to demonstrate the efficacy of incorporating out-of-domain unlabeled samples to decrease the reliance on labeled in-domain data. To achieve this, we propose a novel framework inspired by a fusion of concepts from adversarial robustness and self-training. Specifically, we introduce a unique constraint to the conventional Empirical Risk Minimization (ERM) procedure, focusing exclusively on the unlabeled part of the dataset. Our theoretical and experimental analyses show that the inclusion of unlabeled data reduces the generalization gap for both robust and non-robust loss functions. Importantly, our alternative optimization criteria are computationally efficient and can be solved in polynomial time. We have implemented and validated the effectiveness of our method on various synthetic and real-world datasets.

From a theoretical standpoint, akin to prior research [5–8], we also address the binary classification problem involving two Gaussian models in $\mathbb{R}^d$. This problem has been the center of attention in several recent works on theoretical analysis of both semi-supervised and/or adversarially robust learning paradigms. Despite several recent theoretical investigations, the precise trade-off between the sizes of labeled (m) and unlabeled (n) data, even in this specific case, remains incomplete. A number of works have bounded the labeled sample complexity under the assumption of an asymptotically large n [9], while another series of papers have analyzed this task from a completely unsupervised viewpoint. We endeavor to fill this gap by providing the first empirical trade-off between m and n, even when unlabeled data originates from a slightly perturbed distribution. We derive explicit bounds for both robust and non-robust losses of linear classifiers in this scenario. Our results show that as long as $n \geq \Omega\left(m^2 / d\right)$, our proposed algorithm surpasses traditional techniques that solely rely on labeled data. We also consider the more general case of non-isotropic Gaussian models, as explored in previous studies.

This paper is an extended and refined version of our conference paper accepted at ICLR 2024. In this version, we have improved the algorithm for greater efficiency, added several proofs that were omitted from the conference version, and enhanced the overall rigor of the theoretical analysis.

The remainder of this paper is structured as follows: Section 1.1 provides an overview of related works in distributionally robust optimization and semi-supervised learning. Section 1.3 introduces our notation and definitions. In Section 1.2, we discuss the contributions

made by our work. In Section 3, we present our novel method, followed by a theoretical analysis in Section 4. Section 5 showcases our experimental validations, further supporting our theoretical findings. Finally, we draw conclusions in Section 6.

1.1 Prior Works

One of the challenges in adversarially robust learning is the substantial difficulty in increasing the *robust* accuracy compared to achieving high accuracy in non-robust scenarios [10]. A study by [5] posited that this challenge arises from the larger sample complexity associated with learning robust classifiers in general. Specifically, they presented a simple model where a good classifier with high standard (non-robust) accuracy can be achieved using only a single sample, while a significantly larger training set is needed to attain a classifier with high robust accuracy. Recent works [6–8] demonstrated that the gap in sample complexity between robust and standard learning, as outlined by [5] in the context of a two-component Gaussian mixture model, can be bridged with the inclusion of unlabeled samples. Essentially, unlabeled samples can be harnessed to mitigate classification errors even when test samples are perturbed by an adversary. Another study by [3] achieved a similar result using a different definition of adversarial robustness and a more comprehensive data generation model. Their approach involves the use of 'self-training' to assign soft/hard labels to unlabeled data, contrasting our approach, where unlabeled data is exclusively utilized to constrain the set of classifiers, aiming to avoid crowded regions. While DRO serves as a tool in our approach, it is not necessarily the primary objective. In [11], authors showed that in the setting of [5], out-of-domain unlabeled samples improve adversarial robustness.

Theoretical analysis of Semi-Supervised Learning (SSL) under the so-called cluster assumption has been a long-studied task [12]. However, beyond [3], several recent methods leveraging DRO for semi-supervised learning have emerged [13,14]. Notably, [14] shares similarities with [3]; however, instead of assigning artificial labels to unlabeled samples, [14] employs them to delimit the ambiguity set and enhance understanding of the marginals. Our work primarily focuses on the robustness aspect of the problem rather than advancing the general SSL paradigm.

Defense mechanisms against adversarial attacks usually consider two types of adversaries: i) point-wise attacks similar to [4,15,16], and ii) distributional attacks [17–19], where in the case of the latter adversary can change the distribution of data up to a predefined budget. It has been shown that Distributionally Robust Learning (DRL) achieves a superior robustness compared to point-wise methods [17]. [20] utilized DRL in order to achieve a balance between the bias and variance of classifier's error, leading to faster rates of convergence compared to empirical risk minimization even in the *non-robust* case. In DRL, the learner typically aims to minimize the loss while allowing the data distribution to vary within an uncertainty neighborhood. The central idea used by [20] was to regulate the diameter of this uncertainty neighborhood based on the number of samples. [21] achieved similar results in DRL while utilizing the Wasserstein metric to define the perturbation budget for data distribution. Based on the above arguments, we have also utilized DRL is the main tool in developing our proposed framework.

Over the past decade, a significant body of research has theoretically characterized the Domain Adaptation (DA) problem, progressively elucidating the conditions under which a model can generalize from a labeled source distribution to an unlabeled target distribution. A central theme in this literature is the derivation of generalization bounds, where the target risk is upper-bounded by the sum of the source error and a divergence term that measures the discrepancy between domains using unlabeled samples. PAC-Bayesian frameworks have formalized this by bounding the target risk of stochastic or majority-vote classifiers, emphasizing the trade-off between empirical performance and data-dependent measures of domain divergence [22,23]. These bounds have been further refined to accommodate multiclass adaptation, where non-uniform complexity effects significantly impact generalization [24].

In parallel, Optimal Transport (OT) theory has been integrated into DA to replace generic distribution discrepancies with Wasserstein metrics. These OT-based analyses leverage the underlying geometry of the data, often yielding tighter generalization guarantees for structured shifts [25,26]. Complementing these approaches, information-theoretic perspectives on Unlabeled Domain Adaptation (UDA) decompose the generalization gap into distinct components, offering insights into algorithm design that go beyond simple discrepancy matching [27].

A second major research trajectory investigates the sufficiency of representation alignment and identifies the conditions under which it provably fails. Several studies demonstrate that minimizing source error and marginal feature alignment is insufficient for guaranteeing low target error in the presence of conditional shift (where the relationship between features and labels changes). These works provide refined upper bounds that explicitly account for label-conditional mismatches [28]. Related research addresses label shift, where label distributions differ across domains, proposing principled reweighting mechanisms to maintain theoretical guarantees [29].

The scope of DA theory has recently expanded to address more complex scenarios, such as multi-source transfer, characterizing various notions of invariance and the inherent trade-offs between them [30]. Conversely, the fundamental limitations of UDA have been explored through lower-bound analyses, which formalize the conditions leading to negative transfer and identify when adaptation is theoretically impossible [31]. To mitigate these issues, Gradual Domain Adaptation (GDA) utilizes intermediate distributions to provide improved guarantees; theoretical analyses suggest that placing these distributions along Wasserstein geodesics minimizes error propagation during the transition [32]. Finally, recent work has begun to bridge the gap between DA and adversarial robustness, deriving generalization bounds for robust UDA to quantify the performance cost of maintaining robustness under distribution shift [33].

An important paper that explores the relationship between accuracy and robustness is the so-called TRADES paper [34]. TRADES formalizes adversarial robustness as a trade-off between clean accuracy and the susceptibility of data points to lie near a decision boundary. Specifically, it decomposes the robust classification error into a *natural error* term (standard misclassification on clean inputs) and a boundary error term, which captures how easily small perturbations can flip predictions. This decomposition explains why improving robustness can sometimes reduce clean accuracy. Although TRADES is

related to our work in its use of adversarial training, its objective differs from that of our paper. The primary goal of TRADES is to characterize the trade-off between robustness and accuracy, whereas our aim is not to learn a robust model or to quantify robustness. Instead, we employ adversarial training as a tool to improve accuracy on in-domain samples. Therefore, a direct comparison between the two methods is not appropriate.

To the best of our knowledge, the specific problem setting analyzed theoretically in this paper has not been studied prior to our work. While some existing works share partial similarities, they differ from ours in important respects: either in the type of samples considered (e.g., OOD samples are labeled [35]) or in the problem formulation itself, where no explicit generalization bounds were derived [36].

1.2 Main Contributions

We introduce a novel integration of DRO and Semi-Supervised Learning (SSL), leveraging out-of-domain unlabeled samples to enhance the generalization bound of learning problem. Specifically, we theoretically analyze our method in the setting where samples are generated from a Gaussian mixture model with two components, which is a common assumption in several theoretical analyses in this field. For example, a simpler format, when two Gaussians are isotropic and well-separated, is the sole focus of many papers such as [5–7]. Some of our notable contributions and improvements over recent works in the field include:

(i) In Theorem 6, we present a non-asymptotic bound for adversarially robust learning, leveraging both labeled and unlabeled samples jointly. This result builds upon the work of [6] and [8], which focused on the effectiveness of unlabeled samples when a single labeled sample is sufficient for linear classification of a non-robust classifier. However, these studies do not provide insights into the necessary number of unlabeled samples when multiple labeled samples are involved, particularly in scenarios where the underlying distribution exhibits limited separation between the two classes. Our theoretical bounds address and fill this crucial gap.

(ii) Theorem 7 introduces a novel non-asymptotic bound for integrating labeled and unlabeled samples in SSL. To underscore the significance of our findings, consider the following example: In the realizable setting, where positive and negative samples can be completely separated by a hyperplane in $\mathbb{R}^d$, the sample complexity of supervised learning for a linear binary classifier with at most ϵ true error is known to be $\mathcal{O}(d/\epsilon)$ [37]. However, in the non-realizable setting, this complexity escalates to $\mathcal{O}(d/\epsilon^2)$ [37]. A pivotal question in learning theory revolves around how to approach the sample complexity of $\mathcal{O}(d/\epsilon)$ in the non-realizable setting. Insights provided by [20] delve into this inquiry. Notably, even with the awareness that the underlying distribution is a Gaussian mixture, the optimal sample complexity, as per [38], still exceeds $\mathcal{O}(d/\epsilon^2)$. Our work demonstrates that in scenarios where the underlying distribution is a Gaussian mixture and we possess $m=\mathcal{O}(d/\epsilon)$ labeled samples, coupled with $n=\mathcal{O}\left(\frac{d}{\epsilon^6}\right)$

unlabeled samples (without knowledge of the underlying distribution), one can achieve an error rate lower than or equal to the case of having access to $\mathcal{O}\left(d/\epsilon^2\right)$. labeled samples.

(iii) We formalize the incorporation of *out-of-domain* unlabeled samples into the generalization bounds of both robust and non-robust classifiers in Theorems 6, 7 and 9. We contend that this represents a novel contribution to the field, with its closest counterpart being [11]. Notably, [11] addresses a scenario where the underlying distribution is an isotropic Gaussian mixture with well-separated Gaussian components, while the separation of components is not a prerequisite for our results.

1.3 Notation and Definitions

Let us denote the feature space by $\mathcal{X}\subseteq\mathbb{R}^d$, and assume $\mathcal{H}$ as a class of binary classifiers parameterized by the parameter set Θ: for each $\theta\in\Theta$, we have a classifier $h_\theta\in\mathcal{H}$ where $h_\theta:\mathcal{X}\to\{-1,1\}$. Assume a positive function $\ell:\left(\mathcal{X}\times\{-1,1\}\times\Theta\right)\to\mathbb{R}_{\geq 0}$ as the loss function. Also, let P be the unknown data distribution over $\mathcal{X}\times\{-1,1\}$, and $S=\left\{\left(\boldsymbol{X}_i,y_i\right)\right\}_{i=1}^m$ r $m\in\mathbb{N}$ be a set of i.i.d. samples drawn from P. Then, for all $\theta\in\Theta$ the true risk R and the empirical risk $\hat{R}$ of a classifier w.r.t. P can be defined as follows:

$$R\left(\theta,P\right)=\mathbb{E}_P\left[\ell\left(\boldsymbol{X},y;\theta\right)\right],R\left(\theta,\hat{P}_S^m\right)=\mathbb{E}_{\hat{P}_S^m}\left[\ell\left(\boldsymbol{X},y;\theta\right)\right]=\frac{1}{m}\sum_{i=1}^m\ell\left(\boldsymbol{X}_i,y_i;\theta\right), \tag{1}$$

where $\hat{P}_S^m$ denotes an empirical estimate of P based on the m samples in S. We also need a way to measure the distance between various distributions that are supported over $\mathcal{X}$. A well-known candidate for this goal is the *Wasserstein* distance (Definition 1). Subsequently, we also define a Wasserstein ball in Definition 2 in order to effectively constrain a set of probability measures. It should be noted that throughout this paper, the Wasserstein distance between any two distributions supported over $\mathcal{X}\times\{\pm 1\}$ is defined as the distance between their respective marginals on $\mathcal{X}$.

Definition 1 (Wasserstein Distance). Consider two probability distributions P and Q supported on $\mathcal{X}$, and assume cost function $c:\mathcal{X}\times\mathcal{X}\to\mathbb{R}_+$ is a non-negative lower semi-continuous function satisfying $c\left(\boldsymbol{X},\boldsymbol{X}\right)=0$. for all $\boldsymbol{X}\in\mathcal{X}$. Then, the Wasserstein distance between P and Q w.r.t. c, denoted as $\mathcal{W}_c\left(P,Q\right)$, is defined as

$$\mathcal{W}_c\left(P,Q\right)=\inf_{\mu\in\Gamma\left(\mathcal{X}^2\right)}\mathbb{E}_{\boldsymbol{X},\boldsymbol{X}'\sim\mu}\left[c\left(\boldsymbol{X},\boldsymbol{X}'\right)\right],\text{subject to }\mu\left(\boldsymbol{X},\cdot\right)=P,\mu\left(\cdot,\boldsymbol{X}'\right)=Q, \tag{2}$$

where $\Gamma\left(\mathcal{X}^2\right)$ denotes the set of all couplings over $\mathcal{X}\times\mathcal{X}$.

Definition 2 (ϵ -neighborhood of a Distribution). The ϵ -neighborhood of a distribution P is defined as the set of all distributions that have a Wasserstein distance less than ϵ from P. Mathematically, it can be represented as:

$$\mathcal{B}_\epsilon^c(P) = \{Q : \mathcal{W}_c(P,Q) \le \epsilon\}. \tag{3}$$

The ultimate goal of classical learning is to find the parameter $\theta^* \in \Theta$ such that with high probability, $R(\theta^*)$ is sufficiently close to $\min_\theta R(\theta)$. A well-known approach to achieve this goal is Empirical Risk Minimization (ERM) algorithm, formally defined as follows:

$$\hat{\theta}^{\mathrm{ERM}}(S) = \operatorname{argmin}_{\theta\in\Theta} \mathbb{E}_{\hat{P}_S^m}\left[\ell(\theta; \boldsymbol{X}, y)\right] = \operatorname{argmin}_{\theta\in\Theta} \frac{1}{m}\sum_{i=1}^{m} \ell(\theta; \boldsymbol{X}_i, y_i). \tag{4}$$

A recent variant of ERM, which has gained huge popularity in both theory and practice, is the so-called Distributionally Robust Learning (DRL) which is formulated as follows:

Definition 3 (Distributionally Robust Learning(DRL)). DRL aims at training a classifier which is robust against adversarial attacks on data distribution. In this regard, the *learner* attempts to find a classifier with a small robust risk, denoted as $R^{\mathrm{robust}}(\theta, P)$, which is defined as

$$R_{\epsilon,c}^{\mathrm{robust}}(\theta, P) = \sup_{P' \in \mathcal{B}_\epsilon^c(P)} R(\theta, P'), \tag{5}$$

for all $\theta \in \Theta$ and any $\epsilon \ge 0$. Therefore, DRL solves the following optimization problem:

$$\hat{\theta}_{\epsilon,c}^{\mathrm{DRL}}(S) = \operatorname{argmin}_{\theta\in\Theta} R_{\epsilon,c}^{\mathrm{robust}}\left(\theta, \hat{P}_S^m\right). \tag{6}$$

Surprisingly, the sophisticated minimax optimization problem of Eq. (6) which takes place in a subset of the infinite-dimensional space of probability measures that corresponds to the constraints, can be substantially simplified when is re-written in the dual format:

Lemma 4 (From Blanchet et al. [39]). For a sufficiently small $\epsilon > 0$, the minimax optimization problem of Eq. (6) has the following dual form:

$$\inf_{\theta\in\Theta} \sup_{P' \in \mathcal{B}_\epsilon^c(\hat{P}_S^m)} R(\theta, P') = \inf_{\gamma \ge 0} \left\{ \gamma\epsilon + \inf_{\theta\in\Theta} \frac{1}{m}\sum_{i=1}^{m} \sup_{\boldsymbol{z}\in\mathcal{X}} \ell(\boldsymbol{Z}, y_i; \theta) - \gamma c(\boldsymbol{Z}, \boldsymbol{X}_i) \right\}, \tag{7}$$

where γ and ϵ are dual parameters, and there is a bijective and reciprocal relation between the ϵ and γ^*, i.e., the optimal value which minimizes the r.h.s.

As suggested by [40], the $\inf_{\gamma\ge 0}$ in the r.h.s. part in the above optimization problem can be removed by fixing a user-defined value for γ. This also means that if one attempts to find the optimal value for θ, the additive term $\gamma\epsilon$ is ineffective and can be removed as well.

It should be noted that this also fixes an (unknown) value for ϵ. In practice, the appropriate value for ϵ is not known beforehand and thus can be usually found through a cross-validation stage, while the same procedure can be applied to its dual counterpart, i.e., γ. In other words, the above-mentioned strategy keeps the generality of the problem intact. For the sake of simplicity in relations, throughout the rest of the paper we work with the dual formulation in Eq. (7) and let γ be a fixed and arbitrary value.

2 Problem Definition

At this point, we can formally define our problem. Let $\mathcal{X} \subseteq \mathbb{R}^d$, and assume P_0 be an unknown and arbitrary distribution supported on $\mathcal{X} \times \{\pm 1\}$, i.e., P_0 produces feature-label pairs. For a valid cost function $c : \mathcal{X}^2 \to \mathbb{R}_{\geq 0}$. , let P_1 represent a shifted version of P_0 such that the marginal distributions of P_0 and P_1 on $\mathcal{X}$. are shifted with $\mathcal{W}_c\left(P_{0,X}, P_{1,X}\right) = \alpha$ for some $\alpha > 0$. No assumption on $P_1\left(y \mid \boldsymbol{X}\right)$ is necessary in this work. Here, the subscript X implies the marginal distribution on $\mathcal{X}$. Let us consider the following two sets of samples:

$$S_0 = \left\{\left(\boldsymbol{X}_i, y_i\right)\right\}_{i=1}^{m} \sim P_0^m, S_1 = \left\{\boldsymbol{X}_i\right\}_{i=1}^{n} \sim P_{1,X}^n, \tag{8}$$

where S_0. indicates the labeled set and S_1 represents the unlabeled out-of-domain data. A classical result from VC-theory states that the generazation gap in learning from only S_0 (with high probability) can be bounded as

$$R\left(\hat{\theta}^{\mathrm{ERM}}, P_0\right) \leq \min_{\theta \in \Theta} R\left(\theta, P_0\right) + \mathcal{O}\left(\sqrt{\mathrm{VCdim}\left(\mathcal{H}\right)/m}\right) + \sqrt{\mathcal{O}\left(1\right)/m}, \tag{9}$$

where $\mathrm{VCdim}\left(\mathcal{H}\right)$ denotes the VC-dimension of hypothesis class $\mathcal{H}$ [3. This bound can be prohibitively large when $\mathrm{VCdim}\left(\mathcal{H}\right)$ grows uncontrollably, e.g., the case of linear classifiers in very high dimensions ($d >> 1$).

We aim to propose a general framework that leverages both S_0. and S_1 concurrently, and outputs (in polynomial time) an estimator, denoted by $\hat{\theta}^{\mathrm{RSS}}$, such that the second term in the r.h.s. of Eq. (9) would decay faster as one increases both m and n. We are specially interested in cases where $n >> m$. In the next step, we apply our method on a simplified theoretical example in order to give explicit bounds. Similar to [5–8], we fully focus the binary classification problem of a high-dimensional Gaussian mixture model with two components using linear classifiers. Mathatically speaking, for some $\sigma_0 \geq 0$ and $\boldsymbol{\mu}_0 \in \mathbb{R}^d$, let P_0 be the feature-label joint distribution over $\mathbb{R}^d \times \{-1, 1\}$ as follows:

$$P_0\left(y = 1\right) = \frac{1}{2}, P_0\left(\boldsymbol{X} \mid y\right) = \mathcal{N}\left(y\boldsymbol{\mu}_0, \sigma_0^2 \boldsymbol{I}\right). \tag{10}$$

Also, suppose a shifted version of P_0, denoted by P_1 with $P_{1,X} = (1/2)\sum_{u=-1,1} \mathcal{N}\left(u\boldsymbol{\mu}_1, \sigma_1^2 \boldsymbol{I}\right)$, where $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| \le \mathcal{O}(\alpha)$ and $|\sigma_1 - \sigma_0| \le \mathcal{O}(\alpha)$ [2]. Given the two sample sets S_0 and S_1 in this configuration, the problem is to estimate the optimal linear classifier which achieves the minimum error rate.

3 Proposed Method: Robust Self Supervised (RSS) Training

We propose a solution that unifies two generally independent paradigms in machine learning: self-training [41] and distributionally robust learning, formalized in Eq. (6). In self-training, the learner uses its current hypothesis to generate artificial (pseudo-) labels for unlabeled data. Thus, for any unlabeled instance $\boldsymbol{X}_j^{'}$ and parameter $\theta \in \Theta$, we momentarily regard $h_\theta\left(\boldsymbol{X}_j^{'}\right)$ as its label. Building on this idea, the estimator $\hat{\theta}^{\mathrm{RSS}} = \hat{\theta}^{\mathrm{RSS}}(S_0, S_1)$ is defined below.

Definition 5 (Robust Self-Supervised (RSS) Training). RSS augments the robust empirical-risk-minimisation objective with an additional constraint that is evaluated solely on the out-of-domain unlabeled set S_1. For a transportation cost function c and a robustness parameter $\gamma \ge 0$, define the *robust loss* $\phi_\gamma : \mathcal{X} \times \{\pm 1\} \times \Theta \to \mathbb{R}$ by

$$\phi_\gamma(\boldsymbol{X}, y; \theta) = \sup_{\boldsymbol{Z} \in \mathcal{X}} \ell(\boldsymbol{Z}, y; \theta) - \gamma\, c(\boldsymbol{Z}, \boldsymbol{X}). \tag{11}$$

In this regard, for a given set of parameters $\gamma, \gamma', s \in \mathbb{R}_{\ge 0}$, the proposed RSS estimator is defined as

$$\hat{\theta}^{\mathrm{RSS}} = \operatorname{argmin}_{\theta \in \Theta} \left\{ \frac{1}{m}\sum_{i=1}^{m} \phi_\gamma(\boldsymbol{X}_i, y_i; \theta) : \frac{1}{n}\sum_{j=1}^{n} \phi_{\gamma'}\left(\boldsymbol{X}_j^{'}, h_\theta\left(\boldsymbol{X}_j^{'}\right); \theta\right) \le s \right\}. \tag{12}$$

The loss in Eq. (12) comprises two main components. The first term minimises the empirical robust risk on the labelled data S_0, permitting an adversary to shift each example within a Wasserstein ball whose effective radius is set by γ. As shown later (cf. [21]), γ may grow with m, so the ball radius shrinks, and a modest but non-zero adversarial budget can improve generalisation. The second term acts only on the unlabeled data, which have been assigned pseudo-labels by h_θ. This constraint

[2] Having a Wasserstein distance of α between two high-dimensional Gaussian distributions implies that both mean vectors $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1$ and variances σ_0, σ_1 are within a fraction of at most $\mathcal{O}(\alpha)$ from each other.

discourages decision boundaries in densely populated regions of the feature space; its strength is governed jointly by s and γ'.

3.1 Model Optimization: Algorithm and Theoretical Guarantees

Although the optimization problem in Eq. (12) is generally not convex, and therefore strong duality cannot be guaranteed in general, it can be shown that for a convex loss function ℓ, a convex cost function c, sufficiently small s (such that pseudo-labels of samples do not significantly change within the feasible set of θ), and sufficiently large γ and γ' (i.e., sufficiently small Wasserstein radii), the optimization problem Eq. (12) becomes convex. In this scenario, it can be solved to arbitrary precision in polynomial time. Moreover, even if ℓ is non-convex (e.g., when $\mathcal{H}$ is the set of all neural networks), a simple Stochastic Gradient Descent (SGD) algorithm is still guaranteed to converge to at least a local minimum of the dual problem associated with Eq. (12), around whose optimal point and under certain conditions, the duality gap is likely to be very small.

Specifically, the optimization in Eq. (12) constitutes a minimax problem, consisting of an inner maximization (formulated in Eq. (11)) followed by an outer minimization. Provided that the cost function c is strongly convex and that γ or γ' are chosen sufficiently large, the inner maximization problem in Eq. (11) becomes strongly concave [3,40]. This crucial property holds irrespective of the convexity of ℓ, which is particularly important given that ℓ typically lacks convexity in practical settings.

Furthermore, the outer minimization problem in Eq. (12) is differentiable provided ℓ is sufficiently smooth (again, convexity is not required). Consequently, the gradient of Eq. (12) exists and can be efficiently computed using the *Envelope Theorem*. Explicit bounds on the maximum number of iterations required by a simple SGD algorithm (with a mini-batch size of 1) to reach an ε-neighborhood of the global maximum of Eq. (11), as well as a local minimum of Eq. (12), have been derived by [40]. Additionally, gradient formulations for minimax loss functions like Eq. (12), leveraging the envelope theorem, have been established in prior works such as [3,40]. We adopt a similar gradient formulation in the numerical optimization of our model parameters in Section 5, where experimental results with real datasets and neural network models are demonstrated.

The main algorithm and its training procedure are presented in Algorithms 1 and 2. Additionally, we provide a visual overview illustrating our method in Figure 1.

In the next section, we derive theoretical guarantees for $\hat{\theta}^{\mathrm{RSS}}$ and show that it leads to improved generalization bounds when n is sufficiently large and α is controlled.

Algorithm 1 Finding the adversarial perturbed input for original input data based on gradient ascent

function ADVERSARIAL-PERTURB($x, y, \gamma, \alpha, N_s$)
  $x' = x$
  for step = $1, ..., N_s$ **do** ▷ Gradient ascent loop
    $p = model(x')$
    $\phi = CE(p, y) - \gamma.||x' - x||_2^2$ ▷ CE: Cross entropy loss
    $\alpha = \alpha/s$
    $x' = x' + \alpha \nabla_{x'} \phi$
  end for
  return x'
end function

4 Theoretical Guarantees and Generalization Bounds

In this section, we discuss the theoretical aspects of using the RSS training method, specially for the classification of a two-component Gaussian mixture model using linear classifiers, i.e., $\mathcal{H} = \{\text{sign}(<\boldsymbol{\theta}, \cdot>) : \mathbb{R}^d \to \{\pm 1\} \mid \boldsymbol{\theta} \in \mathbb{R}^d\}$. For the sake of simplicity in results, let us define the loss function ℓ as the zero-one loss:

$$\ell(\boldsymbol{X}, y; \theta) = 1(y < \theta, \boldsymbol{X} > \geq 0). \tag{13}$$

The following theorem shows that the proposed RSS estimator in Eq. (12) can potentially improve the generalization bound in a *robust* learning scenario.

Theorem 6. Consider the setup described in Section 2 for the sample generation process (GMM assumption), and the loss function defined in Eq. (12). Using RSS training with m labeled and n unlabeled samples in S_0 and S_1, respectively, and for any $\gamma, \delta > 0$, there exist s and γ' which can be calculated solely based on input samples such that the following holds with probability at least $1 - \delta$:

$$\mathbb{E}_{P_0}\left[\phi_\gamma\left(\boldsymbol{X}, y; \hat{\theta}^{\text{RSS}}\right)\right] \leq \min_{\theta \in \Theta} \mathbb{E}_{P_0}\left[\phi_\gamma\left(\boldsymbol{X}, y; \theta\right)\right] + O\left(\gamma \sqrt{\frac{2d}{m}\left(\alpha + \sqrt{\frac{2d \log \frac{1}{\delta}}{2n+m}}\right)} + \sqrt{\frac{2\log 1/\delta}{m}}\right). \tag{14}$$

The proof, as well as how to calculate s and γ' can be found in Appendix. Theorem 6 presents a generalization bound for the proposed estimator when one considers the robust loss under an adversarial budget, which is characterized by γ. Larger values of γ correspond to smaller Wasserstein radii for the distributional adversary of Eq. (5). The residual term in the r.h.s. of Eq. (14) converges to zero with a faster rate compared to that of Eq. (9), given n is sufficiently large and α is sufficiently small. We derive explicit

conditions regarding this event in Corollary 8. Before that, let us show that for fixed m, as the number of unlabeled samples n increases, the *non-robust excess risk* of the RSS-trained classifier decreases. An important point, also mentioned in the proof of Theorem 6, is that when $\gamma=\infty$, the upper bound in the theorem simplifies to $\mathcal{O}\left(\sqrt{\frac{d}{m}}\right)$ and ϕ_γ reduces to the original loss function ℓ.

Theorem 7. Consider the setting described in Theorem 6. Then, the estimator $\hat{\theta}^{\mathrm{RSS}}$ of Eq. (12) using respectively m labeled and n unlabeled samples, along with specific values of γ, γ', and s which can be calculated solely from the input samples, satisfies the following non-robust generalization bound with probability at least $1-\delta$:

$$R\left(\hat{\theta}^{\mathrm{RSS}},P\right)-\min_{\theta\in\Theta}R\left(\theta,P\right)\le O\left(\min\left\{\sqrt{\frac{d\log\frac{1}{\delta}}{m}},\frac{e^{\frac{-\|\mu_0\|_2^2}{4\sigma_0^2}}}{\sqrt{2\sigma_0\sqrt{2\pi}}}\left(\frac{2d}{m}\alpha+\frac{2d}{m}\sqrt{\frac{2d\log\frac{1}{\delta}}{2n+m}}\right)^{1/4}+\sqrt{\frac{2\log 1/\delta}{m}}\right\}\right). \tag{15}$$

Again, the proof and the procedure for calculating γ,γ', and s are discussed in Appendix.

Based on the previous results, the following corollary showcases a number of surprising non-asymptotic conditions under which our generalization bound becomes superior to conventional approaches.

Corollary 8. Consider the setting described in Theorem 7. Then, $\hat{\theta}^{\mathrm{RSS}}$ of Eq. (12) with m labeled and n unlabeled samples have an advantage over the traditional ERM, if:

$$\alpha\le\mathcal{O}\left(d/m\right),n\ge\Omega\left(m^2/d\right). \tag{16}$$

Also, the following conditions are sufficient to make the minimum required m (for a given error bound) independent of the dimension

$$d:\ \alpha\le\mathcal{O}\left(d^{-1}\right),n\ge\Omega\left(d^3\right). \tag{17}$$

Algorithm 2 The training loop

Require: Number of epochs N_{ep}, Number of perturbation steps N_s, Set of hyper-parameter $\{\gamma, \gamma', \lambda$, Learning rate of perturbation $\alpha\}$

$L = \{(x_0, y_0), (x_0, y_0), \ldots, (x_N, y_N)\}$ ▷ Labeled data with size of N

$U = \{x'_0, x'_1, \ldots, x'_M\}$ ▷ Unlabeled data with the size of M

$k = 2$

$L_b =$ the set of batches of L ▷ batch size = N/k

$U_b =$ the set of batches of U ▷ batch size = M/k

for epoch = $1, \ldots, N_{ep}$ **do**

  for $(x_i, y_i), x'_j$ in L_b, U_b **do**

    $x_i^p = \text{ADVERSERIAL-PERTURB}(x_i, y_i, \gamma, \alpha, step)$

    $y'_j = model(x'_j)$

    $x_j'^p = \text{ADVERSERIAL-PERTURB}(x'_j, y'_j, \gamma', \alpha, step)$

    $p_i = model(x_i^p)$

    $p_j = model(x_j'^p)$

    $\phi_i = CE(p_i, y_i) - \gamma.||x_i^p - x_i||_2^2$

    $\phi_j = CE(p_j, y'_j) - \gamma'.||x_j'^p - x_j||_2^2$

    $l = \phi_i + \lambda.\phi_j$

    backpropagate loss (l) with gradient decent to the deep net and update the weightes

  end for

end for

Proof is given in Appendix. Finally, Theorem 7 also implies that if unlabeled samples are drawn from the same distribution as that of the labeled ones, i.e., $\alpha = 0$, then the excess risk of RSS-training satisfies the following inequality with probability at least $1-\delta$:

$$R\left(\hat{\theta}^{\text{RSS}}, P\right) - \min_{\theta\in\Theta} R(\theta, P) \le \mathcal{O}\left(\left(\frac{d^3 \log 1/\delta}{m^2(2n+m)}\right)^{1/8} + \sqrt{\frac{\log 1/\delta}{m}}\right), \tag{18}$$

which again shows the previously-mentioned improvements when all samples are in-domain.

The assumption of an isotropic GMM with two components has been already studied in the literature (see Section 1). Next, we present a more general case of Theorem 7 where each Gaussian component can have a non-diagonal covariance matrix. Mathematically speaking, suppose that P_0 and P_1 are defined as follows:

$$\begin{aligned} P_0(y=1) = 1/2, P_0(\boldsymbol{X} \mid y) &= N(y\boldsymbol{\mu}_0, \Sigma_0), \\ P_{1,X} &= \frac{1}{2}N(\boldsymbol{\mu}_1, \Sigma_1) + \frac{1}{2}N(-\boldsymbol{\mu}_1, \Sigma_1), \end{aligned} \tag{19}$$

where $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\| \le \mathcal{O}(\alpha)$, $\|\Sigma_1 - \Sigma_0\|_2 \le \mathcal{O}(\alpha)$ and $\|\boldsymbol{\mu}_1\|_2 \ge \beta\lambda_{\max}(\Sigma_1)$. Assume a set of m labeled samples $S_0 \sim P_0^m$, and a set of n unlabeled samples $S_1 \sim P_{1,X}^n$.

Theorem 9 (Generalization Bound for General Gaussian Mixture Models). Consider the setting described in Eq. (19). Using algorithm in Eq. (12) with m labeled and n unlabeled samples, there exists a set of parameters γ, γ', s for which the following holds with probability at least $1-\delta$:

$$R\left(\hat{\theta}^{\mathrm{RSS}}, P\right) - \min_{\theta\in\Theta} R(\theta, P) \leq$$

$$\mathcal{O}\left(e^{\vartheta^2} \left(\sqrt{\frac{\|\boldsymbol{\mu}_1\|_2^2 + \mathrm{Tr}(\Sigma_1)}{m}} \left(C\alpha + \sqrt{\frac{\log\frac{1}{\delta}}{2n+m}} \right) \frac{d\kappa_1\kappa_1'}{\Delta(\Sigma)} \right)^{1/2} + \sqrt{\frac{\log 1/\delta}{m}} \right), \tag{20}$$

Where:

$$\vartheta = \left| \boldsymbol{\mu}_1 \Sigma_1^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0 \Sigma_0^{-1} \boldsymbol{\mu}_0 \right|, C = \left(\frac{\|\mu_0\|^2 + \lambda_{\min}(\Sigma_1)\|\mu_0\|_2}{\lambda_{\min}^2} \right),$$

$$\kappa_1 = \frac{\lambda_{\max}(\Sigma_1)}{\lambda_{\min}(\Sigma_1)}, \kappa_1' = \frac{\lambda_{\max}(\Sigma_1)}{\Delta(\Sigma_1)},$$

$$\Delta(\Sigma_1) = \min\left\{ \lambda_i(\Sigma_1) - \lambda_j(\Sigma_1) \right\}, \forall i, j : \lambda_i(\Sigma_1) \neq \lambda_j(\Sigma_1), \tag{21}$$

and $\lambda_i(\Sigma)$ is the i (th) eigenvalue of Σ.

Proof can be found in Appendix. One important difference to note between Theorem 9 and Theorem 7 is the choice of γ', which controls the adversarial budget for unlabeled (and out-of-domain) part of the dataset. In the setting of Theorem 7, we prefer to choose γ' as small as possible. However, in the setting of Theorem 9, we consider the eigenvectors and eigenvalues of Σ_1 and Σ_0, as well as the direction of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_0$ in order to find the optimal value for the adversarial budget. In fact, there are cases in which selecting a large γ' (less freedom for the adversary) may actually be the optimal choice.

5 Exprimental Results

The effectiveness of the proposed method has been assessed through experimenting on various datasets, including simulated data, and real-world datasets of histopathology images. Each experiment has been divided into two parts: i) cases in which both labeled and unlabeled data are sampled from the same distribution, and ii) the scenarios where the unlabeled data differs in distribution from the labeled ones. First, let us specify the datasets used in our experiments:

1. **Simulated data** consists of binary-labeled data points with a dimension of $d = 200$, generated according to the setting described in Section 2.

2. **NCT-CRC-HE-100K** consists of 100,000 histopathology images of colon tissue [42]. The images have dimensions of 224×224 and were captured at 20x magnification. The dataset is labeled with 9 distinct classes.

3. **PatchCamelyon** is a widely used benchmark dataset for medical image analysis. It consists of a large collection of 327,680 color histopathology images from lymph node, each with dimensions 96×96. The dataset has binary labels for presence/absence of metastatic tissue.

5.1 Experiment of Simulated Data

To evaluate the effectiveness of our method on simulated data, we first find the optimal classifier using only labeled samples. Then, we apply our method with a varying number of unlabeled samples. The results (see Table 1) show that our proposed method achieves accuracy improvements comparable to models trained only on labeled samples. Moreover, results indicate that our method is more effective when labeled and unlabeled data come from the same distribution. However, it still demonstrates significant improvement even when the unlabeled samples undergo a distribution shift.

5.2 Experiment of Histopathology Data

The processing pipeline over the real-world dataset of histopathology images is based on using a ResNet50 encoder pre-trained on ImageNet [43,44], which extracts and stores 1×1024 embeddings from input images. Such embeddings are then used to train a deep neural network with four layers of size 2048 and one output layer for the class id. Also, we have used a LeakyReLU activation function.

Experimental results in this part are shown in Table 2. Under the “same distribution" setting, both labeled and unlabeled data have been taken from the NCT-CRC-HE-100K dataset. On the other hand, “different distributions" setting implies that labeled data comes from the NCT-CRC-HE-100K dataset (labels are either “Normal" or “Tumor"), while the PatchCamelyon dataset was used for the unlabeled data. As a result, the final labeling is binary. The experimental results demonstrate that increasing the number of unlabeled data leads to an improvement in accuracy for both the ‘same’ and ‘different’ distribution settings.

6 Conclusion and Discussion

In this study, we address the robust and non-robust classification challenges with a limited labeled dataset and a larger collection of unlabeled samples, assuming a slight perturbation in the distribution of unlabeled data. We present the first non-asymptotic tradeoff between labeled (m) and unlabeled (n) sample sizes when learning a two-component Gaussian mixture model. Our analysis reveals that when $n \geq \Omega\left(m^2 / d\right)$, the generalization bound improves compared to using only labeled data, even when

unlabeled data points are slightly out-of-domain. We derive sophisticated results for the generalization error in both robust and non-robust scenarios, employing a technique based on optimizing a robust loss and regularization to avoid crowded and dense areas. Our framework integrates tools from self-training, distributionally robust learning, and optimal transport.

The method developed in this paper may also extend to more general settings, for example when the class priors are unbalanced or when the domain centers are not aligned. Analyzing these variations separately is often straightforward. For instance, when the classes are balanced, we can recenter the samples (i.e., subtract the empirical mean) and proceed with essentially the same analysis. Likewise, when the samples are centered but the class priors differ between the positive and negative classes, the analysis can typically be adapted with only minor modifications. However, when both effects occur simultaneously (misaligned centers and class-prior imbalance) the problem became more challenging.

Experiments on synthetic and real-world datasets validate our theoretical findings, demonstrating improved classification accuracy, even for non-Gaussian cases, by incorporating out-of-domain unlabeled samples. Our methodology hinges on leveraging such data to enhance robust accuracy and adapting the uncertainty neighborhood radius based on labeled and unlabeled sample quantities to strike a balance between bias and variance in classification error.

For future work, there's room for improving and relaxing the conditions for the utility of unlabeled data. Exploring error lower-bounds and impossibility results presents another intriguing avenue. Additionally, relaxing the constraints on the level of distribution shift for out-of-domain samples could be a promising direction.

7 Acknowledgement

Babak Hossein Khalaj is with the Electrical Engineering Department, Sharif University of Technology, Tehran, Iran, also with the School of Computer Science, Institute for Research in Fundamental Sciences, Tehran, Iran (email:khalaj@sharif.edu).

The work of Babak Hossein Khalaj was supported in part by a Grant from the Institute for Research in Fundamental Sciences (IPM).

8 References

[1] Goodfellow IJ, Shlens J, Szegedy C. "Explaining and harnessing adversarial example", *arXive*, (2015). https://doi.org/10.48550/arXiv.1412.6572.

[2] Biggio B, Roli F. "Wild patterns: ten years after the rise of adversarial machine learning" , *ACM* Digital Library, (2018). https://doi.org/10.1016/j.patcog.2018.07.023.

[3] Najafi A, Maeda S, Koyama M, et al. "Robustness to adversarial perturbations in learning from incomplete data" , *arXive*, (2019). https://doi.org/10.48550/arXiv.1905.13021.

[4] Miyato T, Maeda S, Koyama M, et al. “Virtual adversarial training: a regularization method for supervised and semi-supervised learning” , *arXive*, (2018). https://doi.org/10.48550/arXiv.1704.03976.

[5] Schmidt L, Santurkar S, Tsipras D, et al. “Adversarially robust generalization requires more data” , *arXive*, (2018). https://doi.org/10.48550/arXiv.1804.11285.

[6] Carmon Y, Raghunathan A, Schmidt L, et al. “Unlabeled data improves adversarial robustness” , *arXive*, (2022). https://doi.org/10.48550/arXiv.1905.13736.

[7] Zhai R, Cai T, He D, et al. “Adversarially robust generalization just requires more unlabeled data” , *arXive*, (2019). https://doi.org/10.48550/arXiv.1906.00555.

[8] Uesato J, Alayrac J-B, Huang P-S, et al. “Are labels required for improving adversarial robustness? ” , *arXive*, (2019). https://doi.org/10.48550/arXiv.1905.13725.

[9] Kumar A, Ma T, Liang P. “Understanding self-training for gradual domain adaptation” , *arXive*,(2020). https://doi.org/10.48550/arXiv.2002.11361.

[10] Carlini N, Wagner D. “Towards evaluating the robustness of neural networks” , *arXive*, (2017). https://doi.org/10.48550/arXiv.1608.04644.

[11] Deng Z, Zhang L, Ghorbani A, et al. “Improving adversarial robustness via unlabeled out-of-domain data” , *arXive*, (2021). https://doi.org/10.48550/arXiv.2006.08476.

[12] Rigollet P. “Generalization error bounds in semi-supervised classification under the cluster assumption. *JMLR,* (2007);8. https://doi.org/10.48550/arXiv.math/0604233.

[13] Blanchet J, Kang Y. “Semi-supervised learning based on distributionally robust optimization” , *Wiley* Online Library, (2019). https://doi.org/10.1002/9781119721871.ch1.

[14] Frogner C, Claici S, Chien E, et al. “Incorporating unlabeled data into distributionally robust learning” , *arXive*, (2019). https://doi.org/10.48550/arXiv.1912.07729.

[15] Nguyen A, Yosinski J, Clune J. “Deep neural networks are easily fooled: high confidence predictions for unrecognizable images” , *arXive*, (2015). https://doi.org/10.48550/arXiv.1412.1897.

[16] Szegedy C, Zaremba W, Sutskever I, et al. “Intriguing properties of neural networks” , *arXive*, (2014). https://doi.org/10.48550/arXiv.1312.6199.

[17] Staib M, Jegelka S. “Distributionally robust deep learning as a generalization of adversarial training. ” *NIPS* Work. Mach. Learn. Comput. Secur., (2017), p. 4.

[18] Shafieezadeh-Abadeh S, Esfahani PM, Kuhn D. “Distributionally robust logistic regression” , *arXive*, (2015). https://doi.org/10.48550/arXiv.1509.09259.

[19] Esfahani PM, Kuhn D. “Data-driven distributionally robust optimization using the wasserstein metric: performance guarantees and tractable reformulations” , *arXive*, (2017). https://doi.org/10.48550/arXiv.1505.05116.

[20] Duchi J, Namkoong H. “Variance-based regularization with convex objectives” , *arXive*, (2017). https://doi.org/10.48550/arXiv.1610.02581.

[21] Gao R. “Finite-sample guarantees for wasserstein distributionally robust optimization: breaking the curse of dimensionality” , *arXive*, (2022). https://doi.org/10.48550/arXiv.2009.04382.

[22] Germain P, Habrard A, Laviolette F, et al. “A new pac-bayesian perspective on domain adaptation” , *arXive*, (2016). https://doi.org/10.48550/arXiv.1506.04573.

[23] Germain P, Habrard A, Laviolette F, et al. “PAC-Bayes and domain adaptation”. *neurocomputing*

(2020);379:379–97. https://doi.org/10.1016/j.neucom.2019.10.105.

[24] Sicilia A, Atwell K, Alikhani M, et al. “PAC-Bayesian domain adaptation bounds for multiclass Learners” , *arXive*, (2022). https://doi.org/10.48550/arXiv.2207.05685.

[25] Redko I, Habrard A, Sebban M. “Theoretical analysis of domain adaptation with optimal transport”, *Springer* International Publishing, (2017), p. 737–53. https://doi.org/10.1007/978-3-319-71246-8_45.

[26] Kerdoncuff T, Emonet R, Sebban M. “Metric learning in optimal transport for domain adaptation. ” Proc. Twenty-Ninth Int. Jt. Conf. Artif. Intell., California: *ICAIO*, International Joint Conferences on Artificial Intelligence Organization; (2020), p. 2162–8. https://doi.org/10.24963/ijcai.2020/299.

[27] Wang Z, Mao Y. “Information-theoretic analysis of unsupervised domain adaptation” , *arXive*, (2023). https://doi.org/10.48550/arXiv.2210.00706.

[28] Zhao H, Combes RT des, Zhang K, et al. “On learning invariant representation for domain adaptation” , *arXive*, (2019). https://doi.org/10.48550/arXiv.1901.09453.

[29] Tachet R, Zhao H, Wang Y-X, et al. “Domain adaptation with conditional distribution matching and generalized label shift” , *arXive*, (2020). https://doi.org/10.48550/arXiv.2003.04475.

[30] Phung T, Le T, Vuong L, et al. “On learning domain-invariant representations for transfer learning with multiple sources” , *arXive*, (2021). https://doi.org/10.48550/arXiv.2111.13822.

[31] Mehra A, Kailkhura B, Chen P-Y, et al. “Understanding the limits of unsupervised domain adaptation via data poisoning ” , *arXive*,(2021). https://doi.org/10.48550/arXiv.2107.03919.

[32] He Y, Wang H, Li B, et al. “Gradual domain adaptation: theory and algorithms” , *arXive*,(2025). https://doi.org/10.48550/arXiv.2310.13852.

[33] Shi L, Liu W. “Adversarially robust unsupervised domain adaptation.” *Artif Intell* (2025);347:104383. https://doi.org/10.1016/j.artint.2025.104383.

[34] Zhang H, Yu Y, Jiao J, et al. “Theoretically principled trade-off between robustness and accuracy” , *arXive*, (2019). https://doi.org/10.48550/arXiv.1901.08573.

[35] De Silva A, Ramesh R, Priebe CE, et al. “The value of out-of-distribution data” , *arXive*, (2023). https://doi.org/10.48550/arXiv.2208.10967.

[36] Lee S, Park C, Lee H, et al. “Removing undesirable feature contributions using out-of-distribution data” , *arXive*, (2021). https://doi.org/10.48550/arXiv.2101.06639.

[37] Mohri M, Rostamizadeh A, Talwalkar A. “Foundations of machine learning”. The *MIT Press*; (2018). https://mitpress.mit.edu/9780262039406/foundations-of-machine-learning/

[38] Ashtiani H, Ben-David S, Harvey N, et al. “Near-optimal sample complexity bounds for robust learning of gaussians mixtures via compression schemes” , *arXive*, (2020). https://doi.org/10.48550/arXiv.1710.05209.

[39] Blanchet J, Kang Y, Murthy K. “Robust wasserstein profile inference and applications to machine learning”, Journal of Applied Probability (*JAP*), (2019) , pp. 830 - 857. https://doi.org/10.1017/jpr.2019.49.

[40] Sinha A, Namkoong H, Volpi R, et al. “Certifying some distributional robustness with principled adversarial training” , *arXive*, (2020). https://doi.org/10.48550/arXiv.1710.10571.

[41] Amini M-R, Gallinari P. “Semi-supervised logistic regression. ”, *ECAI,* (2002);2:11. https://doi.org/10.5555/3000905.3000988.

[42] Katherm JN, Halama, Niels, et al. “100,000 histological images of human colorectal cancer and healthy tissue 2018. ” https://doi.org/10.5281/zenodo.1214456. (accessed May 18, 2026).

[43] Deng J, Dong W, Socher R, et al. “ImageNet: A large-scale hierarchical image database. ”, *IEEE* Conf. Comput. Vis. Pattern Recognit. (2009), IEEE; 2009, p. 248–55. https://doi.org/10.1109/CVPR.2009.5206848.

[44] He K, Zhang X, Ren S, et al. “Deep residual learning for image recognition.” *IEEE* Conf. Comput. Vis. Pattern Recognit., IEEE; (2016), p. 770–8. https://doi.org/10.1109/CVPR.2016.90.

9 List of Captions

Figure 1: A visual overview of Robust Self Supervised (RSS) Training.

Table 1: Accuracy of the model trained on labeled datasets of sizes 10, 20, 40, and 10,000 with varying amounts of unlabeled data from the same distribution with α=0 (left), and different distribution with α=$\frac{1}{2}||\mu_0||_2$ (right).

Table 2: Accuracy of the model trained on labeled data from NCT-CRC-HE-100K dataset with varying amounts of unlabeled data from the same distribution (left), as well as when unlabeled samples come from a different distribution (PatchCamelyon dataset)(right).

10 List of Figures

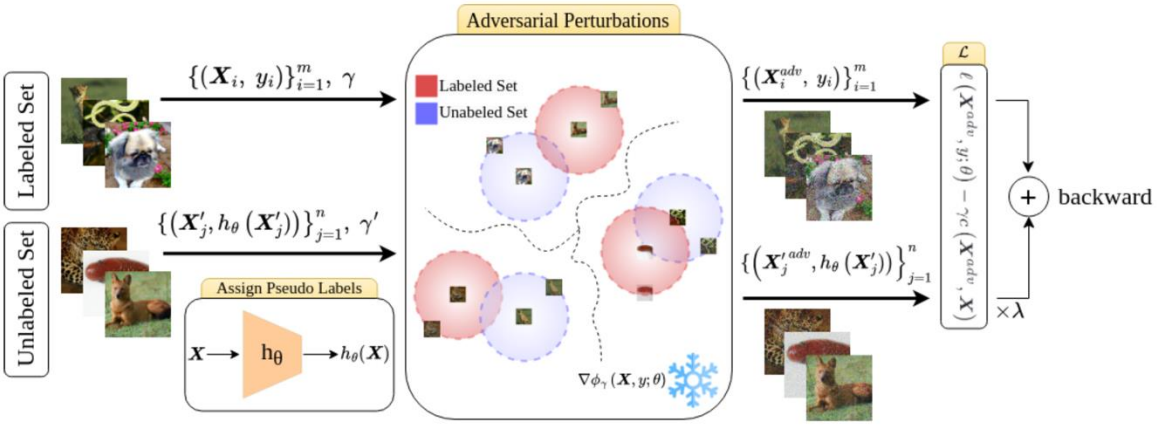

Figure 1: A visual overview of Robust Self Supervised (RSS) Training.

11 List of Tables

Table 1

Same distribution				**Different distribution**			
1-4 (lr)5-8 Labeled size	Acc	Unlabeled size	Acc	Labeled size	Acc	Unlabeled size	Acc
10	0.59	10	0.63	10	0.59	10	0.61
		100	0.66			100	0.65
		1,000	0.79			1,000	0.78
		10,000	**0.82**			10,000	**0.81**
20	0.62	20	0.64	20	0.62	20	0.65
		200	0.69			200	0.65
		2,000	0.80			2,000	0.79
		10,000	**0.82**			10,000	**0.80**
		40	0.65			40	0.65
40	0.65	400	0.71	40	0.65	400	0.73
		4,000	0.81			4,000	0.78
		10,000	**0.82**			10,000	**0.80**
10,000	**0.83**	-	-	10,000	**0.83**	-	-

Table 2

Same distribution				**Different distribution**			
1-4 (lr)5-8 Labeled size	Acc	Unlabeled size	Acc	Labeled size	Acc	Unlabeled size	Acc
48	0.65	200	0.71	25	0.78	100	0.78
		700	0.80			400	0.79
		2,000	**0.82**			2,000	**0.81**
240	0.77	500	0.78	50	0.82	200	0.82
		1,200	0.82			700	0.86
		4,000	**0.83**			3,000	**0.87**
1040	0.83	3,000	0.87	300	0.87	600	0.88
		10,000	0.89			2,000	0.89
		20,000	**0.91**			8,000	**0.90**
50,000	**0.916**	-	-	32,000	**0.94**	-	-

12 Biographies

Seyed Amirhossein Saberi: Seyed Amirhossein Saberi is a researcher in Theoretical Machine Learning and Data Science. He received his B.Sc. degree in Electrical Engineering (Communication Systems) from Sharif University of Technology in 2014, followed by an M.Sc. degree in Bioelectric Engineering from the same university in 2016. He was directly admitted to the Ph.D. program in Electrical Engineering (Communication Systems) at Sharif University in 2018, where he is currently pursuing his doctoral studies. His research lies at the intersection of machine learning theory, signal processing, and data science, with a focus on developing rigorous mathematical frameworks for learning systems and exploring their applications in communication and intelligent data-driven technologies.

Amir Najafi: Amir Najafi is an Assistant Professor in the Department of Computer Engineering at Sharif University of Technology, Tehran, Iran, since 2023. He received his B.Sc. and M.Sc. degrees in Electrical Engineering in 2012 and 2015, respectively, and his Ph.D. in Artificial Intelligence in 2020, all from Sharif University of Technology. Prior to joining Sharif as a faculty member, he was a postdoctoral researcher at the School of Mathematics, Institute for Research in Fundamental Sciences (IPM), from 2020 to 2023. He has also been a visiting research scholar at the Broad Institute of MIT and Harvard in 2017, and a part-time researcher at Preferred Networks Inc. in Tokyo, Japan in 2018. His research interests include machine learning theory and statistics.

Alireza Heidary: Alireza Heidary is an AI Researcher, Data Scientist, and Software Engineer currently pursuing a Master of Science in Computing Science at Simon Fraser University (SFU). Since 2025, he has served as a Research Assistant at SFU, actively contributing to advanced academic research within the department. Prior to his graduate studies, Alireza earned his Bachelor of Science in Computer Engineering from the prestigious Sharif University of Technology in 2024. His academic and professional trajectory reflects a strong intersection of theoretical artificial intelligence research, data analytics, and practical software engineering.

Mohammad Hosein Movasaghinia: Mohammad Hosein Movasaghinia is a Data Scientist and AI researcher. He received his B.Sc. degree in Computer Engineering from Shahed University in 2021 and his M.Sc. degree in Bioinformatics from the Department of Computer Engineering at Sharif University of Technology in 2024, where he ranked first in both programs. He works at the intersection of machine learning research and engineering, with experience leading the development of AI-driven systems, adapting and deploying machine learning models for real-world applications, and building LLM-based multi-agent assistant systems. His research interests include machine learning, deep learning, and bioinformatics.

Seyed Abolfazl Motahari: Seyed Abolfazl Motahari received the B.Sc. degree from the Iran University of Science and Technology (IUST), Tehran, in 1999, the M.Sc. degree from the Sharif University of Technology, Tehran, in 2001, and the Ph.D. degree from the University of Waterloo, Waterloo, Canada, in 2009, all in electrical engineering. He is

currently an Assistant Professor with the Computer Engineering Department, Sharif University of Technology. From August 2000 to August 2001, he was a Research Scientist with the Advanced Communication Science Research Laboratory, Iran Telecommunication Research Center (ITRC), Tehran. From October 2009 to September 2010, he was a Post-Doctoral Fellow with the University of Waterloo. From September 2010 to July 2013, he was a Post-Doctoral Fellow with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley.

Babak Hossein Khalaj: Babak Hossein Khalaj (Senior Member, IEEE) received his B.Sc. degree in electrical engineering from the Sharif University of Technology, Tehran, Iran, in 1989, and the M.Sc. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 1993, and 1996, respectively. Since 1999, he has been a Senior Consultant in the area of data communications, and from 2006 to 2007, he was a Visiting Professor with CEIT, San Sebastian, Spain. He has coauthored many papers in signal processing and digital communications and holds four U.S. patents. He was the recipient of Alexander von Humboldt Fellowship from 2007 to 2008 and Nokia Visiting Professor Scholarship in 2018.