AFRL-AFOSR-JP-TR-2024-0041

Optimal Transport Theory for Machine Learning with Limited and Less Labels

TRUNG LE
MONASH UNIVERSITY
WELLINGTON RD
CLAYTON, VIC, 3168
AUS

01/11/2024
Final Technical Report

Air Force Research Laboratory
Air Force Office of Scientific Research
Asian Office of Aerospace Research and Development
Unit 45002, APO AP 96338-5002

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
20240111	Final	START DATE 20210909	END DATE 20230908

4. TITLE AND SUBTITLE
Optimal Transport Theory for Machine Learning with Limited and Less Labels

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
	FA2386-21-1-4049	61102F
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Trung Le, Dinh Phung, He Zhao

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
MONASH UNIVERSITY
WELLINGTON RD
CLAYTON, VIC 3168
AUS

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AOARD UNIT 45002 APO AP 96338-5002	AFRL/AFOSR IOA	AFRL-AFOSR-JP-TR-2024-0041

12. DISTRIBUTION/AVAILABILITY STATEMENT
A Distribution Unlimited: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Due to the exponential growth of unlabeled data collected and the hardness, labourexpensiveness, errors, and time-consuming of the process to label data, learning with less and limited labels (LwLL) has emerged as an important research endeavor in machine learning and deep learning. LwLL aims to escalate learning process with limited labeled data similar to the way human learns a new concept effortlessly. It covers a set of related problems, notably transfer learning, domain adaptation (DA), meta-learning, domain generalization, and semisupervised learning. In parallel, optimal transport (OT) is a recent powerful mathematical theory that has been rapidly become a mainstream research tool in machine learning. With its attractive geometry interpretation, computational tractability and expressiveness, OT offers a principal tool to address several LwLL problems. However, this connection is still very limited and under-explored in the current literature. To this end, this proposal aims to investigate OT theory for LwLL- a problem which, to our best knowledge, is new and novel. In particular, we plan to discover a bridging theory connecting OT and LwLL that can elegantly exploit the unique characteristics of OT transport (e.g., distribution matching and clustering view) for advancing the current state-of-the-art in LwLL.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U	SAR	46

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
AKIRA NAMATAME	3152277010

Standard Form 298 (Rev.5/2020)
Prescribed by ANSI Std. Z39.18

Final Report for AOARD Grant FA2386-21-1-4049

Optimal Transport Theory for Machine Learning with Limited and Less Labels

INVESTIGATORS

Principal Investigator: Dr **Trung Le**
Monash University, Australia
Email: trunglm@monash.edu

Co-Principal Investigator: Professor **Dinh Phung**
Monash University, Australia
Email: dinh.phung@monash.edu

Co-Investigator: Dr **He Zhao**
Monash University, Australia
Email: ethan.zhao@monash.edu

Abstract

Due to the exponential growth of unlabeled data collected and the hardness, labour-expensiveness, errors, and time-consuming of the process to label data, learning with less and limited labels (LwLL) has emerged as an important research endeavor in machine learning and deep learning. LwLL aims to escalate learning process with limited labeled data similar to the way human learns a new concept effortlessly. It covers a set of related problems, notably transfer learning, domain adaptation (DA), meta-learning, domain generalization, and semi-supervised learning. In parallel, optimal transport (OT) is a recent powerful mathematical theory that has been rapidly become a mainstream research tool in machine learning. With its attractive geometry interpretation, computational tractability and expressiveness, OT offers a principal tool to address several LwLL problems. However, this connection is still very limited and under-explored in the current literature. To this end, this proposal aims to investigate OT theory for LwLL- a problem which, to our best knowledge, is new and novel. In particular, we plan to discover a bridging theory connecting OT and LwLL that can elegantly exploit the unique characteristics of OT transport (e.g., distribution matching and clustering view) for advancing the current state-of-the-art in LwLL.

1 Publication Outcomes

These results have been documented in 5 research papers accepted for publication at the top-notch conferences in machine learning and artificial intelligence including UAI2022, ICLR2023, KDD2023, AISTATS2023, and NeurIPS2023. In what follows, we provide the details of publications and briefly summarize each publication.

1. Tuan Nguyen, Van Nguyen, **Trung Le**, **He Zhao**, Hung-Quan Tran, **Dinh Phung**. "Cycle class consistency with distributional optimal transport and knowledge distillation for unsupervised domain adaptation". Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence (UAI), 2022.
   - This work develops a distributional optimal transport technique that can be applied successfully to single source transfer learning to earn state-of-the-art performance.

2. Vy Vo, Van Nguyen, **Trung Le**, Quan Hung Tran, Gholamreza Haffari, Seyit Camtepe, **Dinh Phung**. "An Additive Instance-Wise Approach to Multi-class Model Interpretation". International Conference on Representation Learning (ICLR), 2023.
   - This work proposes a novel technique for interpretable machine learning and explainable AI that earns state-of-the-art performance for text and tabular data.

3. Hoang Phan, **Trung Le**, Trung Phung, Anh Tuan Bui, Nhat Ho, **Dinh Phung**. "Global-Local Regularization Via Distributional Robustness". Proceedings of The 26th International Conference on Artificial Intelligence and Statistics (AISTATS), 2023
   - This work proposes a novel optimal transport-based distributional robustness technique that can be applied to many applications including domain adaptation, meta-learning, semi-supervised learning, and adversarial machine learning.

4. Vy Vo, **Trung Le**, Van Nguyen, **He Zhao**, Edwin Bonilla, Gholamreza Haffari, **Dinh Phung**. "Feature-based Learning for Diverse and Privacy-Preserving Counterfactual Explanations". In ACM SIGKDD Conference On Knowledge Discovery and Data Mining (KDD), 2023.
   - This work proposes a novel explainable technique to identify counterfactual explanations that are diverse and privacy-preserving. This work achieves ***the student best paper award at KDD 2023*** *which can be regarded as the top 1 conference in data mining.*

5. Van-Anh Nguyen, Tung-Long Vuong, Hoang Phan, Thanh-Toan Do, Dinh Phung, Trung Le. "Flat Seeking Bayesian Neural Networks". In Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS), 2023.
   - This work develops the concept of robustness for Bayesian neural networks to improve the generalization ability of Bayesian neural networks.

2 Acknowledgment, Collaborations, Partnerships, and Human Involved

We deeply appreciate the US Air Force for sponsoring and supporting this fundamental research. This sponsorship is especially precious and valuable to Dr Trung Le at his early research career. This strongly assists us in promoting and strengthening our research and collaboration with our research partners including Dr He Zhao from CSIRO, Dr Hung Bui, Dr Toan Tran, Dr Anh Nguyen from VinAI Research, and Dr Nhat Ho from the University of Texas Austin (USA). We really enjoy the journey of this research and the fruitful outcome from this research opens doors for our further and future research.

Moreover, the US Air Force grant helps us to receive DP23 Grant: "*Exploiting Geometries of Learning for Fast, Adaptive and Robust AI*" (450,000 AUDs) with Professor **Dinh Phung**; Dr Mehrtash Harandi; Professor Richard Hartley; Dr **Trung Le**; Dr Piotr Koniusz. Last but not least, the US Air Force grant strongly assists us in promoting new research problems in our research group which involve and engage our PostDoc and PhD students including Dr Anh Bui (a PostDoc), Mr Tuan Nguyen (a PhD student), Miss Vy Vo (a PhD student), and Miss Van-Anh Nguyen (a PhD student).

Cycle Class Consistency with Distributional Optimal Transport and Knowledge Distillation for Unsupervised Domain Adaptation

Tuan Nguyen[1] **Van Nguyen**[2] **Trung Le**[1] **He Zhao**[1] **Quan Hung Tran**[3] **Dinh Phung**[1,4]

[1]Department of Data Science and AI, Monash University, Australia
[2]The University of Adelaide, Australia
[3]Adobe Research, San Jose, CA, USA
[4]VinAI Research, Vietnam

Abstract

Unsupervised domain adaptation (UDA) aims to transfer knowledge from a model trained on a labeled source domain to an unlabeled target domain. To this end, we propose in this paper a novel cycle class-consistent model based on optimal transport (OT) and knowledge distillation. The model consists of two agents, a teacher and a student cooperatively working in a cycle process under the guidance of the distributional optimal transport and distillation manner. The OT distance is designed to bridge the gap between the distribution of the target data and a distribution over the source class-conditional distributions. The optimal probability matrix then provides pseudo labels to learn a teacher that achieves a good classification performance on the target domain. Knowledge distillation is performed in the next step in which the teacher distills and transfers its knowledge to the student. And finally, the student produces its prediction for the optimal transport step. This process forms a closed cycle in which the teacher and student networks are simultaneously trained to conduct transfer learning from the source to the target domain. Extensive experiments show that our proposed method outperforms existing methods, especially the class-aware and OT-based ones on benchmark datasets including Office-31, Office-Home, and ImageCLEF-DA.

1 INTRODUCTION

Unsupervised domain adaptation (UDA) allows us to transfer knowledge from a model trained on a source domain with labels to a target domain without any labels. To cope with structural data more efficiently and effectively, deep domain adaptation (DDA) [Ganin and Lempitsky, 2015] has been proposed and widely studied [Nguyen et al., 2019, 2020, Phung et al., 2021]. To tackle the data shift issue and learn domain-invariant features, DDA aims to bridge the distribution gap between the source and target domains in a latent space using a feature extractor. Guided by this principle, most of the existing works in DDA propose minimizing a divergence between the source and target distributions in the latent space. Popular choices of divergence include the Jensen-Shannon (JS) divergence [Ganin and Lempitsky, 2015, Tzeng et al., 2015, Shu et al., 2018], the maximum mean discrepancy (MMD) distance [Gretton et al., 2007, Long et al., 2015], and the Wasserstein (WS) distance [Shen et al., 2018, Lee et al., 2019, Le et al., 2021a].

Recently, Optimal transport (OT) [Villani, 2008, Santambrogio, 2015], a powerful tool in mathematics with rich and rigorous theories, has been widely applied in deep domain adaptation [Courty et al., 2017b,a, Damodaran et al., 2018, Redko et al., 2019, Lee et al., 2019, Xie et al., 2019, Xu et al., 2020, Nguyen et al., 2021a,b, Le et al., 2021b, Nguyen et al., 2021c]. From the conceptual perspective, OT-based methods encourage the target samples to move towards the source samples by minimizing a transportation cost. However, since the transportation cost usually engages the pairs of target and source samples without considering label information of the source samples, the movement of the target samples to the source domain seems to be unaware of the class regions in that domain, hence cannot resolve the label shift issue [Tachet des Combes et al., 2020]. Although OT has been initially used for solving this problem [Courty et al., 2017b, Damodaran et al., 2018], the performance of the existing methods is still less satisfactory compared with the state-of-the-art ones.

In this paper, we propose a novel distributional OT that enables the incorporation of the source label information when engaging and matching target and source samples. Specifically, in the source domain we consider that one label is associated with a conditional distribution over all the samples conditioned on that label. Next, we define a distribution over these conditional distributions of all the

Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022), PMLR 180:1519–1529.

labels in the source domain. In the target domain where there are no labels, we also consider a distribution over all the target samples. With the two distributions for the source and target domains respectively, we formulate the DA problem as the computation of the OT distance between the two distributions. The OT transport plan gives us the information of how a target sample related to the source samples by taking into account the source domain labels. The challenge here is how to define the cost function, which indicates the transport cost of OT between a target sample and a source class-conditional distribution. To tackle this challenge, we propose a cycle class consistency framework in which we leverage the advantages of knowledge distillation (KD) which has recently obtained outstanding achievements [Tian et al., 2020, Zhao et al., 2020, Tejankar et al., 2021, Feng et al., 2021]. We name our proposed approach ***C**ycle Class C**O**nsistency with **O**ptimal Transport and **K**nowledge Distillation for Unsupervised Domain Adaptation* (COOK). In summary, our contributions in this paper include:

- We propose a novel distributional OT which seeks the optimal matching between the target and source examples taking into account the source label information for reducing the label and data shift, two challenging problems of UDA.
- We connect KD and OT to further improve the performance of class-aware UDA methods via proposing a cycle class consistency framework where the teacher and student networks cooperatively work in a distillation process and support to reduce the mismatch between the target distribution and the source class-conditional distributions.
- We conduct experiments to compare our proposed COOK with the existing standard UDA methods, especially class-aware UDA methods (e.g., RADA [Wang et al., 2019b] and CAN [Kang et al., 2019]), and OT-based UDA methods (e.g., DeepJDOT [Damodaran et al., 2018], ETD [Li et al., 2020], and RWOT [Xu et al., 2020]). The experimental results show that our proposed method surpasses the baselines on the benchmark datasets including *Office-31*, *Office-Home*, and *ImageCLEF-DA*.

2 RELATED WORK

2.1 STANDARD DA

Deep domain adaptation has been intensively studied and shown appealing performance in various tasks and applications, notably in Ganin and Lempitsky [2015], Long et al. [2015], et al. [2017, 2018]. The core idea of DDA is to bridge the gap between source and target distributions in a joint space by minimizing a divergence between distributions induced from the source and target domains in this space. Popular choices of divergence include Jensen-Shannon divergence [Ganin and Lempitsky, 2015, Tzeng et al., 2015, Shu et al., 2018]; maximum mean discrepancy distance [Gretton et al., 2007, Long et al., 2015]; and WS distance [Shen et al., 2018, Lee et al., 2019, Le et al., 2021a]. Some recent works have exploited different aspects of UDA for improving the performance [Kurmi et al., 2019, Wang et al., 2019a, Chen et al., 2019, Hu et al., 2020]. Typically, CADA [Kurmi et al., 2019] considered the probabilistic certainty estimate of various regions and used these certainty estimate weights for improving the classifier performance on the target dataset. GSDA [Hu et al., 2020] introduced a novel method named Hierarchical Gradient Synchronization to model the synchronization relationship among the local distribution pieces and global distribution, aiming for more precise domain-invariant features.

2.2 OPTIMAL TRANSPORT BASED DA

Optimal transport theory has been applied to domain adaptation in Courty et al. [2017b,a], Damodaran et al. [2018], Redko et al. [2019], Lee et al. [2019], Xie et al. [2019], Xu et al. [2020]. Particularly, Lee et al. [2019] proposed using sliced Wasserstein distance for domain adaption, whereas Xie et al. [2019] proposed SPOT in which the optimal transport plan is approximated by a pushforward of a reference distribution, and cast the optimal transport problem into a minimax problem. Recent OT-based DA work (RWOT) [Xu et al., 2020] leveraged spatial prototypical information and intra-domain structures of image data to reduce the negative transfer caused by target samples near decision boundaries. Moreover, Courty et al. [2017b] proposed an idea to connect the theory of optimal transport and domain adaptation, which later inspired an OT-based deep DA method (DeepJDOT) [Damodaran et al., 2018]. Another recent work (ETD) [Li et al., 2020] tackled the bottlenecks of OT in UDA by developing an attention-aware OT distance to measure the domain discrepancy under the guidance of the prediction-feedback. Our proposed approach is totally different from existing OT based DA approaches in which we examine an OT distance discrete distribution over source class-conditional distributions and the target data distribution. By investigating this specific OT distance and minimizing it, we can guide target examples moving to an appropriate source class on the latent space for mitigating both data and label shifts.

2.3 CLASS-AWARE DA

Some recent approaches [Wang et al., 2019b, Kang et al., 2019] leverage the useful information from the label space to improve the quality of the alignment between the source and target domains. Wang et al. [2019b] proposed a novel relationship-aware adversarial domain adaptation (RADA) algorithm. It first uses a single multi-class domain discrimi-

nator to enforce the learning of inter-class dependency structure during domain-adversarial training. After that, it aligns this structure with the inter-class dependencies that are characterized from training the label predictor on source domain. Furthermore, the authors imposed a regularization term in order to penalize the structure discrepancy between the inter-class dependencies estimated from domain discriminator and label predictor. With this alignment, RADA makes the adversarial domain adaptation aware of the class relationships. Kang et al. [2019] proposed a contrastive adaptation network (CAN) which optimizes a new metric modeling the intra-class domain discrepancy and the inter-class domain discrepancy. In particular, the authors introduced a new contrastive domain discrepancy (CDD) objective to enable class-aware UDA. CAN aims to faciliate the optimization with CDD (established on maximum mean discrepancy (MMD) [Long et al., 2015]).

3 BACKGROUND

In what follows, we present the background of OT for two discrete distributions, which is used in our work. Consider two discrete distributions: $\mathbb{P}^1 = \sum_{i=1}^{M} \pi_i^1 \delta_{\mathbf{x}_i^1}$ and $\mathbb{P}^2 = \sum_{j=1}^{N} \pi_j^2 \delta_{\mathbf{x}_j^2}$ where $\boldsymbol{\pi}^1 = \left[\pi_i^1\right]_{i=1}^{M}$ and $\boldsymbol{\pi}^2 = \left[\pi_j^2\right]_{j=1}^{N}$ are probability masses, $\left\{\mathbf{x}_i^1\right\}_{i=1}^{M}$ and $\left\{\mathbf{x}_j^2\right\}_{j=1}^{N}$ are the sets of atoms, and $\delta_{\mathbf{x}}$ is the Dirac delta distribution concentrated at the atom $\mathbf{x}$. Let $c\left(\mathbf{x}_i^1, \mathbf{x}_j^2\right)$ be a cost function. The OT distance between $\mathbb{P}^1$ and $\mathbb{P}^2$ w.r.t. the cost function c is defined as

$$\min_{A \in \mathbb{R}_+^{M \times N}} \sum_{i=1}^{M} \sum_{j=1}^{N} a_{ij} c\left(\mathbf{x}_i^1, \mathbf{x}_j^2\right), \tag{1}$$

where $A = [a_{ij}] \in \mathbb{R}_+^{M \times N}$ of non-negative elements satisfying $\sum_{j=1}^{N} a_{ij} = \pi_i^1, \forall i \in \{1, ..., M\}$ and $\sum_{i=1}^{M} a_{ij} = \pi_j^2, \forall j \in \{1, ..., N\}$.

In addition, $a_{ij} \in [0; 1]$ is interpreted as the probability to match $\mathbf{x}_i^1$ and $\mathbf{x}_j^2$ or to transport $\mathbf{x}_i^1$ to $\mathbf{x}_j^2$, which suffers the cost $c\left(\mathbf{x}_i^1, \mathbf{x}_j^2\right)$. Therefore, the sum $\sum_{i=1}^{M} \sum_{j=1}^{N} a_{ij} c\left(\mathbf{x}_i^1, \mathbf{x}_j^2\right)$ can be viewed as the total cost to match $\mathbb{P}^1$ and $\mathbb{P}^2$ or to transport $\mathbb{P}^1$ to $\mathbb{P}^2$. By solving the optimization problem in Eq. (1), we aim to find the optimal transportation matrix A^* which minimizes the total cost.

4 DISTRIBUTIONAL OPTIMAL TRANSPORT APPROACH FOR CLASS-AWARE UDA

4.1 PROBLEM FORMULATION

We consider the standard setting of unsupervised domain adaptation in which we have a labeled dataset $\mathbb{D}^S = \left\{\left(\mathbf{x}_i^S, y_i^S\right)\right\}_{i=1}^{N_S}$ from a source domain and an unlabeled dataset $\mathbb{D}^T = \left\{\mathbf{x}_i^T\right\}_{i=1}^{N_T}$ from a target domain. We assume that data examples $\mathbf{x}_i^S, \mathbf{x}_i^T \in \mathbb{R}^d$ and the categorical labels $y_i^S \in \{1, 2, ..., M\}$ where M is the number of classes. For the sake of notion simplification, we overload $\mathbb{D}^S$ and $\mathbb{D}^T$ to represent the empirical joint distributions of the source and target domains. We denote $\mathbb{P}^S$ and $\mathbb{P}^T$ as the data distributions of the source and target domains respectively. Moreover, given a class m, we further denote $\mathbb{P}_m^S$ as the m-th class-conditional distribution of the source domain (i.e., the distribution with the density function $p^S\left(\mathbf{x} \mid y = m\right)$).

4.2 MOTIVATION

For our proposed approach, we consider an OT distance of two discrete distributions. The first one is the discrete distribution whose atoms are the target examples $\mathbf{x}^T$ (i.e., $\mathbf{x}_i^1 = \mathbf{x}_i^T$ in Eq. (1)), while the second one is the discrete distribution whose atoms are the source class-conditional distributions $\mathbb{P}_m^S$ (i.e., $\mathbf{x}_j^2 = \mathbb{P}_m^S$ in Eq. (1)). The cost $c\left(\mathbf{x}_i^T, \mathbb{P}_m^S\right)$ is defined as the negative log likelihood $-\log p_m^S\left(\mathbf{x}_i^T\right) = -\log p^S\left(\mathbf{x}_i^T \mid y = m\right)$. Hence, if a target sample $\mathbf{x}_i^T$ is more likely to be a sample from $\mathbb{P}_m^S$, the log likelihood $\log p_m^S\left(\mathbf{x}_i^T\right)$ is higher, meaning that the cost $c\left(\mathbf{x}_i^T, \mathbb{P}_m^S\right) = -\log p_m^S\left(\mathbf{x}_i^T\right)$ becomes smaller. As shown in Figure 1, by examining the OT distance between two aforementioned distributions, we aim to find the best match between a given target sample $\mathbf{x}_i^T$ and a source class-conditional distribution $\mathbb{P}_m^S$.

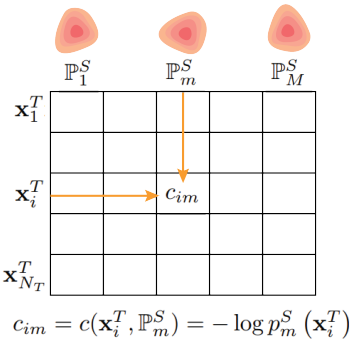


Figure 1: We consider the OT distance between two distributions: the first one has atoms as the target examples $\mathbf{x}^T$ and the second one has atoms as the class-conditional distributions $\mathbb{P}_m^S$. The cost function $c(\mathbf{x}_i^T, \mathbb{P}_m^S) = -\log p_m^S\left(\mathbf{x}_i^T\right) = -\log p^S\left(\mathbf{x}_i^T \mid y = m\right)$.

4.3 DISTRIBUTIONAL OPTIMAL TRANSPORT

We define $\mathcal{P}^S = \sum_{m=1}^{M} \pi_m \delta_{\mathbb{P}_m^S}$, where δ is the Dirac delta distribution and the mixing proportion $\boldsymbol{\pi} \in \Delta_M := \left\{\boldsymbol{\alpha} \in \mathbb{R}^M : \boldsymbol{\alpha} \geq \mathbf{0} \text{ and } \|\boldsymbol{\alpha}\|_1 = 1\right\}$ with the number of

classes M. Obviously, $\mathcal{P}^S$ is a discrete distribution of distributions wherein $\mathcal{P}^S$ takes $\mathbb{P}^S_m$ with the probability π_m. As mentioned in the motivation section, we now examine an OT distance between $\mathbb{P}^T$ and $\mathcal{P}^S$, we aim at matching target examples to the source class-conditional distributions in which a target example is absolutely guided to match the source class-conditional distribution corresponding to its ground-truth label.

In the sequel, we inspect an OT distance between $\mathbb{P}^T$ and $\mathcal{P}^S$ in which we define the cost $c\left(\mathbf{x}_i, \mathbb{P}^S_m\right)$ to match a target sample $\mathbf{x}_i$ to $\mathbb{P}^S_m$ as $-\log p^S_m\left(\mathbf{x}_i\right)$. Let us denote $A = [a_{im}] \in \mathbb{R}^{N_T \times M}$ as the transportation matrix wherein a_{im} represents the probability to match or transport $\mathbf{x}_i$ to $\mathbb{P}^S_m$. The OT distance between $\mathbb{P}^T$ and $\mathcal{P}^S$ w.r.t. the cost function c and the mixing proportion $\boldsymbol{\pi}$ is defined as:

$$\mathcal{W}_{c,\boldsymbol{\pi}}\left(\mathbb{P}^T, \mathcal{P}^S\right) = \min_A \Bigg\{ \sum_{i=1}^{N_T}\sum_{m=1}^{M} a_{im} c\left(\mathbf{x}_i, \mathbb{P}^S_m\right) : \sum_{m=1}^{M} a_{im} = \frac{1}{N_T}, \sum_{i=1}^{N_T} a_{im} = \pi_m \Bigg\}. \tag{2}$$

Similar to other DA works [Pan et al., 2008, Tzeng et al., 2015, Long et al., 2017], we employ a feature extractor G to map both source and target examples to a latent space. We denote $\mathbb{Q}^S, \mathbb{Q}^T, \mathbb{Q}^S_m$, and $\mathcal{Q}^S$ as the corresponding distributions over the latent space induced by $\mathbb{P}^S, \mathbb{P}^T, \mathbb{P}^S_m$, and $\mathcal{P}^S$ via the feature extractor G. The OT distance in Eq. (2) is rewritten as:

$$\mathcal{W}_{c,\boldsymbol{\pi}}\left(\mathbb{Q}^T, \mathcal{Q}^S\right) = \min_A \Bigg\{ \sum_{i=1}^{N_T}\sum_{m=1}^{M} a_{im} c\left(G\left(\mathbf{x}_i\right), \mathbb{Q}^S_m\right) : \sum_{m=1}^{M} a_{im} = \frac{1}{N_T}, \sum_{i=1}^{N_T} a_{im} = \pi_m \Bigg\}. \tag{3}$$

To encourage the target examples $G\left(\mathbf{x}_i\right)$ to move towards proper class regions of the source domain, we propose solving the following optimization problem (OP):

$$\min_{G,\boldsymbol{\pi}} \mathcal{W}_{c,\boldsymbol{\pi}}\left(\mathbb{Q}^T, \mathcal{Q}^S\right). \tag{4}$$

With $c\left(G\left(\mathbf{x}_i\right), \mathbb{Q}^S_m\right) = -\log p^S_m\left(\mathbf{x}_i\right)$, minimizing the OT distance in Eq. (4) encourages the target example $G\left(\mathbf{x}_i\right)$ to move towards a $\mathbb{Q}^S_k\ (1 \le k \le M)$ with a high likelihood and $\mathbf{a}_i = [a_{im}]_m$ inspired to be close to the corresponding scaled one-hot vector $\frac{1}{N_T}\mathbf{1}_k$. Here we denote $\mathbf{1}_k$ as the one-hot vector with the k-th element being one.

5 CYCLE CLASS CONSISTENCY FRAMEWORK

5.1 COST FUNCTION AND KNOWLEDGE DISTILLATION

To define the cost function $c\left(G\left(\mathbf{x}_i\right), \mathbb{Q}^S_m\right)$ in Eq. (4), we build a classifier h^S over the latent space, and rely on its output to compute the cost values. This classifier is first trained using the labeled source dataset $\mathbb{D}^S = \left\{\left(\mathbf{x}^S_i, y^S_i\right)\right\}_{i=1}^{N_S}$ by minimizing the empirical loss:

$$\mathcal{L}^{src} = \frac{1}{N_S}\sum_{i=1}^{N_S} CE\left(\sigma\left(h^S\left(G\left(\mathbf{x}_i\right)\right)\right), y^S_i\right), \tag{5}$$

where σ denotes a softmax function and CE represents a cross-entropy loss. Recap that given a target example $\mathbf{x}_i$, $c\left(G\left(\mathbf{x}_i\right), \mathbb{Q}^S_m\right)$ captures the matching extent of $G\left(\mathbf{x}_i\right)$ and the class-conditional distribution $\mathbb{Q}^S_m$. Therefore, we can reasonably define $c\left(G\left(\mathbf{x}_i\right), \mathbb{Q}^S_m\right) = -\log \sigma_m\left(h^S\left(G\left(\mathbf{x}_i\right)\right)\right)$ (i.e., $\sigma_m\left(h^S\left(G\left(\mathbf{x}_i\right)\right)\right)$ is the predicted probability of $\mathbf{x}_i$ belonging to class m by classifier h^S).

However, we find that h^S is a well-trained classifier on the source domain, and can generalize poorly on the target domain due to the data and label shifts. Therefore, instead of using only one classifier trained to work well on both domains, we leverage knowledge distillation [Hinton et al., 2015, Tian et al., 2020, Tejankar et al., 2021] which includes the two-network architecture, a teacher $h^T$and a student h^S. The teacher h^T aims to be an expert on the target domain, while the student h^S, which classifies accurately on the source domain, is also able to generalize on the target domain via distilling knowledge from its teacher. When the generalization ability of h^S is improved, the cost $c\left(G\left(\mathbf{x}_i\right), \mathbb{Q}^S_m\right)$ is computed more accurately to solve the OP in Eq. (3). Inspired by the work of Hinton et al. [2015], we perform knowledge distillation from the teacher h^T to the student h^S in the target domain by minimizing a distillation loss $\mathcal{L}^{dl}$ w.r.t. a temperature softmax function:

$$\mathcal{L}^{dl} = \frac{1}{N_T}\sum_{i=1}^{N_T} CE\left(\sigma\left(\frac{h^S\left(G\left(\mathbf{x}_i\right)\right)}{\tau}\right), \sigma\left(\frac{h^T\left(G\left(\mathbf{x}_i\right)\right)}{\tau}\right)\right), \tag{6}$$

where τ is a temperature parameter. When setting $\tau > 1$, the teacher and student's predictions become softer, from which the student can capture "dark knowledge" [Hinton et al., 2015] from the teacher and effectively mimic the teacher's behaviour.

The student h^S is now trained well in the source domain via Eq. (5), and is possible to generalize on the target domain via Eq. (6). To achieve this good generalization capability,

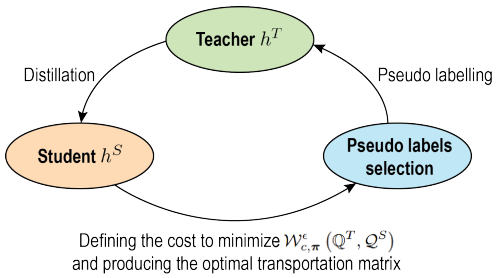


Figure 2: The proposed cycle class consistency framework.

we need to produce a teacher h^T that is with good classification performance on the target domain. To this end, we propose minimizing a cross-entropy loss between the teacher's prediction and pseudo labels computed via the optimal transportation matrix A^* after solving Eq. (3):

$$\mathcal{L}^{pl} = \frac{1}{n_T}\sum_{i=1}^{n_T} CE\left(\sigma\left(h^T\left(G\left(\mathbf{x}_i\right)\right)\right), \hat{y}_i^T\right), \quad (7)$$

where $\hat{y}^T$ are pseudo labels for unlabeled target samples. It is worth noting that only a subset of target samples with high-confidence pseudo labels is selected (i.e., $n_T < N_T$). In the next section, we discuss on how to compute these pseudo labels and our framework.

5.2 PSEUDO-LABEL SELECTION AND OUR FRAMEWORK

We now introduce the strategy to produce pseudo labels for unlabeled target samples. Let us return to the Eq. (3) where directly solving this OP is computationally expensive. Hence, we instead use an entropic regularized version to minimize:

$$\mathcal{W}^{\epsilon}_{c,\boldsymbol{\pi}}\left(\mathbb{Q}^T, \mathcal{Q}^S\right) = \min_{A}\Bigg\{\sum_{i=1}^{N_T}\sum_{m=1}^{M} a_{im}c\left(G\left(\mathbf{x}_i\right), \mathbb{Q}_m^S\right) - \epsilon H(A) : \sum_{m=1}^{M} a_{im} = \frac{1}{N_T}, \sum_{i=1}^{N_T} a_{im} = \pi_m\Bigg\}, \quad (8)$$

where $H(A) := -\sum_{i=1}^{N_T}\sum_{m=1}^{M} a_{im}\log a_{im}$ denotes an entropy of the transportation matrix A, and ϵ is the regularization rate. During the training, we use Sinkhorn algorithm [Cuturi, 2013] to solve this OP and achieve A^* at every mini-batch. Interestingly, the solution of Eq. (8) also provides us $\sum_{m=1}^{M} a^*_{im} = \frac{1}{N_T}$ or in other words, $N_T\sum_{m=1}^{M} a^*_{im} = 1$. Hence, we can define the pseudo label $\hat{y}_i^T := N_T a_i^*$ for a given target sample $\mathbf{x}_i$ and it satisfies $\sum_{m=1}^{M}\hat{y}^T_{im} = N_T\sum_{m=1}^{M} a^*_{im} = 1$. The definition of $\hat{y}_i^T$ is then used for minimizing $\mathcal{L}^{pl}$ in Eq. (7).

One problem when choosing $\hat{y}_i^T := N_T a_i^*$ is that the performance of the teacher h^T can be reduced if some pseudo labels are incorrect, especially at the beginning of the training due to the data and label shifts between the source and target domains. This issue also influences the distillation process since we aim to build a well-classified teacher h^T on the target domain to transfer some of its aspects (e.g, its "dark knowledge") to the student h^S. To avoid this problem, inspired by Yang et al. [2021], we propose only selecting highly confident pseudo labels (i.e., pseudo labels whose entropies are less than a threshold) using an entropy-based selection method. The OP in Eq. (7) is now minimized w.r.t. the weights w_i:

$$\mathcal{L}^{pl}_{w} = \frac{1}{n_T}\sum_{i=1}^{n_T} w_i CE\left(\sigma\left(h^T\left(G\left(\mathbf{x}_i\right)\right)\right), \hat{y}_i^T\right), \quad (9)$$

where $w_i = \mathbb{I}_{\left\{H\left(\hat{y}_i^T\right) < H_\rho\right\}}$ with $\mathbb{I}_C$ representing the indicator function for a statement C (i.e., $\mathbb{I}_C$ returns 1 iff C is true), $H\left(\hat{y}_i^T\right) := -\sum_{m=1}^{M} a_{im}\log a_{im}$ is the entropy of a pseudo label $\hat{y}_i^T$ w.r.t. a target example $\mathbf{x}_i$, and the threshold H_ρ denotes the ρ-th percentile of $H\left(\hat{y}_i^T\right)$.

Additionally, when training our COOK, at each iteration, we sample a mini-batch of target examples and consider $\mathbb{Q}^T$ as the distribution of latent representations corresponding to this mini-batch. Therefore, N_T in Eq. (8) is replaced by the batch size and the threshold H_ρ denotes the ρ-th percentile of $H\left(\hat{y}_i^T\right)$ in the mini-batch.

Finally, we present our framework in Figure 2 which includes three main steps: (i) the teacher is encouraged to be an expert on the target domain using the pseudo labeling technique; (ii) the teacher transfers its knowledge to the student via a distillation process to support the student to generalize well on the target domain; and (iii) the predicted probabilities of the student classifier are utilized for minimizing $\mathcal{W}^{\epsilon}_{c,\boldsymbol{\pi}}\left(\mathbb{Q}^T, \mathcal{Q}^S\right)$ using Sinkhorn algorithm, and offering the optimal transportation matrix A^* to compute pseudo labels. The pseudo labels with low entropies are selected to train the teacher at the first step. This process forms a closed cycle in which target samples are confidently moved towards corresponding source class-conditional distributions $\mathbb{Q}_m^S$ under the consistently cyclic guidance of the key factors including the distributional optimal transport and knowledge distillation, which motivates us to propose our COOK.

5.3 TRAINING PROCEDURE OF COOK

To strengthen h^S for providing better predictions and accelerating matching target samples $\mathbf{x}^T$ to source class-conditional distributions $\mathbb{Q}_m^S$, we enforce the clustering assumption to h^S. Inspired by applying clustering assumption in domain adaptation works [Shu et al., 2018, Kumar et al., 2018], we employ Virtual Adversarial Training (VAT) [Miyato et al., 2019] in conjunction with minimizing entropy [Grandvalet and Bengio, 2005] of the prediction of

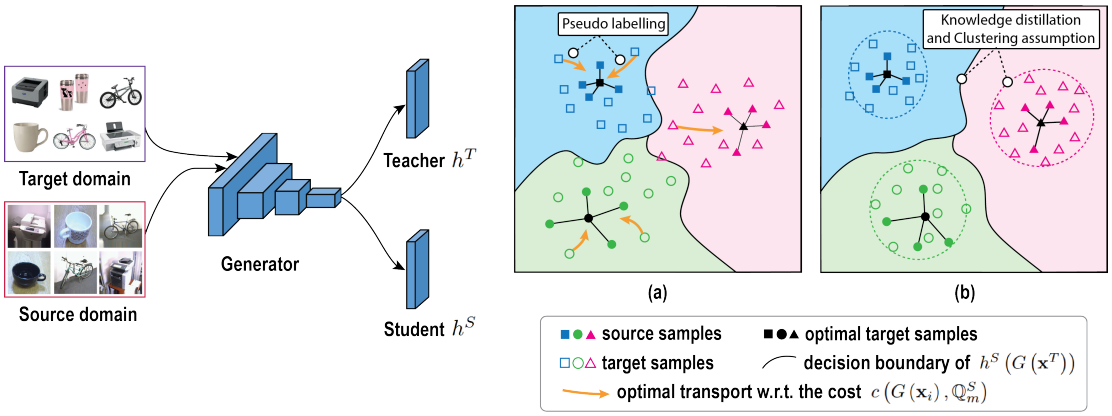


Figure 3: The overall architecture of our proposed method where G is a weight-sharing generator for mapping the source and target data into the latent space. The teacher h^T and the student h^S act in a cyclic process as described in Figure 2 where we apply pseudo labelling, knowledge distillation and enforce clustering assumption: (a) when minimizing pseudo labelling loss $\mathcal{L}_w^{pl}$, target samples are encouraged to move towards the corresponding source class-conditional distributions; (b) minimizing distillation loss $\mathcal{L}^{dl}$ pushes the target samples closer to the source samples due to the distillation process between predictions of the teacher and student classifiers. While minimizing $\mathcal{L}^{clus}$ accelerates transporting target samples, achieves a strong clustering, improves local smoothness and achieves the good generalization ability of h^S on the target domain, from which the pseudo labels are selected with the high confidence.

$h^S\left(G\left(\mathbf{x}^T\right)\right)$. VAT is an effective technique to improve the local distribution robustness [Nguyen-Duc et al., 2022, Phan et al., 2022]. At first, given a target sample $\mathbf{x}$, a perturbation of $\mathbf{x}$, which is $\mathbf{x}'$ that makes the student classifier h^S give a different prediction from $\mathbf{x}$ is chosen. And then h^S is enforced to predict the same label for $\mathbf{x}$ and $\mathbf{x}'$. As a result, the decision boundary of h^S is pushed away from the target sample $\mathbf{x}$, which achieves a better generalization ability for h^S on the target domain.

$$\mathcal{L}^{clus} = \mathcal{L}^{ent} + \mathcal{L}^{vat}, \tag{10}$$

where with H to be the entropy, we have defined:

$\mathcal{L}^{ent} = \mathbb{E}_{\mathbb{P}^T}\left[H\left(\sigma\left(h^S\left(G\left(\mathbf{x}\right)\right)\right)\right)\right]$,

$\mathcal{L}^{vat} = \mathbb{E}_{\mathbb{P}^T}\Big[\max_{\mathbf{x}':\|\mathbf{x}'-\mathbf{x}\|<\theta} D_{KL}\Big(\sigma\left(h^S\left(G\left(\mathbf{x}\right)\right)\right),$

$\sigma\left(h^S\left(G\left(\mathbf{x}'\right)\right)\right)\Big)\Big]$, where D_{KL} denotes the Kullback-Leibler divergence and θ is a hyperparameter set to a very small positive number.

The final optimization problem of our COOK for finding h^S, h^T and G is as follows:

$$\min_{h^S,h^T,G}\left\{\mathcal{L}^{src} + \alpha\mathcal{L}^{dl} + \beta\mathcal{L}_w^{pl} + \gamma\mathcal{L}^{clus}\right\}, \tag{11}$$

where $\alpha, \beta, \gamma > 0$ are trade-off parameters. Conveniently, the cyclic process in Figure 2 is operated synchronously by simultaneously updating h^S, h^T and G during the training. Finally, we present the key steps of our COOK in Algorithm 1 and the overall architecture and motivation of component losses are depicted in Figure 3.

Algorithm 1 Pseudocode for training our proposed COOK.

Input: A source batch $\mathcal{B}^S = \left\{\left(\mathbf{x}_i^S, y_i^S\right)\right\}_{i=1}^b$, a target batch $\mathcal{B}^T = \left\{\mathbf{x}_j^T\right\}_{j=1}^b$ (b denotes the batch size).

Output: Classifiers h^{S*}, h^{T*}, generator G^*.

1: **for** number of training iterations **do**
2: Solve the OP in Eq. (8) using Sinkhorn algorithm to find A^*.
3: Compute $\hat{y_i}^T$ in Eq. (9) based on A^*.
4: Compute w_i in Eq. (9) based on H_ρ.
5: Update h^S, h^T and G according to Eq. (11).
6: **end for**

6 EXPERIMENTS

In this section, we conduct experiments on benchmark datasets including *Office-31*, *Office-Home*, and *ImageCLEF-DA* to compare with existing baselines, especially OT-based and class-aware UDA methods.

6.1 DATASETS

Office-31 [Saenko et al., 2010] is a well-known public dataset used for UDA. It consists of three domains including Amazon (**A**), Webcam (**W**) and Dslr (**D**) with 31 common classes and 4,110 images in total.

Office-Home [Venkateswara et al., 2017] is another and

more challenging dataset for UDA which contains images from four different domains, namely Artistic (**Ar**), Clip Art (**Cl**), Product (**Pr**) and Real-world images (**Re**). This dataset consists of around 15,588 images in total with 65 object categories in office and home scenes.

ImageCLEF-DA [Caputo et al., 2014] includes three domains including Caltech-256 (**C**), ImageNet ILSVRC 2012 (**I**), and Pascal VOC 2012 (**P**), each of which has 12 classes with 50 images per class.

6.2 IMPLEMENTATION DETAILS

In the experiments on the *Office-31*, *Office-Home* and *ImageCLEF-DA* datasets, we use the extracted features (2048 dimensions) from ResNet-50 [He et al., 2016]. The generator includes a fully connected layer that outputs 256 dimensions. We use the same architecture for the student and teacher networks which consists of a fully connected layer for each network.

Some hyperparameters substantially contributes to model performance, namely the temperature τ in Eq. (6), and the percentile ρ in Eq. (9). As suggested in the ablation study, we choose $\tau = 10.0$ to effectively activate the knowledge distillation process from the teacher to the student. The percentile ρ is important to measure how well the student h^S can generalize on the target domain. We empirically find that $\rho = 20$ or in other words, choosing the 20-th percentile of $H\left(\hat{y}_i^T\right)$ is appropriate to select high-confidence pseudo labels. Additionally, setting ϵ less than or equal to 0.1 can achieve better performance and we set ϵ to 0.1. We also select the trade-off parameters $\alpha = \beta = 1.0$ and $\gamma = 0.1$ in our experiments as suggested in the ablation studies.

We apply Adam optimizer [Kingma and Ba, 2015] ($\beta_1 = 0.5, \beta_2 = 0.999$) with Polyak averaging [Polyak and Juditsky, 1992], and the learning rate is set to 10^{-4} for *Office-31* and *Office-Home*, and 5×10^{-5} for *ImageCLEF-DA*. For the baselines, we report the experimental results mentioned in the original papers. It is noticeable that in all experiments, we only train the feature extractor, and the performance of COOK can be further improved when fine-tuning the backbone ResNet-50 is conducted.

6.3 RESULT AND DISCUSSION

We compare our COOK with the standard baseline ResNet-50 [He et al., 2016] and existing works including DAN [Long et al., 2015], DANN [Ganin and Lempitsky, 2015], RTN [Long et al., 2016], iCAN [Zhang et al., 2018], CDAN-E [Long et al., 2018], CDAN-BSP [Chen et al., 2019], CDAN-T [Wang et al., 2019a], TPN [Pan et al., 2019], rRevGrad+CAT [Deng et al., 2019], CADA-P [Kurmi et al., 2019], SymNets [Zhang et al., 2019], especially class-aware DA and OT-based methods, namely RADA [Wang et al., 2019b], CAN [Kang et al., 2019], DeepJDOT [Damodaran et al., 2018], ETD [Li et al., 2020], and RWOT [Xu et al., 2020].

Table 1: Mean accuracy (%) on Office-31 for unsupervised domain adaptation (ResNet-50).

Method	A→W	A→D	D→W	W→D	D→A	W→A	Avg
ResNet-50	68.4	68.9	96.7	99.3	62.5	60.7	76.1
DAN	80.5	78.6	97.1	99.6	63.6	62.8	80.4
DANN	82.0	79.7	96.9	99.1	68.2	67.4	82.2
RTN	84.5	77.5	96.8	99.4	66.2	64.8	81.6
iCAN	92.5	90.1	98.8	**100.0**	72.1	69.9	87.2
CDAN-E	94.1	92.9	98.6	**100.0**	71.0	69.3	87.7
CDAN-BSP	93.3	93.0	98.2	**100.0**	73.6	72.6	88.5
CDAN-T	95.7	94.0	98.7	**100.0**	73.4	74.2	89.3
TPN	91.2	89.9	97.7	99.5	70.5	73.5	87.1
rRevGrad+CAT	94.4	90.8	98.0	**100.0**	72.2	70.2	87.6
SymNets	90.8	93.9	98.8	**100.0**	74.6	72.5	88.4
DeepJDOT	88.9	88.2	98.5	99.6	72.1	70.1	86.2
ETD	92.1	88.0	**100.0**	**100.0**	71.0	69.3	86.2
RWOT	**95.1**	94.5	99.5	**100.0**	77.5	77.9	90.8
RADA	91.5	90.7	98.9	**100.0**	71.5	71.3	87.3
CAN	94.5	95.0	99.1	99.8	78.0	77.0	90.6
COOK	**95.1**	**96.2**	98.3	99.9	**88.7**	**86.2**	**94.1**

The results trained on *Office-31* are reported in Table 1. In general, our proposed method achieves high results with four transfer tasks greater than 95%. Except for the transfer tasks **D**→**W** and **W**→**D**, our model significantly outperforms others on almost adaptation tasks, and obtain 94.1% on average, which is a 3.5% increase compared to the runner-up result. It is worth noting that our COOK outperforms the baselines by a large margin on challenging tasks, e.g., a 10.7% increase on **D**→**A** and **W**→**A** with a 9.2% improvement, in which the background of the training images between the two domains are totally dissimilar.

We present the results trained on *Office-Home* in Table 2. In this dataset, our COOK surpasses 7 over 12 transfer tasks compared with the baselines and achieves the best performance, making a 2.8% improvement on average. More specifically, our model sees a remarkable improvement on more challenging adaptation tasks, namely **Ar**→**Pr** (3.6%), **Cl**→**Pr** (7.6%), **Cl**→**Re** (4.1%).

We further evaluate our COOK on *ImageCLEF-DA* and report the classification accuracy in Table 3. Our COOK outperforms 4 over 6 transfer tasks with an average accuracy of 90.7%, compared to ETD and RWOT with 89.7% and 90.3%, respectively.

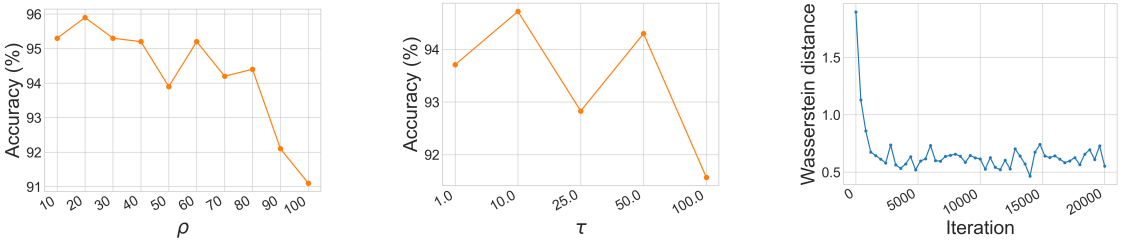
6.4 ANALYSIS

6.4.1 Hyperparameter Sensitivity and Quantitative Evaluation

We conduct experiments to evaluate hyperparameter sensitivity and quantitative result for our proposed COOK in Figure 4. Figure 4a experiences a decrease of the model per-

Table 2: Mean accuracy (%) on Office-Home for unsupervised domain adaptation (ResNet-50).

Method	Ar→Cl	Ar→Pr	Ar→Re	Cl→Ar	Cl→Pr	Cl→Re	Pr→Ar	Pr→Cl	Pr→Re	Re→Ar	Re→Cl	Re→Pr	Avg
ResNet-50	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DAN	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
DANN	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
SymNets	47.7	72.9	78.5	64.2	71.3	74.2	64.2	48.8	79.5	**74.5**	52.6	82.7	67.6
CDAN-E	50.7	70.6	76.0	57.6	70.0	70.0	57.4	50.9	77.3	70.9	56.7	81.6	65.8
CDAN-BSP	52.0	68.6	76.1	58.0	70.3	70.2	58.6	50.2	77.6	72.2	**59.3**	81.9	66.3
CDAN-T	50.2	71.4	77.4	59.3	72.7	73.1	61.0	**53.1**	79.5	71.9	59.0	82.9	67.6
DeepJDOT	48.2	69.2	74.5	58.5	69.1	71.1	56.3	46.0	76.5	68.0	52.7	80.9	64.3
ETD	51.3	71.9	**85.7**	57.6	69.2	73.7	57.8	51.2	79.3	70.2	57.5	82.1	67.3
RWOT	**55.2**	72.5	78.0	63.5	72.5	75.1	60.2	48.5	78.9	69.8	54.8	82.5	67.6
COOK	53.0	**76.5**	81.8	**65.5**	**80.3**	**79.2**	**64.5**	51.8	**82.4**	71.3	54.2	**83.9**	**70.4**



(a) The changes of ρ-percentile. (b) Study of twisting τ. (c) Values of $\mathcal{W}^{\epsilon}_{c,\boldsymbol{\pi}}\left(\mathbb{Q}^T, \mathcal{Q}^S\right)$.

Figure 4: Ablation studies of our proposed method on the transfer task **A**→**W**.

Table 3: Mean accuracy (%) on ImageCLEF-DA for unsupervised domain adaptation (ResNet-50).

Method	I→P	P→I	I→C	C→I	C→P	P→C	Avg
RTN	75.6	86.8	95.3	86.9	72.7	92.2	84.9
iCAN	79.5	89.7	94.7	89.9	78.5	92.0	87.4
CDAN-E	77.7	90.7	97.7	91.3	74.2	94.3	87.7
CDAN-T	78.3	90.8	96.7	92.3	78.0	94.8	88.5
SymNets	80.2	93.6	97.0	93.4	78.7	96.4	89.9
CADA-P	78.0	90.5	96.7	92.0	77.2	95.5	88.3
DeepJDOT	77.7	90.6	95.1	88.5	75.3	94.3	86.9
ETD	81.0	91.7	97.9	93.3	79.5	95.0	89.7
RWOT	**81.5**	93.1	**98.0**	92.8	79.3	96.8	90.3
RADA	79.2	92.4	97.5	91.1	76.6	95.3	88.7
COOK	80.1	**95.5**	97.0	**95.9**	**79.1**	**96.3**	**90.7**

Table 4: Accuracy (%) of ablation study on ImageCLEF-DA.

$\mathcal{L}^{src}$	$\mathcal{L}^{pl}_{w}$	$\mathcal{L}^{dl}$	$\mathcal{L}^{clus}$	I→P	P→I	I→C	C→I	C→P	P→C	Avg
✓	✓			75.9	86.1	93.9	89.0	74.4	87.4	84.5
✓	✓	✓		76.4	86.6	93.9	89.9	76.0	91.2	85.7
✓	✓		✓	78.6	91.0	95.9	92.9	77.9	95.9	88.7
✓		✓	✓	78.5	90.9	96.7	93.3	**79.6**	95.0	89.0
✓	✓	✓	✓	**80.1**	**95.5**	**97.0**	**95.9**	79.1	**96.3**	**90.7**

Table 5: Results (%) on different training strategies.

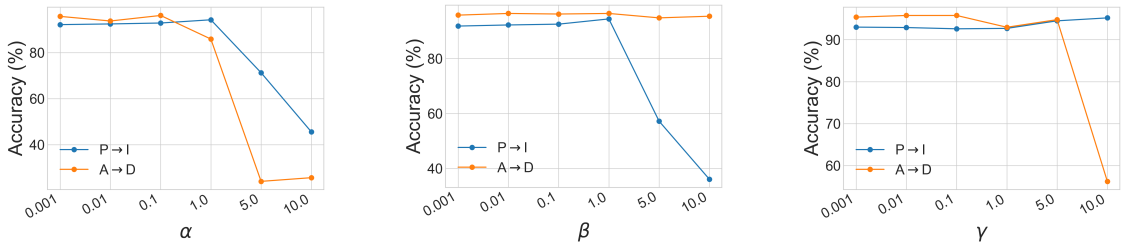
Methods	A→W	A→D	D→W	W→D	D→A	W→A	Avg
Without KD	93.8	94.6	97.8	99.2	86.4	86.1	93.0
With KD	**95.1**	**96.2**	**98.3**	**99.9**	**88.7**	**86.2**	**94.1**

formance when twisting ρ in Eq. (9). Our proposed COOK works well with ρ from 10 to 40. Relying on this investigation, we pick $\rho = 20$ in our experiments. Similarly, Figure 4a shows results with the changes of τ. We search τ in the grid of $\{1.0, 10.0, 25.0, 50.0, 100.0\}$ and find that setting $\tau = 10.0$ achieves the best performance to perform knowledge distillation. Furthermore, we investigate the Wasserstein distance $\mathcal{W}^{\epsilon}_{c,\boldsymbol{\pi}}\left(\mathbb{Q}^T, \mathcal{Q}^S\right)$ in Figure 4c, which sees a reduction during the training. This result shows the success of transporting target samples to their corresponding source class-conditional distributions.

We further evaluate the effects of the trade-off parameters α, β, γ on model performance by twisting their values. Figure 5 shows results when we search α, β and γ in the grid of $\{0.001, 0.01, 0.1, 1.0, 5.0, 10.0\}$ and report the test accuracy on two transfer tasks **P** → **I** (*ImageCLEF-DA*) and **A**→**D** (*Office-31*). The results show that the model yields the stable performance when α, β, γ from 0.001 to 1.0. We find that our COOK can achieve high performance when $\alpha = \beta = 1.0$ and $\gamma = 0.1$, hence we suggest picking these values on most of our experiments.

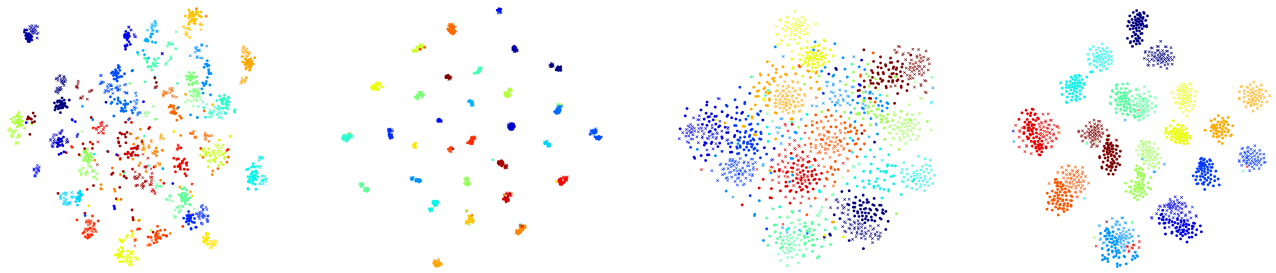
6.4.2 Effect of Losses

We investigate the effectiveness of the pseudo labelling loss $\mathcal{L}^{pl}_{w}$, the distillation loss $\mathcal{L}^{dl}$, and the clustering assumption loss $\mathcal{L}^{clus}$ in Eq. (11). The experimental results are described in Table 4, which shows that all component losses contribute to the model performance since they participate



(a) α (b) β (c) γ

Figure 5: Analysis of hyperparameter sensitivity of α, β and γ on transfer tasks **P** $\rightarrow$ **I** and **A**$\rightarrow$**D**.



(a) ResNet (b) COOK (c) ResNet (d) COOK

Figure 6: The t-SNE visualization of **A**$\rightarrow$**W** (Figure a, b) and **P**$\rightarrow$**C** (Figure c, d) tasks with label and domain information. Each color denotes a class while the circle and cross markers represent the source and target data respectively.

in the cyclic process and support to match target samples to the corresponding source regions. It is noticeable that the model performance is the best when all component losses are activated and participate in the training process.

6.4.3 Effect of Knowledge Distillation

We further testify the contribution of KD to our proposed method in two different scenarios: *Without KD* and *With KD*. For *Without KD* setting, we deploy a model where h^S and h^T are weight-sharing networks and train this model using the final optimization problem where $\mathcal{L}^{dl} = 0$. We compare the *Without KD* setting with our architecture COOK (a.k.a. *With KD*) and report the accuracy score in Table 5. The results show that our COOK with KD outperforms that without KD by nearly 1%, which demonstrates the effectiveness of KD for our framework.

6.4.4 Feature Visualization

We select transfer tasks **A**$\rightarrow$**W** (*Office-31*) and **P**$\rightarrow$**C** (*ImageCLEF-DA*) tasks to visualize their representation in the latent space using *t*-SNE [van der Maaten and Hinton, 2008]. The visualizations in Figure 6a and 6c show that after going through the backbone model ResNet-50, there is still a mismatch between the source and target distributions due to the data and label shifts. However, our proposed COOK (see Figure 6b and 6d) is trained to transport target samples to source samples, which closes this gap and achieves better alignment between the target and the source samples.

7 CONCLUSION

In this paper, we develop a novel framework for class-aware unsupervised domain adaptation. In particular, our proposed method is based on the proposed distributional OT which quantifies an OT distance between a distribution of target data and the source class-conditional distributions. To efficiently train our model with the proposed distributional OT, we develop a novel model operating in a cyclic process. By incorporating knowledge distillation and pseudo labelling technique into this process, our proposed COOK effectively tackles the data and label shifts problem by transporting the target samples to the corresponding source class-conditional distributions in a class-aware manner. The experimental results show that COOK outperforms existing UDA methods, especially the class-aware and OT-based ones on the benchmark datasets.

Acknowledgements

This work was supported by the US Air Force grant FA2386-21-1-4049.

Global-Local Regularization Via Distributional Robustness

Hoang Phan[◇] **Trung Le**[◇†] **Trung Phung**[+] **Anh Bui**[†] **Nhat Ho**[‡] **Dinh Phung**[◇†]
VinAI Research[◇] Monash University, Australia[†] Johns Hopkins University[+] University of Texas, Austin[‡]

Abstract

Despite superior performance in many situations, deep neural networks are often vulnerable to adversarial examples and distribution shifts, limiting model generalization ability in real-world applications. To alleviate these problems, recent approaches leverage distributional robustness optimization (DRO) to find the most challenging distribution, and then minimize loss function over this most challenging distribution. Regardless of having achieved some improvements, these DRO approaches have some obvious limitations. First, they purely focus on local regularization to strengthen model robustness, missing a global regularization effect that is useful in many real-world applications (e.g., domain adaptation, domain generalization, and adversarial machine learning). Second, the loss functions in the existing DRO approaches operate in only the most challenging distribution, hence decouple with the original distribution, leading to a restrictive modeling capability. In this paper, we propose a novel regularization technique, following the veins of Wasserstein-based DRO framework. Specifically, we define a particular joint distribution and Wasserstein-based uncertainty, allowing us to couple the original and most challenging distributions for enhancing modeling capability and enabling both local and global regularizations. Empirical studies on different learning problems demonstrate that our proposed approach significantly outperforms the existing regularization approaches in various domains.

1 Introduction

As the Wasserstein (WS) distance is a powerful and convenient tool of measuring closeness between distributions, Wasserstein Distributional Robustness (WDR) has been one of the most widely-used variants of DR. Here we consider a generic Polish space S endowed with a distribution $\mathbb{P}$. Let $r : S \to \mathbb{R}$ be a real-valued (risk) function and $c : S \times S \to \mathbb{R}_+$ be a cost function. Distributional robustness setting aims to find the distribution $\tilde{\mathbb{P}}$ in the vicinity of $\mathbb{P}$ and maximizes the risk in the expectation form (Blanchet and Murthy, 2019; Sinha et al., 2018):

$$\sup_{\tilde{\mathbb{P}}:\mathcal{W}_c\left(\mathbb{P},\tilde{\mathbb{P}}\right)<\epsilon} \mathop{\mathbb{E}}_{\tilde{Z}\sim\tilde{\mathbb{P}}}\left[r\left(\tilde{Z}\right)\right], \tag{1}$$

where $\epsilon > 0$ and $\mathcal{W}_c\left(\mathbb{P},\tilde{\mathbb{P}}\right) := \inf_{\gamma\in\Gamma\left(\mathbb{P},\tilde{\mathbb{P}}\right)} \int c d\gamma$ denotes an optimal transport (OT) or a WS distance with the set of couplings $\Gamma\left(\mathbb{P},\tilde{\mathbb{P}}\right)$ whose marginals are $\mathbb{P}$ and $\tilde{\mathbb{P}}$.

Direct optimization over the set of distributions $\tilde{\mathbb{P}}$ is often computationally intractable except in limited cases, we thus seek to cast this problem into its dual form. With the assumption that $r \in L^1(\mathbb{P})$ is upper semi-continuous and the cost c is a non-negative and continuous function satisfying $c(Z,\tilde{Z}) = 0$ iff $Z = \tilde{Z}$, (Blanchet and Murthy, 2019; Sinha et al., 2018) showed the *dual* form for Eq. (1) is:

$$\inf_{\lambda\geq 0}\left\{\lambda\epsilon + \mathop{\mathbb{E}}_{Z\sim\mathbb{P}}\left[\sup_{\tilde{z}}\left\{r\left(\tilde{Z}\right) - \lambda c\left(\tilde{Z},Z\right)\right\}\right]\right\}. \tag{2}$$

When applying DR to the supervised learning setting, $\tilde{Z} = \left(\tilde{X},\tilde{Y}\right)$ is a pair of data/label drawn from $\tilde{\mathbb{P}}$ and r is the loss function (Blanchet and Murthy, 2019; Sinha et al., 2018). The fact that r engages only $\tilde{Z} = \left(\tilde{X},\tilde{Y}\right) \sim \tilde{\mathbb{P}}$ certainly restricts the modeling capacity of (2). The reasons are as follows. Firstly, for each anchor Z, the most challenging sample $\tilde{Z}$ is currently defined as the one maximizing $\sup_{\tilde{Z}}\left\{r(\tilde{Z}) - \lambda c(Z,\tilde{Z})\right\}$, where $r(\tilde{Z})$ is inherited from the primal form (1). Hence, it is not suitable to express the risk function r engaging both Z and $\tilde{Z}$ (e.g., Kullback-Leibler divergence $KL\left(p\left(\tilde{Z}\right)\|p(Z)\right)$ between the predictions for Z and $\tilde{Z}$ as in TRADES (Zhang et al., 2019)). Secondly, it is also *impossible* to inject a *global regularization term* involving a batch of samples $\tilde{Z}$ and Z.

Contribution. To empower the formulation of DR for efficiently tackling various real-world problems, in this work,

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

we propose a rich OT based DR framework, named ***G**lobal-**L**ocal **O**ptimal **T**ransport based **D**istributional **R**obustness* (GLOT-DR). Specifically, by designing special joint distributions $\mathbb{P}$ and $\tilde{\mathbb{P}}$ together with some constraints, our framework is applicable to a mixed variety of real-world applications, including domain generalization (DG), domain adaptation (DA), semi-supervised learning (SSL), and adversarial machine learning (AML).

Additionally, our GLOT-DR makes it possible for us to equip not only a *local regularization term* for enforcing a local smoothness and robustness, but also a *global regularization term* to impose a global effect targeting a downstream task. Moreover, by designing a specific WS distance, we successfully develop a closed-form solution for GLOT-DR without using the dual form in (Blanchet and Murthy, 2019; Sinha et al., 2018) (i.e., Eq. (2)).

Technically, our solution turns solving the inner maximization in the dual form (2) into sampling a set of challenging particles according to a local distribution, on which we can handle efficiently using Stein Variational Gradient Decent (SVGD) (Liu and Wang, 2016) approximate inference algorithm. Based on the general framework of GLOT-DR, we establish the settings for DG, DA, SSL, and AML and conduct experiments to compare our GLOT-DR to state-of-the-art baselines in these real-world applications. Overall, our contributions can be summarized as follows:

- We enrich the general framework of DR to make it possible for many real-world applications by enforcing both local and global regularization terms. We note that the global regularization term is crucial for many downstream tasks (see Section 3.1 for more details).
- We propose a closed-form solution for our GLOT-DR without involving the dual form in (Blanchet and Murthy, 2019; Sinha et al., 2018) (i.e., Eq. (2)). We note that the dual form (2) is *not computationally tractable* due to the minimization over λ.
- We conduct comprehensive experiments to compare our GLOT-DR to state-of-the-art baselines in DG, DA, SSL, and AML. The experimental results demonstrate the merits of our proposed approach and empirically prove that both of the introduced local and global regularization terms advance existing methods across various scenarios, including DG, DA, SSL, and AML.

2 Related Work

Distributional robustness (DR). DR is an attractive framework for improving machine learning models in terms of robustness and generalization. Its underlying idea is to find the *most challenging distribution* around a given distribution and then challenge a model with this distribution. To characterize the closeness of a distribution to a center distribution, either a f-divergence (Ben-Tal et al., 2013; Duchi et al., 2021, 2019; Miyato et al., 2015; Namkoong and Duchi, 2016) or Wasserstein distance (Blanchet et al., 2019; Gao and Kleywegt, 2016; Kuhn et al., 2019; Mohajerin Esfahani and Kuhn, 2015; Shafieezadeh-Abadeh et al., 2015) can be employed. Other works (Blanchet and Murthy, 2019; Sinha et al., 2018) developed a dual form for DR, opening the door to incorporate DR into the training of deep learning models.

Adversarial Robustness (AR). Neural networks are generally vulnerable to adversarial attacks, notably FGSM (Goodfellow et al., 2014), PGD (Madry et al., 2018), and Auto-Attack (Croce and Hein, 2020). Among various kinds of defense approaches, Adversarial Training (AT), originating in (Goodfellow et al., 2014), has drawn the most research attention. Given its effectiveness and efficiency, many variants of AT have been proposed with: (1) different types of adversarial examples (e.g., the worst-case examples (Goodfellow et al., 2014) or most divergent examples (Zhang et al., 2019)), (2) different searching strategies (e.g., non-iterative FGSM and Rand FGSM (Madry et al., 2018)), (3) additional regularization (e.g., adding constraints in the latent space (Bui et al., 2020; Zhang and Wang, 2019)). Inspired by the potential of DR, it has been applied to enhance model robustness in (Dong et al., 2020; Levine and Feizi, 2020; Miyato et al., 2018; Sinha et al., 2018; Nguyen-Duc et al., 2022; Bui et al., 2022; Le et al., 2022; Hoang et al., 2020).

Transfer Learning (TL). Domain adaptation (DA) and domain generalization (DG) are two typical settings in TL. As for domain adaptation, (Ganin et al., 2016; Li et al., 2020; Long et al., 2017a; Nguyen et al., 2022; Le et al., 2021; Nguyen et al., 2021b,c,a) aim at training a model based on a labeled source domain to adapt to an unlabeled target domain, while the works in DG (Balaji et al., 2018; Bousmalis et al., 2016; Li et al., 2017, 2018, 2019; Mancini et al., 2018; Phung et al., 2021) aim at training a model based on multiple labeled source domains to predict well on unseen target domains. Finally, in more recent work, it was leveraged with DG in (Zhao et al., 2020) and DA in (Wang et al., 2021).

3 Proposed Approach

In this section, we first introduce the GLOT-DR framework and provide the theoretical development in Section 3.1. Then Section 3.2 presents the general training procedure of our proposed approach, and the detailed formulations of scenarios are described in the remainder of this section.

3.1 Our Framework

We propose a regularization technique based on optimal transport distributional robustness that can be widely applied to many settings including i) *semi-supervised learning*, ii) *domain adaptation*, iii) *domain generalization*, and iv) *adversarial machine learning* . In what follows, we present

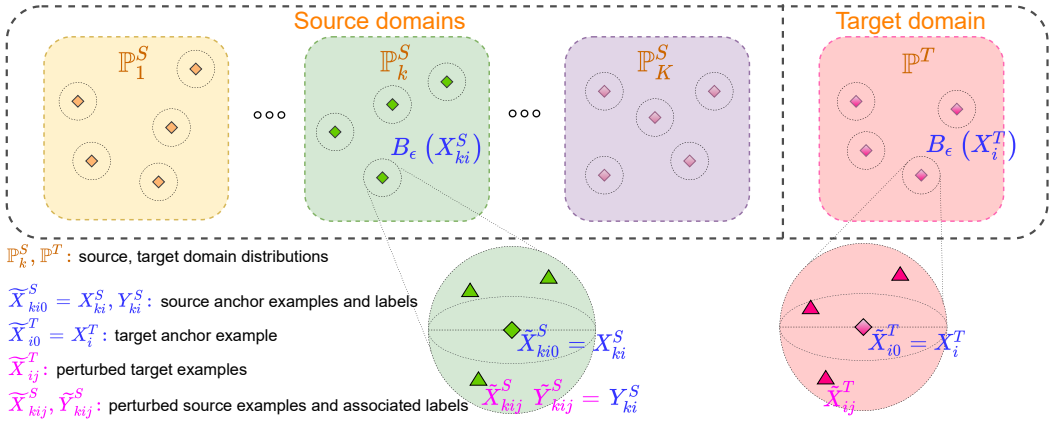


Figure 1: Overview of GLOT-DR. We sample $\left[X_{ki}^S, Y_{ki}^S\right]_{i=1}^{B_k^S}$ for each source domain, $\left[X_i^T\right]_{i=1}^{B^T}$ for the target domain, and define $Z, \tilde{Z}$ as in Eqs. (3,4). For $\left(Z, \tilde{Z}\right) \sim \gamma$ satisfying $\mathbb{E}_\gamma\left[\rho\left(Z, \tilde{Z}\right)\right]^{1/q} \leq \epsilon$, we have $\tilde{X}_{ki0}^S = X_{ki0}^S = X_{ki}^S$, $\tilde{X}_{i0}^T = X_{i0}^T = X_i^T$. Besides, $\tilde{X}_{kij}^S$ with $j \geq 1$ can be viewed as the *perturbed* examples in the ball $B_\epsilon\left(X_{ki}^S\right)$, which have the same label Y_{ki}^S. Similarly, $\tilde{X}_{ij}^T$ with $j \geq 1$ can be viewed as the *perturbed* examples in the ball $B_\epsilon\left(X_i^T\right)$.

the general setting along with the notations used throughout the paper and technical details of our framework.

Assume that we have *multiple labeled source domains* with the *data/label* distributions $\left\{\mathbb{P}_k^S\right\}_{k=1}^K$ and a *single unlabeled target domain* with the *data* distribution $\mathbb{P}^T$. For the k-th source domain, we draw a batch of B_k^S examples as $\left(X_{ki}^S, Y_{ki}^S\right) \overset{\text{iid}}{\sim} \mathbb{P}_k^S$, where $i = 1, \ldots, B_k^S$. Meanwhile, for the target domain, we sample a batch of B^T examples as $X_i^T \overset{\text{iid}}{\sim} \mathbb{P}^T$, $i = 1, \ldots, B^T$. It is worth noting that for the DG setting, we set $B^T = 0$ (i.e., not use any target data in training). Furthermore, we examine the multi-class classification problem with the label set $\mathcal{Y} := \{1, ..., M\}$. Hence, the prediction of a classifier is a prediction probability belonging to the *label simplex* $\Delta_M := \left\{\pi \in \mathbb{R}^M : \|\pi\|_1 = 1 \text{ and } \pi \geq \mathbf{0}\right\}$. Finally, let $f_\psi = h_\theta \circ g_\phi$ with $\psi = (\phi, \theta)$ be parameters of our deep net, wherein g_ϕ is the feature extractor and h_θ is the classifier on top of feature representations.

Constructing Challenging Samples: As explained below, our method involves the construction of a random variable Z with distribution $\mathbb{P}$ and another random variable $\tilde{Z}$ with distribution $\tilde{\mathbb{P}}$, "containing" anchor samples $\left(X_{ki}^S, Y_{ki}^S\right), X_i^T$ and their perturbed counterparts $\left(\tilde{X}_{kij}^S, \tilde{Y}_{kij}^S\right), \tilde{X}_{ij}^T$ (see Figure 1 for the illustration). The inclusion of both anchor samples and perturbed samples allows us to define a unifying cost function containing local regularization, global regularization, and classification loss.

Concretely, we first start with the construction of Z, containing repeated anchor samples as follows:

$$Z := \left[\left[\left[X_{kij}^S, Y_{kij}^S\right]_{k=1}^K\right]_{i=1}^{B_k^S}\right]_{j=0}^{n^S}, \left[\left[X_{ij}^T\right]_{i=1}^{B^T}\right]_{j=0}^{n^T}. \quad (3)$$

Here, each source sample is repeated $n^S + 1$ times $(X_{kij}^S, Y_{kij}^S) = (X_{ki}^S, Y_{ki}^S), \forall j$, while each target sample is repeated $n^T + 1$ times $X_{ij}^T = X_i^T, \forall j$. The corresponding distribution of this random variable is denoted as $\mathbb{P}$. In contrast to Z, we next define random variable $\tilde{Z} \sim \tilde{\mathbb{P}}$, whose form is

$$\tilde{Z} := \left[\left[\left[\tilde{X}_{kij}^S, \tilde{Y}_{kij}^S\right]_{k=1}^K\right]_{i=1}^{B_k^S}\right]_{j=0}^{n^S}, \left[\left[\tilde{X}_{ij}^T\right]_{i=1}^{B^T}\right]_{j=0}^{n^T}. \quad (4)$$

Here we note that for $\tilde{X}_{kij}^S$, the index k specifies the k-th source domain, the index i specifies an example in the k-th source batch, while the index j specifies the j-th perturbed example to the source example X_{ki}^S. Similarly, for $\tilde{X}_{ij}^T$, the index i specifies an example in the target batch, while the index j specifies the j-the perturbed example to the target example X_i^T.

We would like $\tilde{Z}$ to contain both: i) anchor examples, i.e., $\left(\tilde{X}_{ki0}^S, \tilde{Y}_{ki0}^S\right) = \left(X_{ki}^S, Y_{ki}^S\right)$ and $\tilde{X}_{i0}^T = X_i^T$; ii) n^S perturbed source samples $\left\{\left(\tilde{X}_{kij}^S, \tilde{Y}_{kij}^S\right)\right\}_{j=1}^{n^S}$ to $\left(X_{ki}^S, Y_{ki}^S\right)$ and n^T perturbed target samples $\left\{\tilde{X}_{ij}^T\right\}_{i=1}^{n^T}$ to X_i^T. In order to impose this requirement, we only consider sampling $\tilde{Z}$ from distribution $\tilde{\mathbb{P}}$ inside the Wasserstein-ball of $\mathbb{P}$, i.e., sat-

isfying $\mathcal{W}_\rho\left(\mathbb{P},\tilde{\mathbb{P}}\right) := \inf_{\gamma\in\Gamma\left(\mathbb{P},\tilde{\mathbb{P}}\right)} \mathbb{E}_{(Z,\tilde{Z})\sim\gamma}\left[\rho\left(Z,\tilde{Z}\right)\right]^{\frac{1}{q}} \leq \epsilon$, where the cost metric ρ is defined as

$$\rho\left(Z,\tilde{Z}\right) := \infty\sum_{k=1}^{K}\sum_{i=1}^{B_k^S}\left\|X_{ki0}^S - \tilde{X}_{ki0}^S\right\|_p^q + \infty\sum_{i=1}^{B^T}\left\|X_{i0}^T - \tilde{X}_{i0}^T\right\|_p^q + \sum_{k=1}^{K}\sum_{i=1}^{B_k^S}\sum_{j=1}^{n^S}\left\|X_{kij}^S - \tilde{X}_{kij}^S\right\|_p^q + \sum_{i=1}^{B^T}\sum_{j=1}^{n^T}\left\|X_{ij}^T - \tilde{X}_{ij}^T\right\|_p^q + \infty\sum_{k=1}^{K}\sum_{i=1}^{B_k^S}\sum_{j=0}^{n^S}\rho_l\left(Y_{kij}^S,\tilde{Y}_{kij}^S\right),$$

where ρ_l is a metric on the *label simplex* Δ_M and $q \geq 1$. Here we slightly abuse the notion by using $Y \in \mathcal{Y}$ to represent its corresponding one-hot vector. By definition, this cost metric almost surely: i) enforces all 0-th (i.e., $j = 0$) samples in $\tilde{Z}$ to be anchor samples, i.e., $\tilde{X}_{ki0}^S = X_{ki0} = X_{ki}$; ii) allows perturbations on the input data, i.e., $\tilde{X}_{kij}^S \neq X_{ki}^S$ and $\tilde{X}_{ij}^T \neq X_i^T$, for $\forall j \neq 0$; iii) restricts perturbations on labels, i.e., $Y_{kij}^S = \tilde{Y}_{kij}^S$ for $\forall j$ (see Figure 1 for the illustration). The reason is that if either (i) or (iii) is violated on a non-zero measurable set then $\mathcal{W}_\rho\left(\mathbb{P},\tilde{\mathbb{P}}\right)$ becomes infinity.

Learning Robust Classifier: Upon clear definitions of $\tilde{Z}$ and $\tilde{\mathbb{P}}$, we wish to learn good representations and regularize the classifier f_ψ, via the following DR problem:

$$\min_{\theta,\phi}\max_{\tilde{\mathbb{P}}:\mathcal{W}_\rho\left(\mathbb{P},\tilde{\mathbb{P}}\right)\leq\epsilon} \mathbb{E}_{\tilde{Z}\sim\tilde{\mathbb{P}}}\left[r\left(\tilde{Z};\phi,\theta\right)\right]. \tag{5}$$

The cost function $r\left(\tilde{Z};\phi,\theta\right) := \alpha r^l\left(\tilde{Z};\phi,\theta\right) + \beta r^g\left(\tilde{Z};\phi,\theta\right) + \mathcal{L}\left(\tilde{Z};\phi,\theta\right)$ with $\alpha,\beta > 0$ is defined as the weighted sum of a *local-regularization function* $r^l\left(\tilde{Z};\phi,\theta\right)$, a *global-regularization function* $r^g\left(\tilde{Z};\phi,\theta\right)$, and the *loss function* $\mathcal{L}\left(\tilde{Z};\phi,\theta\right)$, whose explicit forms are dependent on the task (DA, SSL, DG, and AML).

Intuitively, the optimization in Eq. (5) iteratively searches for the worst-case $\tilde{\mathbb{P}}$ w.r.t. the cost $r\left(\cdot;\phi,\theta\right)$, then changes the network f_ψ to minimize the worst-case cost.

We now define

$$\Gamma_\epsilon := \left\{\gamma : \gamma \in \bigcup_{\tilde{\mathbb{P}}}\Gamma\left(\mathbb{P},\tilde{\mathbb{P}}\right), \mathop{\mathbb{E}}_{(Z,\tilde{Z})\sim\gamma}\left[\rho\left(Z,\tilde{Z}\right)\right]^{1/q} \leq \epsilon\right\}$$

and show that the inner max problem in Eq. (5) is equivalent to searching in Γ_ϵ.

Lemma 3.1. *The optimization problem in Eq. (5) is equivalent to the following optimization problem:*

$$\min_{\theta,\phi}\max_{\gamma:\in\Gamma_\epsilon}\mathop{\mathbb{E}}_{(Z,\tilde{Z})\sim\gamma}\left[r\left(\tilde{Z};\phi,\theta\right)\right]. \tag{6}$$

To tackle the optimization problem (OP) in Eq. (6), we add the entropic regularization and arrive at the following OP:

$$\min_{\theta,\phi}\max_{\gamma:\in\Gamma_\epsilon}\left\{\mathop{\mathbb{E}}_{(Z,\tilde{Z})\sim\gamma}\left[r\left(\tilde{Z};\phi,\theta\right)\right] + \frac{1}{\lambda}\mathbb{H}\left(\gamma\right)\right\}, \tag{7}$$

where $\lambda > 0$ is the entropic regularization parameter and $\mathbb{H}$ returns the entropy of a given distribution.

It is worth noting that minimizing the entropy $\mathbb{H}\left(\gamma\right)$ encourages more uniform γ. Moreover, when λ becomes bigger, the optimal solution of the OP in Eq. (7) gets closer to that of (6). Additionally, the following theorem indicates the optimal solution of the inner max in the OP in Eq. (7).

Theorem 3.2. *Assuming* $r\left(\tilde{Z};\psi\right) = \alpha r^l\left(\tilde{Z};\psi\right) + \beta r^g\left(\tilde{Z};\psi\right) + \mathcal{L}\left(\tilde{Z};\psi\right)$ *with* $\psi = (\phi,\theta)$. *In addition, Z and $\tilde{Z}$ are constructed as in Eq.(3) and Eq.(4), respectively. Let ℓ denote the loss function, so the expected classification loss becomes*

$$\mathcal{L}\left(\tilde{Z};\psi\right) := \sum_{k=1}^{K}\sum_{i=1}^{B_k^S}\sum_{j=0}^{n_k^S}\ell\left(\tilde{X}_{kij}^S,\tilde{Y}_{kij}^S;\psi\right).$$

Moreover, let the global-regulazation $r^g\left(\tilde{Z};\psi\right) := r^g\left(\left[\tilde{X}_{ki0}^S\right]_{k,i},\left[\tilde{X}_{i0}^T\right]_i;\psi\right)$ *depend only on anchor samples, while the local-regularization depend on the differences between anchor samples and perturbed samples,*

$$r^l\left(\tilde{Z};\psi\right) := \sum_{i=1}^{B^T}\sum_{j=1}^{n^T}s\left(\tilde{X}_{i0}^T,\tilde{X}_{ij}^T;\psi\right) + \sum_{k=1}^{K}\sum_{i=1}^{B_k^S}\sum_{j=1}^{n_k^S}s\left(\tilde{X}_{ki0}^S,\tilde{X}_{kij}^S;\psi\right),$$

where $s\left(\tilde{X}_0,\tilde{X}_j;\psi\right)$ *measures the difference between 2 input samples, and* $s\left(X,X;\psi\right) = 0, \forall X$. *To this end, the inner max in the OP when* $q = \infty$ *has the following solution*

$$\gamma^*\left(Z,\tilde{Z}\right) = \prod_{k=1}^{K}\prod_{i=1}^{B_k^S}\prod_{j=0}^{n_k^S}p_k^S\left(X_{ki}^S,Y_{ki}^S\right)\prod_{i=1}^{B^T}\prod_{j=0}^{n^T}p^T\left(X_i^T\right) \prod_{k=1}^{K}\prod_{i=1}^{B_k^S}\prod_{j=0}^{n_k^S}q_{ki}^S\left(\tilde{X}_{kij}^S \mid X_{ki}^S,Y_{ki}^S;\psi\right)\prod_{i=1}^{B^T}\prod_{j=1}^{n^T}q_i^T\left(\tilde{X}_{ij}^T \mid X_i^T;\psi\right), \tag{8}$$

where $B_\epsilon\left(X\right) := \left\{X' : \left\|X'-X\right\|_p \leq \epsilon\right\}$ *is the ϵ-ball around X,* $\left(X_{ki}^S,Y_{ki}^S\right)_{i=1}^{B_k^S} \overset{iid}{\sim} \mathbb{P}_k^S, \forall k$, $X_{1:B^T}^T \overset{iid}{\sim} \mathbb{P}^T$, p_k^S *is the density function of* $\mathbb{P}_k^S$, p^T *is the density function of* $\mathbb{P}^\mathbb{T}$, $q_{ki}^S\left(\tilde{X}_{kij}^S \mid X_{ki}^S,Y_{ki}^S;\psi\right) \propto \exp\left\{\lambda[\alpha s(X_{ki}^S,\tilde{X}_{kij}^S;\psi) + \ell(\tilde{X}_{kij}^S,Y_{ki}^S;\psi)]\right\}$ *is* ***the local***

distribution over $B_\epsilon\left(X_{ki}^S\right)$ *around the anchor example* X_{ki}^S*, and* $q_i^T\left(\tilde{X}_{ij}^T \mid X_i^T;\psi\right) \propto \exp\left\{\lambda\alpha s\left(X_i^T,\tilde{X}_{ij}^T;\psi\right)\right\}$ *is* ***the local distribution over*** $B_\epsilon\left(X_i^T\right)$ *around the anchor example* X_i^T.

The optimal γ^* in Eq. (8) involves the local distributions q_{ki}^S around the anchor example X_{ki}^S and q_i^T around the anchor example X_i^T. By substituting the optimal solution in Eq. (8) back to Eq. (6), we reach the following OP with $\psi = (\phi, \theta)$:

$$\min_{\psi} \mathop{\mathbb{E}}_{\forall k:\left(X_{ki}^S,Y_{ki}^S\right)_{i=1}^{B_k^S}\overset{\text{iid}}{\sim}\mathbb{P}_k^S, X_{1:B^T}^T\overset{\text{iid}}{\sim}\mathbb{P}^T} \left[r\left(\tilde{Z};\psi\right)\right], \quad (9)$$

where $r\left(\tilde{Z};\psi\right)$ is defined as

$$\begin{aligned}\mathop{\mathbb{E}}_{\left[\tilde{X}_{kij}^S\right]_j\sim q_{ki}^S}&\left[\alpha s(X_{ki}^S,\tilde{X}_{kij}^S;\psi)+\ell(\tilde{X}_{kij}^S,Y_{ki}^S;\psi)\right]\\ &+\mathop{\mathbb{E}}_{\left[\tilde{X}_{ij}^T\right]_j\sim q_i^T}\left[\alpha s\left(X_i^T,\tilde{X}_{ij}^T;\psi\right)\right]\\ &+\beta r^g\left(\left[X_{ki}^S\right]_{k,i},\left[X_i^T\right]_i;\psi\right)\end{aligned} \quad (10)$$

with the *local distribution* q_{ki}^S over $B_\epsilon\left(X_{ki}^S\right)$ and the *local distribution* q_i^T over $B_\epsilon\left(X_i^T\right)$.

As shown in Eq. (10), the perturbed examples $\tilde{X}_{kij}^S$ are sampled from the local distribution q_{ki}^S over the ball $B_\epsilon\left(X_{ki}^S\right)$, while the perturbed examples $\tilde{X}_{ij}^T$ are sampled from the local distribution q_i^T over the ball $B_\epsilon\left(X_i^T\right)$. Due to the formula of q_{ki}^S, the perturbed examples $\tilde{X}_{kij}^S$ tend to reach the high-likelihood region of q_{ki}^S or high-valued region for $\exp\left\{\lambda[\alpha s(X_{ki}^S,\tilde{X}_{kij}^S;\psi)+\ell(\tilde{X}_{kij}^S,Y_{ki}^S;\psi)]\right\}$. We hence can interpret $\tilde{X}_{kij}^S$ as adversarial examples that maximize $\lambda[\alpha s(X_{ki}^S,\tilde{X}_{kij}^S;\psi)+\ell(\tilde{X}_{kij}^S,Y_{ki}^S;\psi)]$. Subsequently, in (10), we update ψ to minimize $\lambda[\alpha s(X_{ki}^S,\tilde{X}_{kij}^S;\psi)+\ell(\tilde{X}_{kij}^S,Y_{ki}^S;\psi)]$ w.r.t. the perturbed adversarial examples. Similarly, we can interpret the perturbed examples $\tilde{X}_{ij}^T$.

Additionally, we can equip the global-regularization function $r^g\left(\left[X_{ki}^S\right]_{k,i},\left[X_i^T\right]_i;\psi\right)$ to suit various characteristics for the task, e.g., bridging the distribution shift between source and target domains in DA, between labeled and unlabeled portions in SSL, and between benign and adversarial data examples in AML, as well as learning domain invariant features in DG. Moreover, our global and local regularization terms can be naturally applied to the latent space induced by the feature extractor g_ϕ. Furthermore, the theory development for this case is similar to that for the data space except replacing X in the data space by $g_\phi(X)$ in the latent space.

3.2 Training Procedure of Our Approach

In what follows, we present how to solve the OP in Eq. (9) efficiently. Accordingly, we first need to sample $\left(X_{ki}^S,Y_{ki}^S\right)_{i=1}^{B_k^S}\overset{\text{iid}}{\sim}\mathbb{P}_k^S, \forall k$ and $X_{1:B^T}^T\overset{\text{iid}}{\sim}\mathbb{P}^T$. For each source anchor $\left(X_{ki}^S,Y_{ki}^S\right)$, we sample $\left[\tilde{X}_{kij}^S\right]_{j=1}^{n^S}\overset{\text{iid}}{\sim}q_{ki}^S$ in the ball $B_\epsilon\left(X_{ki}^S\right)$ with the *density function proportional* to $\exp\left\{\lambda[\alpha s(X_{ki}^S,\bullet;\psi)+\ell(\bullet,Y_{ki}^S;\psi)]\right\}$. Furthermore, for each target anchor X_i^T, we sample $\left[\tilde{X}_{ij}^T\right]_{j=1}^{n^T}\overset{\text{iid}}{\sim}q_i^T$ in the ball $B_\epsilon\left(X_i^T\right)$ with the d*ensity function proportional* to $\exp\left\{\lambda\alpha s\left(X_i^T,\bullet;\psi\right)\right\}$.

To sample the particles from their local distributions, we use Stein Variational Gradient Decent (SVGD) (Liu and Wang, 2016; Phan et al., 2022) with a RBF kernel with kernel width σ. Obtained particles $\tilde{X}_{kij}^S$ and $\tilde{X}_{ij}^T$ are then utilized to minimize the objective function in Eq. (9) for updating $\psi = (\phi,\theta)$. Specifically, we utilize cross-entropy for the classification loss term ℓ and the symmetric Kullback-Leibler (KL) divergence for the local regularization term $s\left(X,\tilde{X};\psi\right)$ as $\frac{1}{2}KL\left(f_\psi(X)\,\|f_\psi\left(\tilde{X}\right)\right)+\frac{1}{2}KL\left(f_\psi\left(\tilde{X}\right)\|f_\psi(X)\right)$.

Finally, the global-regularization function of interest $r^g\left(\left[X_{ki}^S\right]_{k,i},\left[X_i^T\right]_i;\psi\right)$ is defined accordingly depending on the task and explicitly presented in the sequel.

3.3 Setting for Domain Adaptation and Semi-supervised Learning

By considering the single source domain as the labeled portion and the target domain as the unlabeled portion, the same setting can be employed for DA and SSL. Particularly, we denote the data/label distribution of the source domain or labeled portion by $\mathbb{P}_1^{S|l}$ and the data distribution of target domain or unlabeled portion by $\mathbb{P}^{T|u}$. Notice that for SSL, $\mathbb{P}^{T|u}$ could be the marginal of $\mathbb{P}^{S|l}$ by marginalizing out the label dimension. Evidently, with this consideration, DA and SSL are special cases of our general framework in Section 3.1, where the global-regularization function of interest $r^g\left(\left[X_i^S\right]_i,\left[X_j^T\right]_j;\psi\right)$ is defined as

$$\mathcal{W}_d\left(\frac{1}{B^S}\sum_{i=1}^{B^S}\delta_{U_i^S},\frac{1}{B^T}\sum_{j=1}^{B^T}\delta_{U_j^T}\right), \quad (11)$$

where $U_i^S = \left[g_\phi\left(X_i^S\right), h_\theta\left(g_\phi\left(X_i^S\right)\right)\right]$, $U_j^T = \left[g_\phi\left(X_j^T\right), h_\theta\left(g_\phi\left(X_j^T\right)\right)\right]$, and δ is the Dirac delta distribution. The cost metric d is defined as

$$\begin{aligned}d\left(U_i^S,U_j^T\right) :=\ &\rho_d\left(g_\phi\left(X_i^S\right),g_\phi\left(X_j^T\right)\right)\\ &+\gamma\rho_l\left(h_\theta\left(g_\phi\left(X_i^S\right)\right),h_\theta\left(g_\phi\left(X_j^T\right)\right)\right),\end{aligned} \quad (12)$$

Table 1: Single domain generalization accuracy (%) on CIFAR-10-C and CIFAR-100-C datasets with different backbone architectures. We use the **bold** font to highlight the best results.

Datasets	Backbone	Standard	Cutout	CutMix	AutoDA	Mixup	AdvTrain	ADA	ME-ADA	GLOT-DR
CIFAR-10-C	AllConvNet	69.2	67.1	68.7	70.8	75.4	71.9	73	78.2	**82.5**
	DenseNet	69.3	67.9	66.5	73.4	75.4	72.4	69.8	76.9	**83.6**
	WideResNet	73.1	73.2	72.9	76.1	77.7	73.8	79.7	83.3	**84.4**
	ResNeXt	72.5	71.1	70.5	75.8	77.4	73	78	83.4	**84.5**
	Average	71	69.8	69.7	74	76.5	72.8	75.1	80.5	**83.7**
CIFAR-100-C	AllConvNet	43.6	43.2	44	44.9	46.6	44	45.3	51.2	**54.8**
	DenseNet	40.7	40.4	40.8	46.1	44.6	44.8	45.2	47.8	**53.2**
	WideResNet	46.7	46.5	47.1	50.4	49.6	44.9	50.4	52.8	**56.5**
	ResNeXt	46.6	45.4	45.9	48.7	48.6	45.6	53.4	57.3	**58.4**
	Average	44.4	43.9	44.5	47.5	47.4	44.8	48.6	52.3	**55.7**

where ρ_d is a metric on the latent space and $\gamma > 0$.

With the global term in Eq. (11), we aim to reduce the discrepancy gap between the *source (labeled)* domain and the *target (unlabeled)* domain for learning domain-invariant representations. It is worth noting that this global term in Eq. (11) was inspected in DeepJDOT (Damodaran et al., 2018) for DA setting. Our approach is different from that approach in the local regularization term.

3.4 Setting for Domain Generalization

By setting $B^T = 0$ (i.e., not use any target data in training), our general framework in Section 3.1 is applicable to DG, wherein the global-regularization function of interest $r^g\left(\left[X_{ki}^S\right]_{k,i}, \left[X_i^T\right]_i; \psi\right)$ is

$$\sum_{m=1}^{M}\sum_{k=1}^{K}\frac{1}{K}\mathcal{W}_d\left(\tilde{\mathbb{P}}_{km}, \tilde{\mathbb{P}}_m\right), \tag{13}$$

where the cost metric $d = \rho_d$ is a metric on the latent space, $\tilde{\mathbb{P}}_{km}$ is the empirical distribution over $g_\phi\left(X_{ki}^S\right)$ with $Y_{ki}^S = m$, and $\tilde{\mathbb{P}}_m = \frac{1}{K}\sum_{k=1}^{K}\tilde{\mathbb{P}}_{km}$.

3.5 Setting for Adversarial Machine Learning

For AML, we have only *single source domain* and need to train a deep model which is robust to adversarial examples. We denote the data/label distribution of the source domain by $\mathbb{P}_1^S$ and propose using a dynamic and pseudo target domain of the *on-the-fly adversarial examples* $\left[\left[X_{1ij}^S\right]_{i=1}^{B^S}\right]_{j=1}^{n^S}$. In addition to the local and loss terms as in Eq. (9), to strengthen model robustness, we propose the following global term to move adversarial examples ($\sim \mathbb{P}^T$) to benign examples ($\sim \mathbb{P}_1^S$):

$$\mathcal{W}_d\left(\frac{1}{B_1^S}\sum_{i=1}^{B_1^S}\delta_{U_i^S}, \frac{1}{B_1^S n^S}\sum_{i=1}^{B_1^S}\sum_{j=1}^{n^S}\delta_{U_{ij}^S}\right), \tag{14}$$

where $U_i^S = \left[g_\phi\left(X_{1i}^S\right), h_\theta\left(g_\phi\left(X_{1i}^S\right)\right)\right]$, $U_{ij}^S = \left[g_\phi\left(X_{1ij}^S\right), h_\theta\left(g_\phi\left(X_{1ij}^S\right)\right)\right]$, and the metric d is

$$d\left(U_i^S, U_{\bar{i}j}^S\right) = \mathbb{I}_{Y_{1i}^S = Y_{1\bar{i}}^S}\Bigg[\rho_d\left(g_\phi\left(X_{1i}^S\right), g_\phi\left(X_{1\bar{i}j}^S\right)\right) + \gamma\rho_l\left(h_\theta\left(g_\phi\left(X_{1i}^S\right)\right), h_\theta\left(g_\phi\left(X_{1\bar{i}j}^S\right)\right)\right)\Bigg], \tag{15}$$

where $\mathbb{I}$ is the indicator function. Here we note that $X_{1\bar{i}j}^S$ is an adversarial example of $X_{1\bar{i}}^S$ which has the ground-truth label $Y_{1\bar{i}}^S$, hence by using the cost metric as in Eq. (15), we encourage the adversarial example $X_{1\bar{i}j}^S$ to move to a group of the benign examples with the same label.

Finally, to tackle the WS-related terms in equations. (11,13, and 14), we employ the entropic regularization dual form of WS, which was demonstrated to have favorable computational complexities (Lin et al., 2020, 2019a,b).

4 Experiments

To demonstrate the effectiveness of our proposed method, we evaluate its performance on various experiment protocols, including DG, DA, SSL, and AML. Due to the space limitation, the detailed setup regarding the architectures and hyperparameters are presented in the supplementary material[1]. We tried to use the exact configuration of optimizers and hyper-parameters for all experiments and report the original results in prior work, if possible.

4.1 Experiments for DG

In DG experiments, our setup closely follows (Zhao et al., 2020). In particular, we validate our method on the CIFAR-C single domain generalization benchmark: train the model

[1]Our codes are available at https://github.com/VietHoang1512/GLOT

Table 2: Multi-source domain generalization accuracy (%) on PACS datasets. Each column title indicates the target domain used for evaluation, while the rest are for training.

	DSN	L-CNN	MLDG	Fusion	MetaReg	Epi-FCR	AGG	HEX	PAR	ADA	ME-ADA	GLOT-DR
Art	61.1	62.9	66.2	64.1	69.8	64.7	63.4	66.8	66.9	64.3	**67.1**	66.1
Cartoon	66.5	67.0	66.9	66.8	70.4	**72.3**	66.1	69.7	67.1	69.8	69.9	**72.3**
Photo	83.3	89.5	88.0	90.2	91.1	86.1	88.5	87.9	88.6	85.1	88.6	**90.4**
Sketch	58.6	57.5	59.0	60.1	59.2	65.0	56.6	56.3	62.6	60.4	63.0	**65.4**
Average	67.4	69.2	70.0	70.3	72.6	72.0	68.7	70.2	71.3	69.9	72.2	**73.5**

on either CIFAR-10 or CIFAR-100 dataset (Krizhevsky et al., 2009), then evaluate it on CIFAR-10-C or CIFAR-100-C (Hendrycks and Dietterich, 2019), correspondingly. In terms of network architectures, we use the exact backbones from (Zhao et al., 2020) to examine the versatility of our method that can be adopted in any type of classifier. GLOT-DR is compared with other state-of-the-art methods in image corruption robustness: Mixup (Zhang et al., 2018), Cutout (DeVries and Taylor, 2017) and Cutmix (Yun et al., 2019), AutoDA (Cubuk et al., 2019), ADA (Volpi et al., 2018), and ME-ADA (Zhao et al., 2020).

Table 1 shows the average accuracy when we alternatively train the model on one category and evaluate on the rest. In every setting, GLOT-DR outperforms other methods by large margins. Specifically, our method exceeds the second-best method ME-ADA (Zhao et al., 2020) by 3.2% on CIFAR-10-C and 3.4% on CIFAR-100-C. The substantial gain in terms of the accuracy on various backbone architectures demonstrates the high applicability of our GLOT-DR.

Furthermore, we examine multi-source DG where the classifier needs to generalize from multiple source domains to an unseen target domain on the PACS dataset (Li et al., 2017). Our proposed method is applicable in this scenario since it is designed to better learn domain invariant features as well as leverage the diversity from generated data. We compare GLOT-DR against DSN (Bousmalis et al., 2016), L-CNN (Li et al., 2017), MLDG (Li et al., 2018), Fusion (Mancini et al., 2018), MetaReg (Balaji et al., 2018), Epi-FCR, AGG (Li et al., 2019), HEX (Wang et al., 2019b), and PAR (Wang et al., 2019a). Table 2 shows that our GLOT-DR outperforms the baselines for three cases and averagely surpasses the second-best baseline by 0.9%. The most noticeable improvement is on the Sketch domain ($\approx 2.4\%$), which is the most challenging due to the fact that the styles of the images are colorless and far different from the ones from Art Painting, Cartoon or Photos (i.e., larger domain shift).

4.2 Experiments for DA

In this section, we conduct experiments on the commonly used dataset for real-world unsupervised DA - Office-31 (Saenko et al., 2010), comprising images from three domains: Amazon (A), Webcam (W) and DSLR (D). Our proposed GLOT-DR is compared against baselines: ResNet-50 (He et al., 2016), DAN (Long et al., 2015), RTN (Long et al., 2016), DANN (Ganin et al., 2016), JAN (Long et al., 2017b), GTA (Sankaranarayanan et al., 2018), CDAN (Long et al., 2017a), DeepJDOT (Damodaran et al., 2018) and ETD (Li et al., 2020). For a fair comparison, we follow the training setups of CDAN and compare with other works using this configuration. As can be seen from Table 3, GLOT-DR achieves the best overall performance among baselines with 87.8% accuracy. Compared with ETD, which is another OT-based domain adaptation method, our performance significantly increase by 4.1% on A→W task, 2.1% on W→A and 1.6% on average.

Table 3: Accuracy (%) on Office-31 (Saenko et al., 2010) of ResNet50 model (He et al., 2016) in unsupervised DA methods.

Method	A→W	D→W	W→D	A→D	D→A	W→A	Avg
ResNet	68.4	96.7	99.3	68.9	62.5	60.7	76.1
DAN	80.5	97.1	99.6	78.6	63.6	62.8	80.4
RTN	70.2	96.6	95.5	66.3	54.9	53.1	72.8
DANN	84.5	96.8	99.4	77.5	66.2	64.8	81.6
JAN	82	96.9	99.1	79.7	68.2	67.4	82.2
GTA	89.5	97.9	99.8	87.7	72.8	**71.4**	86.5
CDAN	93.1	98.2	**100**	**89.8**	70.1	68	86.6
DeepJDOT	88.9	98.5	99.6	88.2	**72.1**	70.1	86.2
ETD	92.1	**100**	**100**	88	71	67.8	86.2
GLOT-DR	**96.2**	98.9	**100**	90.6	69.9	69.6	**87.8**

We further extensively investigate the role of different components in GLOT-DR. Specifically, the elimination of the global-regularization term in equation (11) downgrades our method to Local Optimal Transport based Distributional Robustness (LOT-DR). Similarly, when discarding the local distribution robustness term, the attained method is denoted by GOT-DR. We then compare these 2 variants of GLOT-DR to the well-known adversarial machine learning method VAT (Miyato et al., 2018). To be more specific, in the adversarial samples generation, we apply VAT by perturbing on the: (i) input space, (ii) latent space. Figure 2 shows that the employment of VAT on latent space (orange) is more effective than on the input space (purple), 83% and 80.6%. However, using GOT-DR or LOT-DR is even more effective: performance is boosted to 84.3% and 85.4%, respectively. Lastly, using the full method GLOT-DR yields the highest average accuracy score among all.

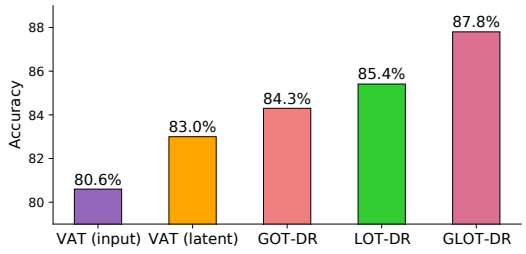


Figure 2: Average accuracy of ResNet50 (He et al., 2016) on Office-31: Comparision between GLOT-DR's variants and VAT (Miyato et al., 2018) on the input and latent spaces.

4.3 Experiments for SSL

Sharing a similar objective with DA, which utilizes the unlabeled samples for improving the model performance, SSL methods can also benefit from our proposed technique. We present the empirical results on CIFAR-10 benchmark with ConvLarge architecture, following VAT's protocol (Miyato et al., 2018), which serves as a strong baseline in this experiment. We refer readers to the supplementary material for more details on the architecture of ConvLarge. Results in Figure 3 (when training with 1,000 and 4,000 labeled examples) demonstrate that, with only $n^S = n^T = 1$ perturbed sample per anchor, the performance of LOT-DR slightly outperforms VAT with $\sim 0.5\%$. With more perturbed samples per anchor, this gap increases: approximately 1% when $n^S = n^T = 2$ and 1.5% when $n^S = n^T = 4$. Similar to the previous DA experiment, adding the global regularization term helps increase accuracy by $\sim 1\%$ in this setup.

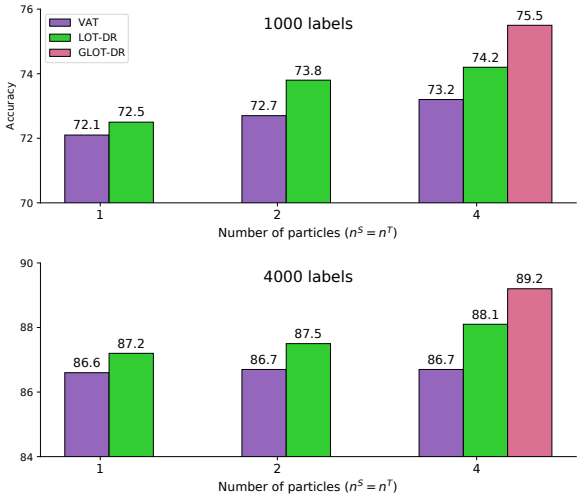


Figure 3: Accuracy (%) on CIFAR-10 of ConvLarge model in SSL settings when using 1,000 and 4,000 labeled examples (i.e. 100 and 400 labeled samples each class). Best viewed in color.

4.4 Experiments for AML

Table 4 shows the evaluation against adversarial examples. We compare our method with PGD-AT (Madry et al., 2018) and TRADES (Zhang et al., 2019), two well-known defense methods in AML and SAT (Bouniot et al., 2021). For the sake of fair comparison, we use the same adversarial training setting for all methods, which is carefully investigated in (Pang et al., 2020). We also compare with adversarial distributional training methods (Dong et al., 2020) (ADT-EXP and ADT-EXPAM), which assume that the adversarial distribution explicitly follows normal distribution. It can be seen from Table 4 that our GLOT-DR method outperforms all these baselines in both natural and robustness performance. Specifically, compared to PGD-AT, our method has an improvement of 0.8% in natural accuracy and around 1% robust accuracies against PGD200 and AA attacks. Compared to TRADES, while achieving the same level of robustness, our method has a better performance with benign examples with a gap of 2.5%. Especially, our method significantly outperforms ADT by around 7% under the PGD200 attack.

Table 4: Adversarial robustness evaluation on CIFAR10 of ResNet18 model. PGD, AA and B&B represent the robust accuracy against the PGD attack (with 10/200 iterations) (Madry et al., 2018), Auto-Attack (Croce and Hein, 2020) and B&B attack (Brendel et al., 2019), respectively, while NAT denotes the natural accuracy. Note that * results are taken from Pang et al. (Pang et al., 2020), while $^\diamond$ results are our reproduced results.

Method	NAT	PGD10	PGD200	AA	B&B
PGD-AT*	82.52	53.58	-	48.51	-
TRADES*	81.45	53.51	-	49.06	-
PGD-AT$^\diamond$	83.36	53.52	52.21	49.00	48.50
TRADES$^\diamond$	81.64	53.73	53.11	49.77	49.02
ADT-EXP	83.02	-	45.80	45.80	46.50
ADT-EXPAM	84.11	-	46.10	44.50	45.83
SAT	83.45	53.95	51.37	48.80	**49.40**
GLOT-DR	**84.13**	**54.13**	**53.1**8	**49.94**	**49.40**

5 Conclusion

Although DR is a promising framework to improve neural network robustness and generalization capability, its current formulation shows some limitations, circumventing its application to real-world problems. Firstly, its formulation is not sufficiently rich to express a global regularization effect targeting many applications. Secondly, the dual form is not readily trainable to incorporate into the training of deep learning models. In this work, we propose a rich OT based DR framework, named ***G**lobal-**L**ocal **O**ptimal **T**ransport based **D**istributional **R**obustness* (GLOT-DR) which is sufficiently rich for many real-world applications including DG, DA, SSL, and AML and has a closed-form solution. Finally, we conduct comprehensive experiments to compare our GLOT-DR with state-of-the-art baselines accordingly. Empirical results have demonstrated the merits of our GLOT-DR on standard benchmark datasets .

AN ADDITIVE INSTANCE-WISE APPROACH TO MULTI-CLASS MODEL INTERPRETATION

Vy Vo[1] **Van Nguyen**[1] **Trung Le**[1]
Quan Hung Tran[2] **Gholamreza Haffari**[1] **Seyit Camtepe**[3] **Dinh Phung**[1,4]
[1]Monash University, Australia
[2]Adobe Research, USA
[3]CSIRO's Data61, Australia
[4]VinAI Research, Vietnam

ABSTRACT

Interpretable machine learning offers insights into what factors drive a certain prediction of a black-box system. A large number of interpreting methods focus on identifying explanatory input features, which generally fall into two main categories: attribution and selection. A popular attribution-based approach is to exploit local neighborhoods for learning instance-specific explainers in an additive manner. The process is thus inefficient and susceptible to poorly-conditioned samples. Meanwhile, many selection-based methods directly optimize local feature distributions in an instance-wise training framework, thereby being capable of leveraging global information from other inputs. However, they can only interpret single-class predictions and many suffer from inconsistency across different settings, due to a strict reliance on a pre-defined number of features selected. This work exploits the strengths of both methods and proposes a framework for learning local explanations simultaneously for multiple target classes. Our model explainer significantly outperforms additive and instance-wise counterparts on faithfulness with more compact and comprehensible explanations. We also demonstrate the capacity to select stable and important features through extensive experiments on various data sets and black-box model architectures.

1 INTRODUCTION

Black-box machine learning systems enjoy a remarkable predictive performance at the cost of interpretability. This trade-off has motivated a number of interpreting approaches for explaining the behavior of these complex models. Such explanations are particularly useful for high-stakes applications such as healthcare (Caruana et al., 2015; Rich, 2016), cybersecurity (Nguyen et al., 2021) or criminal investigation (Lipton, 2018). While model interpretation can be done in various ways (Mothilal et al., 2020; Bodria et al., 2021), our discussion will focus on feature importance or saliency-based approach - that is, to assign relative importance weights to individual features w.r.t the model's prediction on an input example. Features here refer to input components interpretable to humans; for high-dimensional data such as texts or images, features can be a bag of words/phrases or a group of pixels/super-pixels (Ribeiro et al., 2016). Explanations are generally made by selecting top K features with the highest weights, signifying K most important features to a black-box's decision. Note that this work tackles feature selection locally for an input data point, instead of generating global explanations for an entire dataset.

An abundance of interpreting works follows the removal-based explanation approach (Covert et al., 2021), which quantifies the importance of features by removing them from the model. Based on how feature influence is summarized into an explanation, methods in this line of works can be broadly categorized as *feature attribution* and *feature selection*. In general, attribution methods produce relative importance scores to each feature, whereas selection methods directly identify the subset of features most relevant to the model behavior being explained. One popular approach to learn attribution is through an **Additive** model (Ribeiro et al., 2016; Zafar & Khan, 2019; Zhao et al., 2021). The underlying principle is originally proposed by LIME (Ribeiro et al., 2016) which

learns a regularized linear model for each input example wherein each coefficient represents feature importance scores. LIME explainer takes the form of a linear model $\boldsymbol{w}.\boldsymbol{z}$ where $\boldsymbol{z}$ denotes neighboring examples sampled heuristically around the input [1]. Though highly interpretable themselves, additive methods are inefficient since they optimize individual explainers for every input. As opposed to the instance-specific nature of the additive model, most of the feature selection methods are developed instance-wisely (Chen et al., 2018; Bang et al., 2021; Yoon et al., 2019; Jethani et al., 2021a). **Instance-wise** frameworks entail global training of a model approximating the local distributions over subsets of input features. Post-hoc explanations can thus be obtained simultaneously for multiple instances.

Contributions. In this work, we propose a novel strategy integrating both approaches into an **additive instance-wise** framework that simultaneously tackles all issues discussed above. The framework consists of 2 main components: an explainer and a feature selector. The explainer first learns the local attributions of features across the space of the response variable via a multi-class explanation module denoted as $\boldsymbol{W}$. This module interacts with the input vector in an additive manner forming a linear classifier locally approximating the black-box decision. To support the learning of local explanations, the feature selector constructs local distributions that can generate high-quality neighboring samples on which the explainer can be trained effectively. Both components are jointly optimized via backpropagation. Unlike such works as (Chen et al., 2018; Bang et al., 2021) that are sensitive to the choice of K as a hyper-parameter, our learning process eliminates this reliance (See Appendix G for a detailed analysis on why this is necessary).

Our contributions are summarized as follows

- We introduce **AIM** - an **A**dditive **I**nstance-wise approach to **M**ulti-class model interpretation. Our model explainer inherits merits from both families of methods: model-agnosticism, flexibility while supporting efficient interpretation for multiple decision classes. To the best of our knowledge, we are the first to integrate *additive* and *instance-wise* approaches into an end-to-end amortized framework that produces such a multi-class explanation facility.
- Our model explainer is shown to produce remarkably faithful explanations of high quality and compactness. Through quantitative and human assessment results, we achieve superior performance over the baselines on different datasets and architectures of the black-box model.

2 Related work

Early interpreting methods are gradient-based in which gradient values are used to estimate attribution scores, which quantifies how much a change in an input feature affects the black-box's prediction in infinitesimal regions around the input. It originally involves back-propagation for calculating the gradients of the output neuron w.r.t the input features (Simonyan et al., 2014). This early approach however suffers from vanishing gradients during the backward pass through ReLU layers that can downgrade important features. Several methods are proposed to improve the propagation rule (Bach et al., 2015; Springenberg et al., 2014; Shrikumar et al., 2017; Sundararajan et al., 2017). Since explanations based on raw gradients tend to be noisy highlighting meaningless variations, a refined approach is sampling-based gradient, in which sampling is done according to a prior distribution for computing the gradients of probability Baehrens et al. (2010) or expectation function (Smilkov et al., 2017; Adebayo et al., 2018). Functional Information (FI) (Gat et al., 2022) is the state-of-the-art in this line of research applying functional entropy to compute feature attributions. FI is shown to work on auditory, visual and textual modalities, whereas most of the previous gradient-based methods are solely applicable to images.

A burgeoning body of works in recent years can be broadly categorized as removal-based explanation (Covert et al., 2021). Common removal techniques include replacing feature values with neutral or user-defined values such as zero or Gaussian noises (Zeiler & Fergus, 2014; Dabkowski & Gal, 2017; Fong & Vedaldi, 2017; Petsiuk et al., 2018; Fong et al., 2019), marginalization of distributions over input features (Lundberg & Lee, 2017; Covert et al., 2020; Datta et al., 2016), or substituting held-out feature values with samples from the same distribution (Ribeiro et al., 2016). The output explanations are often either attribution-based or selection-based. In addition to additive models

[1] $\boldsymbol{z}$ is a binary representation vector of an input indicating the presence/absence of features. The dot-product operation is equivalent to summing up feature weights given by the weight vector $\boldsymbol{w}$, giving rise to additivity.

that estimate feature important via the coefficients of the linear model, feature attributions can be calculated using Shapley values (Datta et al., 2016; Lundberg & Lee, 2017; Covert et al., 2020) or directly obtained by measuring the changes in the predictive probabilities or prediction losses when adding or excluding certain features (Zeiler & Fergus, 2014; Schwab & Karlen, 2019). On the other hand, selection-based works straightforwardly determine which subset of features are important or unimportant to the model behavior under analysis (Chen et al., 2018; Yoon et al., 2019; Bang et al., 2021; Jethani et al., 2021a; Nguyen et al., 2021; 2022). Explanations are made by either selecting features with the highest logit scores obtained from the learned feature distribution, or specifying a threshold between 0 and 1 to decide on the most probable features. Most selection methods adopt amortized optimization, thus post-hoc inference of features for multiple inputs can be done very efficiently. In contrast, attribution-based approaches are mostly less efficient since they process input examples individually. There have been methods focusing on improving the computational cost of these models (Dabkowski & Gal, 2017; Schwab & Karlen, 2019; Jethani et al., 2021b).

Recently, there is an emerging interest in integrating the *instance-wise* property into an *additive* framework to better exploit global information. For a given input, Plumb et al. (2018); Yoon et al. (2022) in particular learn a surrogate model assigning weights to training examples such that those more similar or relevant to the input are given higher weights. A locally interpretable model (that is often LIME-based) is subsequently trained on these samples to return feature attributions. Agarwal et al. (2021) further proposes a neural additive framework that constructs an explainer in a form of a linear combination of neural networks. Despite their potential, these methods have only been reported to work on tabular data.

3 Proposed Method

3.1 Problem Setup

In the scope of this paper, we limit the current discussion to classification problems. Consider a data set of pairs (X, Y) where $X \sim \mathbb{P}_X(\cdot)$ is the input random variable and Y is characterized by the conditional distribution $\mathbb{P}_m(Y \mid X)$ obtained as the predictions of a pre-trained black-box classifier for the response variable. The notation m stands for *model*, indicating the predictive distribution of the black-box model, to be differentiated from the ground-truth distribution. We denote $\boldsymbol{x} \in \mathbb{R}^d$ as an input realization with d interpretable features and predicted label $Y = c \in \{1, ..., C\}$. Given an input $\boldsymbol{x}$, we obtain the hard prediction from the black-box model as $y_m = \text{argmax}_\text{c}\ \mathbb{P}_m(Y = c \mid X = \boldsymbol{x})$.

While earlier methods generate a single $d-$dimensional weight vector $\boldsymbol{w}_{\boldsymbol{x}}$ assigning the importance weights to each feature, we define an **explainer** $\mathcal{E} : \mathbb{R}^d \mapsto \mathbb{R}^{d \times C}$ mapping an input $\boldsymbol{x}$ to a weight matrix $\boldsymbol{W}_{\boldsymbol{x}} \in \mathbb{R}^{d \times C}$ with the entry $W^{i,j}_{\boldsymbol{x}}$ representing the relative weights of the ith feature of $\boldsymbol{x}$ to the predicted label $j \in \{1, ..., C\}$.

$$\mathcal{E}(\boldsymbol{x}) = \boldsymbol{W}_{\boldsymbol{x}}.$$

Given a training batch, LIME (Ribeiro et al., 2016) in particular trains separate explainers for every input, thus cannot take advantage of the global information from the entire dataset. In line with the *instance-wise* motivation, our explainer $\mathcal{E}$ is trained globally over all training examples to produce local explanations with respect to individual inputs simultaneously, which also seeks to effectively enable global behavior (e.g., two similar instances should have similar explanations). As $\mathcal{E}$ is expected to be locally faithful to the black-box model, we optimize $\mathcal{E}$ on the local neighborhood around the input $\boldsymbol{x}$. This region is constructed via a **feature selection** module. We now explain how this is done.

3.2 Training Objectives

Let $\mathbf{z} \in \{0, 1\}^d$ be a random variable with the entry $\text{z}^i = 1$ indicating the feature ith is important to the black-box's predictions. With respect to $\boldsymbol{x}$, we employ a **selector** $\mathcal{S} : \mathbb{R}^d \mapsto [0, 1]^d$ that outputs $\mathcal{S}(\boldsymbol{x}) = \pi_{\boldsymbol{x}}$ such that $\pi^i_{\boldsymbol{x}} := \mathbb{P}(\text{z}^i = 1 \mid X = \boldsymbol{x}), i = 1, ..., d$.

Through the probability vector $\pi_{\boldsymbol{x}}$, the selector helps define a local distribution on a local space of samples $\boldsymbol{z}_{\boldsymbol{x}} \odot \boldsymbol{x}$ with $\boldsymbol{z}_{\boldsymbol{x}} \sim \text{MultiBernoulli}(\pi_{\boldsymbol{x}})$ and element-wise product $\odot$. The selector $\mathcal{S}$ is also a learnable module, and we want it to generate well-behaved local samples that focus more on valuable features/attributions of $\boldsymbol{x}$. Intuitively, if the feature i of $\boldsymbol{x}$ contributes more to the predictions of the black-box model, i.e., $\pi^i_{\boldsymbol{x}} \approx 1$, the explainer is expected to give higher assignments to the row

vector $W_{\boldsymbol{x}}^{i,:}$. To mimic how the black-box model behaves towards different attributions, we propose to minimize the cross-entropy loss between the prediction of the black-box model on local examples $\boldsymbol{z_x} \odot \boldsymbol{x}$ and the prediction of the explainer on binary vectors $\boldsymbol{z_x}$ via the weight matrix $\boldsymbol{W_x}$ as

$$\mathcal{L}_1 = \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{z_x}} \Big[\mathrm{CE} \left(\tilde{y}_m, \mathrm{softmax}(\boldsymbol{W}_{\boldsymbol{x}}^T \boldsymbol{z_x}) \right) \Big], \tag{1}$$

where CE is the cross-entropy function and $\tilde{y}_m = \mathrm{argmax}_c \, \mathbb{P}_m(Y = c \mid \boldsymbol{z_x} \odot \boldsymbol{x})$.

To make the process continuous and differentiable for training, the temperature-dependent Gumbel-Softmax trick (Jang et al., 2016; Maddison et al., 2016) is applied for relaxing Bernoulli variables $\mathrm{z}_{\boldsymbol{x}}^i$. In particular, the continuous representation $\tilde{z}_{\boldsymbol{x}}^i$ is sampled from the Concrete distribution as $\left[\tilde{\mathrm{z}}_{\boldsymbol{x}}^i, 1 - \tilde{\mathrm{z}}_{\boldsymbol{x}}^i\right] \sim \mathrm{Concrete}(\pi_{\boldsymbol{x}}^i, 1 - \pi_{\boldsymbol{x}}^i)$:

$$\tilde{z}_{\boldsymbol{x}}^i = \frac{\exp\{\big(\log \pi_{\boldsymbol{x}}^i + G_{i1} \big)/\tau\}}{\exp\{(\log(1 - \pi_{\boldsymbol{x}}^i) + G_{i0})/\tau\} + \exp\{(\log \pi_{\boldsymbol{x}}^i + G_{i1})/\tau)\}},$$

with temperature τ, random noises G_{i0} and G_{i1} independently drawn from **Gumbel** distribution $G_t = -\log(-\log u_t),\ u_t \sim \textbf{Uniform}(0, 1)$.

Given the corresponding prediction $\tilde{y}_m = \mathrm{argmax}_c \, \mathbb{P}_m(Y = c \mid \tilde{\boldsymbol{z}}_{\boldsymbol{x}} \odot \boldsymbol{x})$, $\mathcal{L}_1$ now becomes

$$\mathcal{L}_1 = \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\tilde{\boldsymbol{z}}_{\boldsymbol{x}}} \Big[\mathrm{CE} \left(\tilde{y}_m, \mathrm{softmax}(\boldsymbol{W}_{\boldsymbol{x}}^T \tilde{\boldsymbol{z}}_{\boldsymbol{x}}) \right) \Big]. \tag{2}$$

Since $\boldsymbol{z_x}$ is a binary vector indicating the absence/presence of features, $\boldsymbol{z_x} \odot \boldsymbol{x}$ indeed acts as a local perturbation, which generally concurs with the principle of LIME model. However, different from LIME, we amortize the explainer $\mathcal{E}$ to produce the weight matrices $\boldsymbol{W_x}$ locally approximating the black-box model with linear classifiers operating on input neighborhoods. Furthermore, we replace LIME's uniform sampling strategy with a learnable local distribution offered by the selector $\mathcal{S}$.

We argue that heuristic sampling is inadequate for our purpose. As d gets large, realizing the space of 2^d possible binary patterns is infeasible. Given the fact that the number of binary patterns that actually approximate the original prediction is arbitrarily small, it is thus very difficult for such a simple linear separator as one used in LIME to learn useful patterns within finite sampling rounds. While diversity in these samples is desirable for learning attributions for individual decision classes, we also want the explainer $\mathcal{E}$ to focus more on relevant features to the original prediction y_m. To encourage the selector to yield more of the samples that contain the features that best approximate the model behavior on the original input, we propose the following information-theoretic approach.

Let $\boldsymbol{x}_{\mathbb{S}}$ denote the sub-vector formed by the subset of K most important features $\mathbb{S} = \{i_1, \ldots, i_K\} \subset \{1, \ldots, d\}$ $(i_1 < i_2 < \cdots < i_K)$. Thus, $\pi_{\boldsymbol{x}}^i$ can now be viewed as the probability that the ith feature of $\boldsymbol{x}$ appears in $\mathbb{S}$. Given a random vector $X_{\mathbb{S}} \in \mathbb{R}^K$, we maximize the mutual information.

$$\mathbb{I}(X_{\mathbb{S}}; Y) = \mathbb{E}\Big[\log \frac{\mathbb{P}_m(Y \mid X_{\mathbb{S}})}{\mathbb{P}_m(Y)} \Big] = \mathbb{E}_X \mathbb{E}_{\mathbb{S}|X} \mathbb{E}_{Y|X_{\mathbb{S}}} \Big[\log \mathbb{P}_m(Y \mid X_{\mathbb{S}}) \Big] + \text{Constant}. \tag{3}$$

Based on the following inequality, we can obtain a variational lower bound for $\mathbb{I}(X_{\mathbb{S}}; Y)$ via a generic choice of conditional distribution $\mathbb{Q}_{\mathbb{S}}(Y \mid X_{\mathbb{S}})$

$$\begin{aligned} \mathbb{E}_{Y|X_{\mathbb{S}}} \left[\log \mathbb{P}_m \left(Y \mid X_{\mathbb{S}} \right) \right] &= \mathbb{E}_{Y|X_{\mathbb{S}}} \left[\log \mathbb{Q}_{\mathbb{S}} \left(Y \mid X_{\mathbb{S}} \right) \right] + \mathrm{KL} \left(\mathbb{P}_m \left(Y \mid X_{\mathbb{S}} \right), \mathbb{Q}_{\mathbb{S}} \left(Y \mid X_{\mathbb{S}} \right) \right) \\ &\geq \mathbb{E}_{Y|X_{\mathbb{S}}} \left[\log \mathbb{Q}_{\mathbb{S}} \left(Y \mid X_{\mathbb{S}} \right) \right], \end{aligned}$$

where KL represents the Kullback-Leibler divergence.

It is worth noting that the purpose of using the mutual information in L2X (Chen et al., 2018) and our AIM framework are different. L2X uses the mutual information to directly select valuable features/attributions. Meanwhile, in our work, the role of the selector is to balance exploration with exploitation. Stochastic sampling yields various examples that produce different predictions, and $\mathcal{E}$ exploits such variation to learn feature attributions w.r.t multiple classes. Simultaneously, maximizing $\mathbb{I}(X_{\mathbb{S}}; Y)$ encourages the selector $\mathcal{S}$ to produce a meaningful probability vector focusing more on

the selected subset of attributions that can well approximate the full-input decision. In Appendix D, we show that an explainer with learnable local distributions performs significantly better than one optimized on heuristic examples.

Maximizing the mutual information in Eq. (3) can therefore be relaxed to maximizing the variational lower bound $\mathbb{E}_X \mathbb{E}_{\mathbb{S}|X} \mathbb{E}_{Y|X_\mathbb{S}} \Big[\log \mathbb{Q}_\mathbb{S}(Y \mid X_\mathbb{S}) \Big]$. We parametrize $\mathbb{Q}$ with a function approximator $\mathcal{G}$ such that $\mathbb{Q}_\mathbb{S}(\boldsymbol{x}_\mathbb{S}) := \mathcal{G}(\boldsymbol{x}_\mathbb{S})$. Notice that we can now use the element-wise product $\tilde{\boldsymbol{z}}_{\boldsymbol{x}} \odot \boldsymbol{x}$ to approximate $\boldsymbol{x}_\mathbb{S}$. If $\boldsymbol{x}$ contains discrete features (e.g., words), we embed a feature (e.g., a selected word) in $\mathbb{S}$ with a learnable embedding vector, wherein a feature not in $\mathbb{S}$ is replaced with a zero vector. With the prediction $y_m = \text{argmax}_\text{c}\, \mathbb{P}_m(Y = c \mid X = \boldsymbol{x})$, our second objective is given as

$$\mathcal{L}_2 = \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\tilde{\boldsymbol{z}}_{\boldsymbol{x}}} \Big[\text{CE}\left(y_m, \mathcal{G}(\tilde{\boldsymbol{z}}_{\boldsymbol{x}} \odot \boldsymbol{x})\right) \Big]. \tag{4}$$

The final objective. We parametrize $\mathcal{E}$, $\mathcal{S}$ and $\mathcal{G}$ with three neural networks of appropriate capacity. All networks $\mathcal{E}, \mathcal{S}$ and $\mathcal{G}$ are jointly optimized over total parameters θ and globally on the training set. We further introduce a regularization term over $\boldsymbol{W}$ to encourage sparsity and accordingly compact explanations. The final objective function is now given as

$$\min_\theta \Big[\mathcal{L}_1 + \alpha\, \mathcal{L}_2 + \beta\, \mathbb{E}_{\boldsymbol{x}}[||\boldsymbol{W}_{\boldsymbol{x}}||_{2,1}] \Big], \tag{5}$$

where $\| \cdot \|_{2,1}$ is the group norm $2, 1$, and α, β are balancing coefficients on loss terms. α and β are subject to tuning since a highly compressed representation can cause information loss and harm faithfulness. Figure 1 summarizes our framework as follows

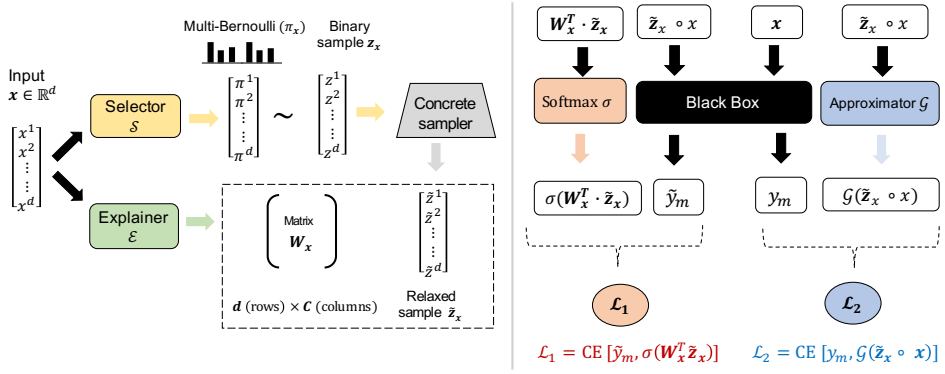


Figure 1: An illustration of AIM pipeline. **Left:** Given an input $\boldsymbol{x}$, the explainer $\mathcal{E}$ produces a local multi-class explanation module $\boldsymbol{W}_{\boldsymbol{x}}$ in which each entry $W_{\boldsymbol{x}}^{i,j}$ representing the relative weight of the ith feature of $\boldsymbol{x}$ to the predicted label $j \in \{1, ..., C\}$. $\mathcal{E}$ is optimized on a local space of perturbations around $\boldsymbol{x}$. Such a space is constructed via the feature selector $\mathcal{S}$ that is simultaneously optimized to generate a high-quality local distribution containing well-behaved neighboring samples. The binary sample $\boldsymbol{z}_{\boldsymbol{x}} \sim \text{Multi-Bernoulli}(\pi_{\boldsymbol{x}})$ is passed through a Gumbel-Softmax sampler for relaxation. We end up with the explanation matrix $\boldsymbol{W}_{\boldsymbol{x}}$ and relaxed samples $\tilde{\boldsymbol{z}}_{\boldsymbol{x}}$. **Right:** The figure illustrates how these output components interact with each other and the input $\boldsymbol{x}$ to form the first and second loss objectives given in Eq. (2) and (4). The final objective in Eq. (5) combines $\mathcal{L}_1$ and $\mathcal{L}_2$ with an additional sparsity term to induce compactness. CE is the cross-entropy function. $\odot$, $\cdot$, and σ denote the element-wise product, inner product and softmax operation respectively.

3.3 Inference

A standard inference strategy is to choose top K features with the highest weights, with K determined in advance. In our framework, the explainer outputs a weight matrix $\boldsymbol{W}_{\boldsymbol{x}}$ size $d \times C$ (recall that d is the number of features and C is the number of target classes). We obtain the black-box's predicted label $j = y_m = \text{argmax}_\text{c} \mathbb{P}_m(Y = c \mid X = \boldsymbol{x})$ and select the corresponding column $W_{\boldsymbol{x}}^{:,j}$ as the weight vector. Features can then be derived accordingly. Though it is intuitive to use $\pi_{\boldsymbol{x}}$ directly for the explanation, doing this may require specifying a certain threshold $\theta \in [0; 1]$. Since $\pi_{\boldsymbol{x}}$ represents

local distributions, choosing the thresholds individually for each input is daunting while setting a global threshold for all inputs is sub-optimal. Moreover, the selection of an ith feature using $\pi_{\boldsymbol{x}}$ (i.e., $\pi^i_{\boldsymbol{x}} \geq \theta$) is independent for each feature, so when combined, they do not guarantee the resulting subsets of features can well approximate the black-box's decisions. On the other hand, the explainer looks into input features all-in-once to settle with good subsets of features. Appendix D provides evidence that inference according to $\boldsymbol{W}_{\boldsymbol{x}}$ is the optimal strategy.

4 EXPERIMENTS

We conducted experiments on various machine learning classification tasks. In the main paper, we focus on NLP classifiers since we believe text data is the most challenging modality. In the following, we discuss the experimental design for textual data.

- **Sentiment Analysis:** The Large Movie Review Dataset **IMDB** (Maas et al., 2011) consists of $50,000$ movie reviews with positive and negative sentiments. The black-box classifier is a bidirectional GRU (Chen et al., 2018) that achieves an 85.4% test accuracy.
- **Hate Speech Detection: HateXplain** is an annotated dataset of Twitter and Gab posts for hate speech detection (Mathew et al., 2021). The task is to classify a post either to be normal or to contain hate/offensive speech. The black-box model is a bidirectional LSTM (Gers et al., 2000) stacked under a standard Transformer encoder layer (Vaswani et al., 2017) of 4 attention heads. The best test accuracy obtained is 69.6%.
- **Topic Classification:** AG is a collection of more than 1 million news articles. **AG News** corpus (Zhang et al., 2015) is constructed by selecting 4 largest classes from the original dataset: World, Sports, Business, and Sci/Tech. We train a word-level convolution neural network (CNN) (LeCun et al., 1995) as a black-box model. It obtains 89.7% accuracy on the test set.

See Appendix A for additional details on our experimental setup and model design. Appendix E further demonstrates the remarkable capability of AIM for generalizing on images and tabular data. Code and data for reproducing our experiments are published at `https://github.com/isVy08/AIM/`.

4.1 PERFORMANCE METRICS & BASELINE METHODS

The task of a saliency-based explainer is to find the subset of input features $\mathbb{S}$ that best mimics the black-box's predictions on the original input. Following the suggestions from Robnik-Šikonja & Bohanec (2018) on desiderata of explanations and related works (Ribeiro et al., 2016; Chen et al., 2018; Schwab & Karlen, 2019; Situ et al., 2021; Gat et al., 2022), Table 1 presents the metrics for quantitative evaluation of word-level explanations. See Appendix B for implementation details.

For text classification tasks, we compare our method against baselines that have done extensive experiments on textual data: L2X (Chen et al., 2018), LIME (Ribeiro et al., 2016), VIBI (Bang et al., 2021) and FI (Gat et al., 2022). Regarding model architectures, note that AIM has been intentionally designed to match those of L2X and VIBI to assure fair comparison. For each method, we tune the remaining hyper-parameters over a wide range of settings and report the results for which Faithfulness is highest (See Appendix H).

4.2 RESULTS

We compare the performance of methods by assessing the extent to which the set of 10 best features satisfies the criteria discussed in Table 1. Except for AIM and FI that do not treat K as a hyper-parameter, all the other baselines are optimized at $K = 10$. Table 2 reports the average results over 5 model initializations. We here show that our method AIM consistently outperforms the baselines on all metrics while achieving a remarkably high level of faithfulness of over 90% across datasets. AIM effectively approximates the black-box predictions with only 10 features, which demonstrates the sufficiency of the selected feature sets. Given the vast combinatorial space of possible subsets of

[2]VIBI (Bang et al., 2021) measures fidelity via the prediction accuracy of the approximator model, whereas we conduct a post-hoc comparison of the black-box's predictions on the original and masked input.

Table 1: Description of quantitative evaluation metrics.

Property	Definition	Metric	Description
Fidelity	How well does the explanation approximate the prediction of the black box model?	Faithfulness / Post-hoc Accuracy[2]	Degree of agreement between the prediction given the full document and the prediction given the selected words in $\mathbb{S}$. A higher value means the explanations are strongly relevant to the black-box's prediction.
Brevity	How concise is the explanation?	Brevity	Number of clusters of duplicates or semantically related words formed over $\mathbb{S}$. A lower value means the tokens are less semantically polarizing and more compact.
Comprehensibility	How well do humans understand the explanation?	Purity	Proportion of stopwords and punctuation included in $\mathbb{S}$. A lower proportion is equivalent to a more meaningful feature set.
Stability	How similar are the explanation for similar instances?	Intersection over Union (IoU)	Proportion of overlapping words in the explanations of two similar documents. We expect the selected features for two such examples overlap in great quantity.
Degree of importance	How well does the explanation reflect the importance of features or parts of the explanation?	Positive Δ log-odds	Difference in the confidence of the black-box model in a prediction before and after masking important words given in $\mathbb{S}$. A higher value indicates $\mathbb{S}$ contains important features.
Degree of importance	How well does the explanation reflect the importance of features or parts of the explanation?	Negative Δ log-odds	Difference in the confidence of the black-box model in a prediction before and after masking unimportant words i.e., words not in $\mathbb{S}$. A lower value indicates features not contained in $\mathbb{S}$ are unimportant.

features, we believe the capacity to efficiently search for a sufficient set of features is what makes AIM stand out from the existing works. Examining Δ log-odds, it is observed that our top 10 features are deemed more important since removing them causes the largest drops in confidence of the black-box model in the original prediction (on the full document). Given an input containing only important features, interestingly there is even a slight increase in confidence when the black-box models make that prediction. Table 2 also reports the average training time (in minutes) for each method. Since AIM is trained in an *instance-wise* manner, AIM matches L2X and VIBI in terms of time efficiency, whereas LIME and FI are extraordinarily time-consuming due to the nature of *additive* models. Learning local explanations instance-wisely also enables AIM to leverage global information, thereby supporting stability (via % IoU) better the baselines.

Table 2: Performance of all methods on 3 datasets at $K = 10$. $\uparrow$ Higher is better. $\downarrow$ Lower is better.

Explainer	**AIM (ours)**	**L2X**	**LIME**	**VIBI**	**FI**
			IMDB		
Purity (%) $\downarrow$	**8.22±0.20**	12.89±0.27	36.55±0.13	30.86±0.20	30.27±0.86
Brevity $\downarrow$	**2.48±0.01**	2.51±0.14	7.73±0.16	3.86±0.23	3.66±0.75
Faithfulness (%) $\uparrow$	**99.62±0.02**	84.80±0.08	79.00±0.21	56.80±0.09	71.70±0.36
IoU (%) $\uparrow$	**6.11±0.09**	4.50±0.14	1.44±0.02	0.59±0.01	3.04±0.01
Positive Δ log-odds $\uparrow$	**7.53±0.03**	2.92±0.11	2.25±0.18	0.09±0.34	2.38±2.35
Negative Δ log-odds $\downarrow$	**-0.20±0.05**	2.63±0.33	5.74±0.06	8.47±0.36	7.15±1.26
Training time (minutes) $\downarrow$	11.19	**6.90**	551.42	8.48	311.42
			HateXplain		
Purity (%) $\downarrow$	**19.78±2.54**	21.87±0.14	37.73±0.17	33.91±0.29	33.13±0.09
Brevity $\downarrow$	**3.88±0.21**	4.36±0.15	7.59±0.20	4.56±0.28	4.23±0.00
Faithfulness (%) $\uparrow$	**92.98±1.17**	75.32±0.03	80.56±0.11	67.25±0.29	66.28±0.68
IoU (%) $\uparrow$	**6.66±0.30**	3.42±0.74	4.40±0.00	2.37±0.08	3.04±0.06
Positive Δ log-odds $\uparrow$	**4.98±0.25**	2.81±0.11	3.18±0.11	1.41±0.13	1.41±0.02
Negative Δ log-odds $\downarrow$	**-1.40±0.16**	1.15±0.27	1.07±0.13	2.50±0.11	2.59±0.02
Training time (minutes) $\downarrow$	3.13	2.43	222.17	**2.15**	162.17
			AG News		
Purity (%) $\downarrow$	**3.83±0.05**	6.64±0.25	29.15±0.06	21.15±0.25	27.61±0.03
Brevity $\downarrow$	**3.39±0.00**	3.94±0.18	8.76±0.16	4.07±0.03	4.91±0.00
Faithfulness (%) $\uparrow$	**97.92±0.05**	90.13±0.26	86.64±0.10	66.58±0.36	76.10±0.11
IoU (%) $\uparrow$	**6.48±0.00**	6.01±0.45	3.52±0.02	1.68±0.02	3.18±0.03
Positive Δ log-odds $\uparrow$	**7.14±0.01**	4.36±0.28	1.38±0.17	0.03±0.22	0.49±0.02
Negative Δ log-odds $\downarrow$	**-1.09±0.02**	0.28±0.29	0.60±0.13	3.88±0.10	2.24±0.00
Training time (minutes) $\downarrow$	**11.44**	22.08	137.02	15.36	30.22

As K increases, the explanation is expected to be more faithful to the black-box model. Since the mechanism of L2X, VIBI, or LIME requires training a new model with the corresponding K, as shown in Figure 2, it may however not guarantee the monotonic behavior. Appendix G analyzes this property in L2X explanations in more detail. We choose to investigate L2X here only since it is the best-performing among the baselines. It is shown that the performance of L2X does not always satisfy monotonicity w.r.t K on a newly trained model: different choices of K can yield different

feature rankings e.g., the features picked by a model trained on top 5 may be considered irrelevant by one trained on top 10. AIM strictly avoids such inconsistency as our framework is not sensitive to K. Table 3 additionally provides 4 examples of the features chosen by AIM in IMDB dataset. This helps shed light on why the black-box model makes a certain prediction, especially the wrong one. A comprehensive qualitative comparison with the baselines on multiple examples can further be found in Appendix C. [3] While explanations from *additive* models (LIME and FI) are contaminated with a larger volume of neutral words, *instance-wise* methods (L2X and VIBI) tend to select more meaningful features. AIM stands out with the strongest compactness by picking up all duplicates and synonyms without compromising predictive performance. Note that LIME suffers from low brevity mainly because its algorithm extracts unique words as explanations. This also means LIME's feature sets tend to be more diverse than the other methods and thus should be more faithful. Our experiment nevertheless shows that this is not the case.

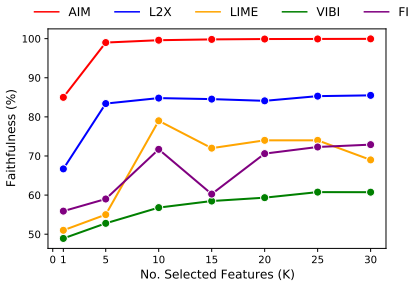


Figure 2: Faithfulness of explanation models at different values of K on IMDB dataset.

Table 3: Ground-truth labels and labels predicted by the black-box model on IMDB movie reviews are given in the first two columns. 10 most relevant words selected by AIM are highlighted in yellow.

Truth	Model	Key words
positive	positive	this movie was a pleasant surprise for me. in all honesty, the previews looked horrible, up until the point where emma thompson and alan rickman appeared. so i rented it with reservation, but i thoroughly enjoyed this movie. it had great acting, a few good plot twists, and, of course, emma thompson and alan rickman. it's definitely worth checking out.
negative	negative	this may just be the worst movie ever produced. worst plot, worst acting, worst special effects...be prepared if you want to watch this. the only way to get enjoyment out of it is to light a match and burn the tape of it, knowing it will never fall into the hands of any sane person again.
positive	negative	to me, "anatomie" is certainly one of the better movies i have seen. i don't think "anatomie" was primarily intended to be a horror movie but a movie questioning the ethics of science. if you watch it with that in mind, it turns into a really good film. the only annoying bit was the awful voice dubbing for the english version. how can you expect any non-german person to listen to these unbearable german accents for two hours ? let native english speakers do the talking or use subtitles instead!!
negative	positive	i have seen this movie several times, it sure is one of the cheapest action flicks of the eighties. so, i think many viewers would definitely change the channel when they come across this one. but, if you are into great trash, "dragon hunt" is made for you. the main characters (the mcnamara twins) are sporting great moustaches and look so ridiculous in their camouflage dresses. one of the best scenes is when one of then gets shot in the leg and is still kicking his enemies into nirvana. this movie is really awful, but then again, it is a great party tape!

4.3 Human Evaluation

We additionally conduct a human experiment to evaluate whether the words selected as an explanation convey sufficient information about the original document to human users. We ask 30 university students to infer the sentiments of 50 IMDB movie reviews, given only 10 key words obtained from an explainer for each review. To avoid confusion, only examples where the black-box model predicts correctly are considered (See Appendix F for the setup). We assess whether the sentiment

Table 4: Human evaluation results on IMDB dataset of AIM, L2X, and LIME.

Explainer	**AIM**	**L2X**	**LIME**
Human accuracy	90.10%	83.03%	84.13%
% Neutral	8.41%	12.22%	19.22%

inferred by humans is consistent with the actual label of a movie review: *human accuracy*. Some reviews are judged as "neutral / can't decide", because the selected key words are neutral, or because positive and negative words are comparable in quantity. We exclude these neutral examples when computing the average accuracy for a participant, but record the proportion of such examples as a proxy measure for purity. The final accuracy is averaged over multiple participants and reported in

[3] All qualitative examples presented in our work are randomly selected from the outputs of the model initialization with the best Faithfulness.

Table 4. It is consistent with our quantitative results that explanations from AIM are perceived to be more informative and contain fewer neural features, thus being more comprehensible to human users.

4.4 MULTI-CLASS EXPLANATION

A novel contribution of our work is the capability of simultaneously explaining multiple decision classes from a single matrix $\boldsymbol{W_x}$. Whereas existing methods often require re-training or re-optimization to predict a different class, our explainer produces class-specific explanations in a single forward pass: given a learned $\boldsymbol{W_x}$, select the column j ($W_{\boldsymbol{x}}^{:,j}$) corresponding to the target class to be explained. To the best of our knowledge, we are the first to propose an explanation module with such a facility.

We assess the quality of multi-class explanations via two modifications of Faithfulness and IoU. The former metric **Class-specific Faithfulness** measures whether the black-box prediction on the explanations aligns with the class being interpreted. The latter **Pairwise IoU** evaluates the overlapping ratio of words in the explanations for a pair of decision classes. Table 5 provides the average results for these metrics, in comparison with LIME and FI. AIM performs surprisingly well on binary classification tasks with the selected feature sets nearly distinctive to each class i.e., overlapping words account only for less than 4%. Faithfulness 98.09% of on the first class, for example, means that given the explanations, the black-box model predicts label 0 for 98.09% of testing examples. Meanwhile, the performance of LIME and FI seems to be no better than random and sensitive to the distribution of classes in the datasets. However, the task gets more challenging as more classes are involved. Since AG News is a dataset of news articles from 4 topics, it is sometimes difficult to clearly distinguish a text between two classes, which we suspect leads to a higher overlapping ratio, thereby harming faithfulness. Regardless, the success on IMDB and HateXplain demonstrates the potential of supporting *counterfactual* explanations that seek to determine which features a black-box classifier attends to when predicting a certain class.

Table 5: Quality of multi-class explanations from AIM, LIME and FI.

Metric	**Class-specific Faithfulness (%)** ↑				**Pairwise IoU (%)** ↓
Target label	0	1	2	3	
IMDB					
AIM	**98.09±0.06**	**98.96±0.05**	-	-	**0.41±0.02**
LIME	50.02±0.15	50.48±0.16	-	-	15.69±0.01
FI	86.32±0.29	15.62±0.36	-	-	92.34±0.08
HateXplain					
AIM	**87.76±0.29**	**88.59±1.88**	-	-	**3.69±0.11**
LIME	24.30±0.12	74.15±0.11	-	-	66.91±0.02
FI	41.21±0.50	60.35±0.40	-	-	76.52±0.05
AG News					
AIM	**73.88±0.47**	**79.13±0.56**	**54.56±0.18**	**84.73±0.38**	**9.24±0.03**
LIME	22.89±0.01	26.45±0.06	25.00±0.02	26.71±0.01	51.09±0.09
FI	25.07±0.43	25.00±0.44	26.18±0.20	26.71±0.01	35.60±0.07

5 CONCLUSION AND FUTURE WORK

We developed **AIM** - a novel model interpretation framework that integrates local *additivity* with *instance-wise* feature selection. The approach focuses on learning attributions across the target output space, based on which to derive important features maximally faithful to the black-box model being explained. We provide empirical evidence further proving the quality of our explanations: compact yet comprehensive, distinctive to each decision class and comprehensible to human users. Exploring causal or counterfactual explanations, especially within our multi-class module is a potential research avenue. Though extension to regression problems and other modalities such as audio or graphical data is straightforward, our future work will conduct thorough experiments on these modalities along with comprehensive comparisons with related baselines. Furthermore, our paper currently focuses on word-level explanations, so there is a chance of discarding positional or phrasal information (e.g., idioms, phrasal verbs). This can be resolved through chunk-level or sentence-level explanations, which will be tackled in future works of ours.

Feature-based Learning for Diverse and Privacy-Preserving Counterfactual Explanations

Vy Vo
Monash University, Australia
CSIRO's Data61, Australia
Tran.Vo@monash.edu

Trung Le
Monash University, Australia

Van Nguyen
Monash University, Australia
CSIRO's Data61, Australia

He Zhao
CSIRO's Data61, Australia

Edwin V. Bonilla
CSIRO's Data61, Australia
Australian National University, Australia

Gholamreza Haffari
Monash University, Australia

Dinh Phung
Monash University, Australia
VinAI Research, Vietnam

ABSTRACT

Interpretable machine learning seeks to understand the reasoning process of complex black-box systems that are long notorious for lack of explainability. One flourishing approach is through counterfactual explanations, which provide suggestions on what a user can do to alter an outcome. Not only must a counterfactual example counter the original prediction from the black-box classifier but it should also satisfy various constraints for practical applications. Diversity is one of the critical constraints that however remains less discussed. While diverse counterfactuals are ideal, it is computationally challenging to simultaneously address some other constraints. Furthermore, there is a growing privacy concern over the released counterfactual data. To this end, we propose a feature-based learning framework that effectively handles the counterfactual constraints and contributes itself to the limited pool of private explanation models. We demonstrate the flexibility and effectiveness of our method in generating diverse counterfactuals of actionability and plausibility. Our counterfactual engine is more efficient than counterparts of the same capacity while yielding the lowest re-identification risks.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Security and privacy** → **Privacy protections**.

KEYWORDS

Explainable AI, Algorithmic Recourse, Privacy

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08…$15.00
https://doi.org/10.1145/3580305.3599343

ACM Reference Format:
Vy Vo, Trung Le, Van Nguyen, He Zhao, Edwin V. Bonilla, Gholamreza Haffari, and Dinh Phung. 2023. Feature-based Learning for Diverse and Privacy-Preserving Counterfactual Explanations. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3580305.3599343

1 INTRODUCTION

The eminence of deep neural networks in recent years has proliferated the use of machine learning in various real-world applications. Such models provide remarkable predictive performance yet often at a cost of transparency and interpretability. This has sparked controversy over whether to rely on algorithmic predictions for high-stakes decision making such as graduate admission [1, 54], job recruitment [3], credit assessment [26] or criminal justice [15, 36]. Progress in interpretable machine learning offers interesting solutions to explaining the underlying behavior of black-box models [37, 41, 53]. One useful approach is through counterfactual examples[1], which sheds light on what modifications to be made to an individual's profile that can counter an unfavorable decision outcome from a black-box classifier. Such explanations explore what-if scenarios that suggest possible recourses for future improvement. Counterfactual explainability indeed has important social implications at both personal and organizational levels. For instance, feedback like 'getting 1 more referral' or 'being fluent in at least 2 languages' would help unsuccessful candidates better prepare for future job applications. By advocating for transparency in decision making, organizations can improve their attractiveness to top talents while inspecting possible prejudice implicitly introduced in historical data and consequentially embedded in the classifiers producing biased decisions.

The ultimate goal of this line of research is to provide realistic guidelines as to what actions an individual can take to achieve a desired outcome. Desiderata of counterfactual explanations have been extensively discussed in previous literature [18, 21, 51, 52].

[1]Counterfactual explanation is sometimes referred to as algorithmic recourse.

To be of practical use, a counterfactual explanation should at least satisfy the following characteristics:

- *Validity:* By definition, a counterfactual example must change the original black-box outcome to a desired one.
- *Sparsity:* Counterfactuals should be close to the original example where a minimal number of features are modified.
- *Actionability:* Counterfactual explanations should only suggest actionable or feasible changes. In particular, changes should be made on mutable features e.g., Work Experience or SAT scores, while leaving immutable features unchanged e.g., Gender or Ethnicity.
- *Diversity:* Diverse explanations are preferable to capture different preferences from the same user so that they can freely explore multiple options to select the best fit.
- *Plausibility*[2]: Plausible or realistic counterfactuals are to obey the input domain and constraints within/among features. For example, Age cannot decrease or be above 200.
- *Scalability:* Inference should be done simultaneously and efficiently for multiple input examples.

Among these desiderata, *diversity* emerges as a non-trivial property to address. Given an instance, a diverse counterfactual engine returns a set of different counterfactual profiles that should all lead to the desired outcome. Ensuring that the entire explanation set satisfies *validity* while dealing with constraints given by *actionability* and *plausibility* poses a computational challenge. *Scalability* becomes another important consideration mainly due to the fact that most of the existing approaches process counterfactuals separately for each input data point. Furthermore, strongly enforcing *sparsity* results in a smaller subset of features that can be changed. This hence can compromise *diversity* since we expect counterfactual states to differ from one to another substantially. On the other hand, there has been a growing concern over the privacy risks of model explanation [32, 46, 48]. Aïvodji et al. [2] points out that diverse counterfactual explanations make the system more vulnerable as the released examples reveal the model decision boundaries and could disclose sensitive information such as health conditions or financial data. A *linkage attack* is one such malicious attempt, which refers to the action of recovering the identity (i.e., re-identifying) of an anonymized record in the published dataset using background knowledge. It is often done by linking records to an external dataset of the population based on the combination of several attributes [12, 28, 44]. Netflix $1M Machine Learning Contest is a notorious data breach, in which the company disclosed a dataset of 100 million subscribers with their movie ratings and preferences. Narayanan and Shmatikov [34] revealed a successful attack of 68% that was easily achieved by cross-referencing the users' dates and precise ratings of 2 movies with a non-anonymous dataset published by IMDb (Internet Movie Database).

Despite an overwhelming number of counterfactual explanation approaches, only a few works tackle diverse counterfactual generation [5, 22, 33, 40, 43]. However, the trade-offs between *diversity* and the aforementioned constraints, including privacy protection, have not been well studied in previous papers (See Table 1 for comparison). Filling this gap, our work proposes a novel learning-based framework that effectively addresses all the above desiderata while mitigating the re-identification risk. From a methodological perspective, **our method diverges markedly from existing approaches in the following ways:**

Firstly, we reformulate the combinatorial search task into a stochastic optimization problem to be solved via gradient descent. Unlike most previous methods that perform optimization per-input basis, we employ amortized inference to generate diverse counterfactual explanations efficiently. Amortization has been previously adopted wherein a counterfactual generative distribution is modelled via Markov Decision Processes [52] or Variational Auto-encoders [9, 30, 38]. On one hand, none of these amortized methods addresses *diversity*. On the other, we here take a different approach: we construct a learnable generation module that directly models the conditional distributions of individual features such that they form a valid counterfactual distribution when combined.

Another point of difference of ours lies in the usage of Bernoulli sampling to ensure *sparsity*. In prior works, standard metrics such as L1 or L2 are often used to penalize the distance between the counterfactual and original data point. Verma et al. [51] criticizes this approach as unnatural, especially for categorical features. Avoiding the use of distance measures, we optimize along a feature selection module to output the likelihood of the feature being mutated. This module can be adapted to any user-defined constraints about the mutability of features.

Finally, we go beyond existing approaches by tackling the constraint of privacy preservation exposed to diverse explanations. The key strategy is to discretize continuous features and operate the counterfactual generation engine in the categorical feature space. Discretization is closely related to the generalization technique used in privacy-preserving data mining (PPDM) [12, 31]. It is also treated as a subroutine to analyze the composition of differential privacy algorithms [14, 17]. The idea is that it is by nature easier to uniquely identify a profile based on continuous features, so discretization is expected to increase the quantities of profiles linked back to a certain group of attributes. Another defense effort against linkage attack can be found in [16]. The paper proposes an algorithm named CF-K that heuristically searches for an equivalence class for each counterfactual instance such that every record is indistinguishable from at least $k-1$ others. CF-K is viewed as an add-on that theoretically can be implemented on top of any counterfactual generative system. However, in practice, this strategy is extremely expensive since it requires repetitively querying the model explainer for a possibly larger number of counterfactuals than requested. Though sharing the same motivation, we here contribute a counterfactual explanation model with a built-in privacy preservation functionality.

Our contributions can be summarized as follows:

- We introduce **L**earning to **C**ounter (**L2C**) - a stochastic feature-based approach for learning counterfactual explanations that address the counterfactual desirable properties in a single end-to-end differentiable framework.
- Through extensive experiments on real-world datasets, L2C is shown to balance the counterfactual trade-offs more effectively than the existing methods and achieve diverse explanations with the lowest re-identifiability risk. To the best of our knowledge,

[2]The terms *plausibility* and *feasibility* are often used interchangeably.

L2C is the first amortized engine that supports diverse counterfactual generations with privacy-preservation capability.

2 RELATED WORKS

Recent years have seen an explosion in the literature on counterfactual explainability, from works that initially focused on one or two characteristics or families of models to those that can deal with multiple constraints and various model types. There have been many attempts to summarize major themes of research and discuss open challenges in great depth. We therefore refer readers to [18, 21, 51] for excellent surveys of methods in this area. We here focus on reviewing algorithms that can support diverse (or at least multiple) local counterfactual generations.

Dealing with the combinatorial nature of the task, earlier works commonly adopt mixed integer programming [43], genetic algorithms [45], or SMT solvers [20]. Another recent popular approach is gradient-based optimization [5, 33], which involves iteratively perturbing the input data point according to an objective function that incorporates desired constraints. The whole idea of *diversity* is to explore different combinations of features and feature values that can counter the original prediction while accommodating various user needs. To support *diversity*, Russell [43] in particular enforces hard constraints on the current generations to be different from the previous ones. Such a constraint will however be removed whenever the solver cannot be satisfied. Meanwhile, Mothilal et al. [33] and Bui et al. [5] add another loss term for *diversity* using Determinantal Point Processes [25], whereas the other works only demonstrate the capacity to generate multiple counterfactuals via empirical results. All of the aforementioned algorithms are computationally expensive in that input data points are handled singly and individual runs are additionally required to produce several counterfactuals. Reducing computational costs, Redelmeier et al. [40] attempts to model the conditional likelihood of mutable features given the immutable features using the training data. They then adopt Monte Carlo sampling to generate counterfactuals from this distribution and filter out samples that do not meet counterfactual constraints. Given such a generative distribution, sampling of counterfactuals can therefore be done straightforwardly. Amortized optimization is another strategy to improve inference speed [9, 30, 38, 52].

In response to the privacy warning about model explanations [32, 46, 48], several defense strategies have been introduced to alleviate the risks. With strong theoretical guarantees, differential privacy [11] stands out as the promising solution to preventing *member inference attack* and *model stealing attack* [2, 32, 47]. With regards to linkage attacks, CF-K [16] is the only work we are aware of that tackles linkage attack in counterfactual explanations.

3 STOCHASTIC FEATURE-BASED COUNTERFACTUAL LEARNING

3.1 Problem setup

Let $\mathcal{X}$ denote the input space where $\boldsymbol{x} = [x_i]_{i=1}^N$ is an input vector with N features of both continuous and categorical types. As discussed previously, we discretize the continuous features into equal-sized buckets, which gives us an input of N categorical features wherein each feature x_i has c_i levels. We apply one-hot encoding on each feature and flatten them into a single input vector $\boldsymbol{z} \in \{0,1\}^D$ where $D = \sum_{i=1}^N c_i$. Concretely, feature x_i is now represented by the vector $\boldsymbol{z}_i \in \mathbb{O}_{c_i}$ where the set of one-hot vectors $\mathbb{O}_{c_i}$ is defined as $\{0,1\}^{c_i} : \sum_{j=1}^{c_i} z_{ij} = 1$.

Let f be the black-box classifying function and $y = f(\boldsymbol{x})$ be the decision outcome on the input $\boldsymbol{x}$. A valid counterfactual example $\widetilde{\boldsymbol{x}}$ associated with $\boldsymbol{x}$ is one that alters the original outcome y into a desired outcome $y' \neq y$ with $y' = f(\widetilde{\boldsymbol{x}})$. Let $\widetilde{\boldsymbol{z}}$ denote the corresponding one-hot representation of $\widetilde{\boldsymbol{x}}$.

Actionability indicates that some features can be *mutable* (i.e., changeable), while others should be kept *immutable* (i.e., unchangeable). Without loss of generality, let us impose an ordering on the set of N features such that the first K features are mutable features (i.e., the ones that can be modified) and denote $\mathbb{K} := \{1, ..., K\} \subset \{1, ..., N\}$. For each mutable feature (i.e., x_i or the one-hot vector $\boldsymbol{z}_i$ with $i \in \mathbb{K}$), we aim to learn a local feature-based perturbation distribution $P(\widetilde{\boldsymbol{z}}_i \mid \boldsymbol{z})$ where $\widetilde{\boldsymbol{z}}_i \in \mathbb{O}_{c_i}$, while leaving the immutable features unchanged.

It is worth noting that our method functions equally well on heterogeneous data where only categorical features are one-hot encoded while continuous features are retained at their original values. However, we believe that performing data discretization (or generalization in terms of PPDM) initially and deploying the classifiers in the discrete feature space would provide better privacy protection. For the purpose of comparing our prototype with existing approaches, we follow the standard practice of explaining classification models trained on the mixed dataset of continuous and categorical features. To make it compatible with the discretization subroutine of our framework, we represent the prediction on a transformed (fully categorical) input vector with the prediction on the input where the categorical values associated with mutable continuous features are substituted with the middle point of the corresponding intervals. We refer to this mechanism as **one-hot decoding**, which will be detailed shortly.

3.2 Methodology

We now detail how L2C works and addresses each counterfactual constraint. The explanation pipeline of L2C is depicted in Figure 1.

For each mutable feature $\boldsymbol{z}_i$ with $i \in \mathbb{K}$, we learn a **local feature-based perturbation distribution** $P(\widetilde{\boldsymbol{z}}_i \mid \boldsymbol{z})$ (i.e., $\widetilde{z}_i \in \mathbb{O}_{c_i}$), which is a categorical distribution $\text{Cat}(\boldsymbol{p}_i \mid \boldsymbol{z})$ with category probability $\boldsymbol{p}_i = \left[p_{i1}, p_{i2}, ..., p_{ic_i}\right]$. We form a counterfactual example $\widetilde{\boldsymbol{z}}$ by concatenating $\widetilde{\boldsymbol{z}}_i \sim \text{Cat}(\boldsymbol{p}_i \mid \boldsymbol{z})$ for the mutable features and $\boldsymbol{z}_i$ for the immutable features. To achieve *validity*, we learn the local feature-based perturbation distribution by maximizing the chance that the counterfactual examples $\widetilde{\boldsymbol{z}}$ counter the original outcome on $\boldsymbol{x}$. Additionally, learning local feature-based perturbation distributions over the mutable features allows us to conduct a global counterfactual distribution $P(\widetilde{\boldsymbol{z}} \mid \boldsymbol{z})$ over the counterfactual examples $\widetilde{\boldsymbol{z}}$ defined above. Sampling from this distribution naturally leads to multiple counterfactual generations efficiently, and we also expect that individual samples $\widetilde{\boldsymbol{z}}_i$ together can form diverse combinations of features, thereby promoting *diversity* within the generative examples.

Table 1: Desiderata comparison of related counterfactual explanation methods. *Privacy refers to whether the output data is protected against linkage attack or re-identification risk. L2C satisfies all of these critical constraints.

Method	Sparsity	Actionability	Diversity	Plausibility	Scalability	Privacy*
L2C (Ours)	✓	✓	✓	✓	✓	✓
DICE [33]	✓	✓	✓			
COPA [5]	✓	✓	✓			
MCCE [40]	✓	✓	✓		✓	
Coherent CF [43]	✓		✓			
MACE [20]	✓	✓	✓			
MOC [8]	✓	✓		✓		
CERTIFAI [45]		✓				
Feasible-VAE [30]	✓	✓		✓	✓	
FastAR [52]	✓	✓		✓	✓	
CRUDS [9]		✓		✓	✓	
C-CHVAE [38]		✓			✓	
CF-K [16]						✓

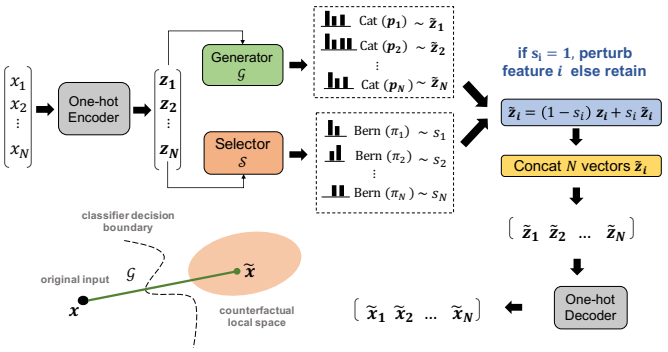


Figure 1: For illustration purposes only, all features are assumed mutable in the figure. We first discretize the continuous features of input x and one-hot encode all features into representations z. For every feature i, the generator $\mathcal{G}$ learns a local perturbation distribution $\text{Cat}(p_i|z)$ so that together they form a distribution of diverse counterfactual representations $\widetilde{z}$. Simultaneously, the selector $\mathcal{S}$ learns to output the distribution Multi-Bernoulli$(\pi|z)$ capturing the probability of each feature i being modified. Every feature sample pair $(\widetilde{z}_i, s_i)$ is passed through an operation in the blue box, which decides whether to accept the change being made to the feature i given by $\widetilde{z}_i$. The output is then decoded into the representations $\widetilde{x}$ compatible with the black-box system. $\mathcal{G}$ and $\mathcal{S}$ are jointly trained via back-propagation according to Eq. (4). Intuitively, $\mathcal{G}$ aims to construct a "bridge" across the decision boundary travelling from the input to a local space of counterfactuals.

As previously discussed, too much of *diversity* can compromise *sparsity*. Dealing with this constraint, for each mutable feature z_i, we propose to learn a **local feature-based selection distribution** that generates a random binary variable $s_i \sim \text{Bernoulli}(\pi_i \mid z)$ wherein we replace z_i by $\widetilde{z}_i \sim \text{Cat}(\boldsymbol{p}_i \mid z)$ if $s_i = 1$ and leave $\tilde{z}_i = z_i$ if $s_i = 0$. Therefore, the formula to update $\widetilde{z}_i$ is

$$\widetilde{z}_i = (1 - s_i)z_i + s_i\widetilde{z}_i.$$

The benefit of having $\boldsymbol{\pi} = [\pi_i]_{i\in\mathbb{K}}$ is thus to control *sparsity* by adding one more channel to decide if we should modify a mutable feature z_i. Appendix D presents an ablation study showing that without the selection distribution, the perturbation distribution alone can generate diverse counterfactuals yet requires changing plenty of mutable features. Meanwhile, optimizing the selection distribution jointly helps harmonize the trade-off between *diversity* and *sparsity*.

3.3 Optimization Objective

In this section, we explain how to design the building blocks of our framework L2C. As shown in Figure 1, our framework consists of two modules: a **counterfactual generator** $\mathcal{G}$ and a **feature selector** $\mathcal{S}$. The **counterfactual generator** $\mathcal{G}$ is used to model the feature-based perturbation distribution, while **feature selector** $\mathcal{S}$ is employed to model the feature-based selection distribution.

Specifically, given a one-hot vector representation $\boldsymbol{z}$ of a data example $\boldsymbol{x}$, we feed $\boldsymbol{z}$ to $\mathcal{G}$ to form $\mathcal{G}(z) = [\mathcal{G}_i(z)]_{i\in\mathbb{K}}$. We then apply the softmax activation function to $\mathcal{G}_i(z)$ to define the feature-based local distribution (i.e., $\text{Cat}(\boldsymbol{p}_i \mid z)$)) for z_i as

$$p_{ij}(z) = \frac{\exp\{\mathcal{G}_{ij}(z)\}}{\sum_{k=1}^{c_i}\exp\{\mathcal{G}_{ik}(z)\}}, \forall j = 1, ..., c_i. \tag{1}$$

The module $\mathcal{S}$ takes $\boldsymbol{z}$ to form $\mathcal{S}(z) = [\mathcal{S}_i(z)]_{i\in\mathbb{K}}$. We then apply the Sigmoid function to $\mathcal{S}_i(z)$ to define the feature-based selection distribution (i.e., $\text{Bernoulli}(\pi_i \mid z)$) for z_i as

$$\pi_i(z) = \frac{1}{1 + \exp\{-\mathcal{S}_i(z)\}}.$$

To encourage *sparsity* by reducing the number of mutable features chosen to be modified, we regularize $\mathcal{S}$ through L1-norm $\|\boldsymbol{\pi}(z)\|_1$ with $\boldsymbol{\pi}(z) = [\pi_i(z)]_{i\in\mathbb{K}}$.

To summarize, given a one-hot vector representation $\boldsymbol{z}$ of a data example $\boldsymbol{x}$, we use $\mathcal{G}$ to work out the local feature-based perturbation distribution $\text{Cat}(\boldsymbol{p}_i(\boldsymbol{z}))$ for every $i \in \mathbb{K}$. We then sample $\widetilde{\boldsymbol{z}}_i \sim \text{Cat}(\boldsymbol{p}_i(\boldsymbol{z}))$ for every $i \in \mathbb{K}$. Subsequently, we use $\mathcal{S}$ to work out the local feature-based selection distribution $\text{Bernoulli}(\pi_i(\boldsymbol{z}))$ for every $i \in \mathbb{K}$. We then sample $s_i \sim \text{Bernoulli}(\pi_i \mid \boldsymbol{z})$ and update $\widetilde{\boldsymbol{z}}_i = (1 - s_i)\boldsymbol{z}_i + s_i\widetilde{\boldsymbol{z}}_i$ for every $i \in \mathbb{K}$. Finally, we concatenate $\widetilde{\boldsymbol{z}}_i$ for $i \in \mathbb{K}$ and $\boldsymbol{z}_i$ for $i \notin \mathbb{K}$ to form the counterfactual example $\widetilde{\boldsymbol{z}}$.

$\mathcal{G}$ and $\mathcal{S}$ are parameterized with neural networks over total parameters θ. For $\widetilde{\boldsymbol{z}}$ to be a *valid* and *sparse* counterfactual associated with a desired outcome y', we propose the following criterion

$$\min_\theta \Big[\mathbb{E}_{\widetilde{\boldsymbol{z}}}\big[\text{CE}(f(\widetilde{\boldsymbol{z}}), y')\big] + \alpha\, \mathbb{E}_{\boldsymbol{z}}\big[\|\boldsymbol{\pi}(\boldsymbol{z})\|_1\big]\Big], \tag{2}$$

where f is the black-box function, CE is the cross-entropy loss, $\|\cdot\|_1$ is L1-norm, α is a loss weight.

One-hot decoding. Recall that $\widetilde{\boldsymbol{z}}$ formed by concatenating many one-hot vectors is an incompatible representation to the classifier f, which in fact requires both continuous and one-hot features. We make a design choice of reconstructing the continuous features by taking the middle point of the range corresponding to the selected level. Specifically, the input to the one-hot decoder is $\widetilde{\boldsymbol{z}} = [\widetilde{\boldsymbol{z}}_i]_{i=1}^N$. If the feature i originally is a categorical feature, we set $\widetilde{x}_i = \widetilde{\boldsymbol{z}}_i$. Otherwise, we set $\widetilde{x}_i = a_i + \frac{(2k-1)(b_i-a_i)}{2c_i}$, which is the middle point of the k-th interval $[a_i + (k-1)(b_i-a_i)/c_i, a_i + k(b_i-a_i)/c_i]$ where $[a_i, b_i]$ is the original value range of the feature i and $\widetilde{\boldsymbol{z}}_i$ corresponds to the level $k \in \{1, ..., c_i\}$ (i.e., $\widetilde{\boldsymbol{z}}_{ik} = 1$ and $\widetilde{\boldsymbol{z}}_{ij} = 0$ if $j \neq k$). Note that one-hot decoding is only applied when a continuous feature is indicated by $\mathcal{S}$ to be mutable ($s_i = 1$). Otherwise, we revert to the original continuous value. Formally, we rewrite

$$\widetilde{x}_i = (1 - s_i)x_i + s_i \sum_{j=1}^{c_i} \widetilde{\boldsymbol{z}}_{ij}\Big[a_i + \frac{(2j-1)(b_i-a_i)}{2c_i}\Big]. \tag{3}$$

To assure model differentiability for training, the one-hot vector $\widetilde{\boldsymbol{z}}_{ij}$ is relaxed into its continuous representation by using Gumbel-Softmax reparametrization trick, which is detailed in Section 3.4.

The final optimization objective is now given as

$$\min_\theta \Big[\mathbb{E}_{\widetilde{\boldsymbol{x}}}\big[\text{CE}(f(\widetilde{\boldsymbol{x}}), y')\big] + \alpha\, \mathbb{E}_{\boldsymbol{z}}\big[\|\boldsymbol{\pi}(\boldsymbol{z})\|_1\big]\Big]. \tag{4}$$

Plausibility. A counterfactual generative engine needs to ensure explanations are realistic for real-world applications. There are two common types of plausibility constraints: **Unary** and **Binary** monotonicity constraints. The former deals with individual features (e.g., Age cannot decrease) while the latter is concerned with the correlation of a pair of features (e.g., increasing in Education level increases Age). To handle binary constraints on heterogeneous data is not as straightforward as unary constraints. We thus delay the discussion on binary constraints until Section 5.4 and focus on unary constraints in the main analysis.

Dealing with such a constraint, one can simply eliminate any counterfactuals violating the constraints during inference time. This however creates additional computational overhead and may compromise some desiderata. There are only few attempts addressing feature constraints in counterfactuals, notably Mahajan et al. [30], Verma et al. [52]: Verma et al. [51] incorporate hard conditions in a Markov Decision process where a feature gets updated only if the corresponding action does not violate the constraints. Meanwhile, Mahajan et al. [30] includes a hinge loss into the loss function for unary features, while specifically learning a separate linear model for every feature pair subject to a binary constraint. For L2C, the learnable local distributions can be used for this purpose conveniently. Our proposed strategy is to impose rule-based unary constraints on related features in the optimization process. Technically, for every feature to be perturbed with a non-decreasing (non-increasing) constraint, we penalize the probabilities corresponding to lower (higher) levels towards zero by multiplying them with a positive infinitesimal quantity. Concretely, given a mutable feature i under monotonic constraints, let $l \in \{1, ..., c_i\}$ denote the current state - the level corresponding to $\boldsymbol{z}_i$ (i.e., $z_{il} = 1$ and $z_{ij} = 0$ if $j \neq l$). Let us denote the restricted set of levels as $\mathbf{C}_i = \{1, ..., l-1\}$ if the feature is non-decreasing and $\mathbf{C}_i = \{l+1, ..., c_i\}$ if it is non-increasing. The perturbation distribution $\text{Cat}(\boldsymbol{p}_i \mid \boldsymbol{z})$ for $\boldsymbol{z}_i$ given in Eq. (1) now becomes

$$p_{ij}(\boldsymbol{z}) \propto \varepsilon^{\mathbf{1}_{\mathbf{C}_i}(j)} \times \frac{\exp\{\mathcal{G}_{ij}(\boldsymbol{z})\}}{\sum_{k=1}^{c_i} \exp\{\mathcal{G}_{ik}(\boldsymbol{z})\}}, \forall j = 1, ..., c_i, \tag{5}$$

where $\mathbf{1}_{\mathbf{C}_i}(.)$ is the indicator function such that $\mathbf{1}_{\mathbf{C}_i}(j) = 1$ if $j \in \mathbf{C}_i$ and $\mathbf{1}_{\mathbf{C}_i}(j) = 0$ otherwise, meaning that the probabilities at the other levels are untouched. We here explicitly force the model to generate more samples at the higher (lower) levels while maintaining differentiability of the objective function. We choose $\varepsilon = e^{-10}$ in our experiments, but any positive value arbitrarily close to zero would suffice.

3.4 Reparameterization for Continuous Optimization

Our L2C involves multiple sampling rounds back and forth to optimize the networks. To make the process continuous and differentiable for training, we adopt the reparameterization tricks [19, 29]:

1) Sampling $\widetilde{\boldsymbol{z}}_i \sim Cat(\boldsymbol{p}_i \mid \boldsymbol{z})$. : To obtain differentiable counterfactual samples, we adopt the classic temperature-dependent Gumbel-Softmax trick [19, 29]. Given the categorical variable $\mathbf{z}_i$ with category probability $\big[p_{i1}, p_{i2}, ..., p_{ic_i}\big]$. The relaxed representation is sampled from the Categorical Concrete distribution as $\widetilde{\boldsymbol{z}}_i \sim \text{Cat-Concrete}(\log p_{i1}, ..., \log p_{ic_i})$ by

$$\widetilde{\boldsymbol{z}}_{ij} = \frac{\exp\big\{(\log p_{ij}(\boldsymbol{z}) + G_j)/\tau\big\}}{\sum_{k=1}^{c_i} \exp\big\{(\log p_{ik}(\boldsymbol{z}) + G_k)/\tau\big\}}.$$

with temperature τ, random noises G_j independently drawn from Gumbel distribution $G_t = -\log(-\log u_t)$, $u_t \sim \text{Uniform}(0, 1)$. As discussed, we apply this mechanism consistently to the one-hot representations of all features. The continuous relaxation of Eq. (3) can be gained by simply using the one-hot relaxation $\widetilde{\boldsymbol{z}}_i$.

2) Sampling $s_i \sim Bernoulli(\pi_i \mid \boldsymbol{z})$. : We again apply the Gumbel-Softmax trick to relax Bernoulli variables of 2 categories. With temperature τ, random noises G_{i0} and $G_{i1} \sim G_t = -\log(-\log u_t)$, $u_t \sim \text{Uniform}(0, 1)$, the continuous representation s_i is sampled from Binary Concrete distribution as $s_i \sim \text{Bin-Concrete}(\pi_i, 1 - \pi_i)$ by

$$s_i = \frac{\exp\{\big(\log \pi_i(\boldsymbol{z}) + G_{i1}\big)/\tau\}}{\exp\{\big(\log(1 - \pi_i(\boldsymbol{z})) + G_{i0}\big)\}/\tau\} + \exp\{\big(\log \pi_i(\boldsymbol{z}) + G_{i1}\big)/\tau)\}}.$$

4 EXPERIMENTAL SETUP

We experiment with 4 popular real-word datasets: German Credit [10], Adult Income [24] Graduate Admission [1] and Student Performance [7]. For each dataset, we select a fixed subset of immutable features based on our domain knowledge and suggestions from [52]. We reserve the privacy analysis for German Credit and Adult Income datasets, which contain personal financial information and various attributes through which data subjects can be re-identified [16]. While implementing the black-box classifiers and the baseline methods, we standardize numerical features to unit variance and one-hot encode categorical features. Note again that, for our method only, we discretize numerical features into equal-sized buckets and decode the numerical features back to their original representations whenever necessary to consult the black-box model. Appendix A describes our tasks and model design in greater detail. Our code repository can be accessed at **https://github.com/isVy08/L2C/**.

Table 2: Description of quantitative evaluation metrics. $\mathbb{C}$ denotes a set of counterfactual examples generated by an algorithmic recourse approach for a given input instance.

Desiderata	Metric	Description
Validity	`Validity`	Proportion of samples in $\mathbb{C}$ can counter the original black-box decision outcome.
	`Coverage`	`Coverage` = 100% if there exists at least 1 valid counterfactual in $\mathbb{C}$.
Sparsity/ Actionability	`Sparsity`	Proportion of features kept unchanged, averaged over the number of samples in $\mathbb{C}$.
Diversity	`Diversity`	Hamming distance of a pair of counterfactual samples across all features where numerical features are discretized. The metric is averaged over all pairs of samples in $\mathbb{C}$.
Sparsity – Diversity Balance	`Harmonic mean`	F-measure of `Diversity` and `Sparsity` = 2 · `Diversity` · `Sparsity`/(`Diversity` + `Sparsity`).
Plausibility	`Unary`	Proportion of examples $\mathbb{C}$ meeting the unary monotonic constraints, averaged over the number of features subject to constraints.

Performance metrics. Following the past works Mothilal et al. [33], Redelmeier et al. [40], Verma et al. [52], Table 2 outlines the commonly used metrics for quantitatively assessing the desirability of counterfactual explanations. As for *diversity*, a widely adopted measure is the pairwise distance between counterfactual examples, with distance defined separately for numerical and categorical features [33, 40]. Though this approach is meaningful for interpreting categorical features, we however find it quite obscure for numerical features. This motivates us to discretize numerical features again when computing `Diversity`, which captures how often a feature gets altered as well as how much the change is - specifically via how often it switches to a different categorical level. The computation of `Diversity` only considers valid counterfactuals, so if valid counterfactuals are none, `Diversity` is set to zero. It is worth noting that there fundamentally exists a trade-off between *sparsity* and *diversity*. To quantify how well a method can balance these two properties, we suggest taking `Harmonic mean` of `Diversity` and `Sparsity`, motivated by the development of F1-score in measuring Precision against Recall. For metrics used in privacy analysis, refer to Section 5.2.

Evaluation setup. We consider a general setting of binary classification where a counterfactual outcome y' is opposite to the original outcome y, whether y has label 1 or 0. From each method, we generate a set of 100 counterfactual explanations. During generation, most methods, including ours, require multiple iterations of searching for the optimal set of counterfactuals based on the optimization constraints. To assure a fair comparison on efficiency, a global maximum time budget of 5 minutes is imposed to search for a set of 100 counterfactuals per input sample. We compare our method top-performing baselines that support diverse counterfactual explanations: DICE [33], MCCE [40] and COPA [5]. DICE offers several search strategies: Random, KD tree, or Genetic algorithm. DICE-KDTree was consistently reported to fail across datasets [52], so we exclude it from our evaluation. We do not consider MACE [20] since it is extremely expensive on large datasets [52] and often fails to converge in our experiments, nor MOC [8] due to the lack of Python implementation.

5 RESULTS AND DISCUSSION

5.1 Counterfactual Explanation Desiderata

We first study whether an algorithmic recourse approach generates a set of diverse counterfactuals without sacrificing the other desiderata. Note that COPA has only been shown to work effectively on linear classifiers. Table 3 reports the average results over 5 model initializations.

Under the same time budget, our method L2C succeeds in generating 100% valid counterfactuals with full coverage. Together with DICE, L2C first satisfies the most important criterion of a counterfactual explanation and resolves the trade-off against *validity*. Recall that we have specified a fixed set of immutable features for each dataset, based on which we can work out the minimum sparsity threshold a counterfactual explanation should adhere to (i.e., % immutable features). *Actionability* can then be assessed by comparing `Sparsity` with this level to determine if a method satisfies the mutability of features. An adequate explanation must achieve at least this level of sparsity. MCCE evidently fails to fulfill this constraint on Adult Income and Student Performance datasets.

Our reported results here are obtained under no other conditions than the constraints related to feature immutability and monotonicity described in Table **??**. Nevertheless, we would like to highlight the flexibility of our framework in controlling the quality of counterfactual generations during inference. Users can freely specify any sparsity threshold or additional conditions of interest to filer out unsatisfactory examples without re-training or re-optimization as in methods like DICE. Specifically, DICE employs gradient search directly on each query according to a selected set of weighting hyperparameters for each term in the objective function. To get a less sparse or more diverse example than the current generation, one needs to activate a new search routine.

Too many constraints or too much sparsity clearly affects the diversity level of the counterfactual set. Maintaining a high `Harmonic mean` scores while satisfying almost all feature constraints demonstrates that L2C can effectively manage these trade-offs. The fact that L2C converges to valid counterfactuals with minimal violation in such a short inference time can be attributed to the practice of injecting hard constraints during optimization and global training

does enhance the effect. It certainly helps circumvent the burden of heuristically eradicating violated samples. Notice also that our quantitative results align with the descriptions in Table 1. None of the diverse counterfactual explanation approaches address *plausibility* thoroughly whereas those reported to support the feasibility of features do not guarantee *diversity*. Appendix C provides empirical evidence for this claim, in which we compare L2C with popular amortized algorithmic recourse approaches and demonstrate our consistent superiority in generating diverse explanations efficiently without violating the required constraints (See Table 7).

5.2 Re-identification Risk Analysis

Preliminaries. We start by reviewing the fundamental concepts related to a public dataset:

- *Identifiers*: Attributes that uniquely identify an individual. Identifiers can be a person's full name, government tax number or driver's license number.
- *Quasi-identifiers*: Attributes that themselves do not uniquely identify a person, but when combined are sufficiently correlated to at least one individual record. For example, the combination of gender, birth dates and ZIP codes can re-identify 87% of American residents [50].
- *Sensitive attributes*: Attribute that are protected against unauthorized access. Sensitive data is confidential and if leaked could harm personal safety or emotional well-being. Examples are salary, medical conditions, salary, criminal histories, or phone numbers.
- *Equivalence class*: An equivalence class is a group of records with identical quasi-identifiers.

Every public dataset must first be anonymized by removing identifiers. However, the data may still be vulnerable to re-identification attacks due to the potential existence of quasi-identifiers. To quantify the level at which a dataset is susceptible to re-identification risk, the following 3 metrics are commonly used:

- `k-Anonymity` [44]: A dataset satisfies k-anonymity if for each record in the dataset, the quasi-identifiers are indistinguishable from at least $k-1$ other people also in the dataset.
- `l-Diversity` [28]: A dataset has l-diversity if, for every equivalence class, there are at least l distinct values for each sensitive attribute.
- `k-Map` [12]: Given an auxiliary dataset used for re-identification (e.g., US Census or IMDb dataset in the Netflix example), so-called the 'attack' dataset, a dataset satisfies k-map if every equivalence class is mapped to at least k records in the 'attack' dataset.

Evaluation metrics. Suppose a company releases an API that permits users to query a set of counterfactual examples. We now analyze the level of privacy leakage associated with the output data, by quantifying the percentage of successful attacks w.r.t the aforementioned metrics. Respectively, we measure (1) `1-Anonymity`: *% equivalence classes with only* $k = 1$ *member*, (2) `1-Diversity`: *% equivalence classes with* $l = 1$ *value for a sensitive attribute*, (3) `1-Map`: *% counterfactual examples exactly matched with any single record in an 'attack' dataset*. Notice that given a k-anonymized dataset, the existence of a one-to-one mapping with the 'attack' dataset means the released dataset fails `k-Map`.

Experiments. By definition, violations w.r.t `k-Anonymity` and `l-Diversity` are computed against the output set of examples, while `k-Map` requires an external dataset. For Adult Income, we choose the validation set as the 'attack' set. For German Credit, there exists multiple versions of this dataset across the literature. The one used in our main analysis is adopted from [52], which has been subject to pre-processing. We use another version published by Penn State University[3] for re-identification. We assume these hold-out sets belong to some larger datasets of population availably accessed by the public. Quasi-identifiers and sensitive attributes are given in Appendix A. Following Goethals et al. [16], we consider the black-box predicted label as part of the quasi-identifiers.

Let's first look into the attacks on the raw output explanations presented in Table 4. We here assume that the attacker's goal is to collect as many examples as possible without caring about which one is valid. If the attacker has no information about how the data is discretized, it is much less likely to find exact matches in the 'attack', thereby reducing the re-identifiability of L2C data. Now we assume the attacker gets access to both the API and our discretization mechanism. They therefore could correspondingly discretize the data in the 'attack' set and retrieve the matches. Our analysis assumes this worse scenario, meaning that for L2C only, we compute `1-Map` against the discretized data. The entries `N/A` in Table 4 are due to the fact that no counterfactual example in the data of DICE-Genetic or COPA has matches, so their robustness remains unverifiable in our experiment. We must highlight that no match does not necessarily translate to zero privacy risk. We also note that such a case is different from L2C, whose result is nearly 0.00%. L2C in fact still returns matches for some records wherein we achieve `k-Map` of $2-3$ specifically. Overall, L2C yields the lowest re-identifiability risk. Another interesting observation on German Credit is that although DICE performs well on `k-Anonymity`, the number of attacks on `l-Diversity` is dramatically high. This sheds light on the limitation of `k-Anonymity` discussed in [28] about *Homogeneity attacks* and *Background knowledge attacks*. Basically, attacking a k-anonymized dataset with a high k can still reveal some private information of the data subjects because profiles in the same equivalence class are similar (very few distinct values in sensitive attributes), or the adversary has some background knowledge that any help narrow down possible values.

Privacy under CF-K. We here investigate the effectiveness of the idea behind CF-K proposed in [16]. CF-K searches for an equivalence class for each counterfactual example and suggests only publishing profiles of at least k-sized equivalence class. Since the authors do not publish their codes, plus it is hugely time-consuming to run on models like DICE, we extend the above experiment and examine the effect when every output counterfactual set is 2-anonymized. Specifically, for every set of 100 generations, we remove records not belonging to any equivalence classes. Given now that the data is now 2-anonymized, we evaluate attacks against `l-Diversity` and `k-Map`. We also measure % of valid counterfactuals left in the set, assuming that a user requests for 100 per instance. The more valid examples lost from a model explainer imply that it will be more costly to search for sufficient equivalence classes for every instance. Figure 2 depicts that in most cases, k-anonymization enhances the

[3] https://online.stat.psu.edu/stat857/node/215/

Table 3: Desirability of counterfactual explanation methods. ↓ Lower is better. ↑ Higher is better. Bold / Underline indicates the best / second-best performance for each dataset. `Time` records total inference time in seconds.

Method	Sparsity (%)↑	Diversity (%)↑	Harmonic Mean (%)↑	Validity (%)↑	Coverage (%)↑	Unary(%)↑	Time(s)↓
German Credit (Logistic Regression) - Min Sparsity: 20.00%							
L2C (Ours)	61.35	37.31	**46.39**	**100.00**	**100.00**	**99.06**	18
DICE-Random	**88.23**	15.29	26.06	**100.00**	**100.00**	90.81	1,150
DICE-Genetic	43.45	**37.56**	40.29	62.87	90.24	56.66	17,615
COPA	57.88	18.88	28.47	44.00	44.00	84.31	17,583
MCCE	28.76	33.40	30.91	48.74	**100.00**	58.76	**2**
Adult Income (Neural Network) - Min Sparsity: 30.77%							
L2C (Ours)	45.70	**28.11**	**34.80**	**100.00**	**100.00**	**97.62**	444
DICE-Random	**89.26**	9.05	16.44	**100.00**	**100.00**	87.15	12,332
DICE-Genetic	41.48	26.27	32.14	92.64	**100.00**	72.70	505,174
MCCE	24.93	4.58	7.74	30.63	74.76	45.79	**98**
Graduate Admission (Neural Network) - Min Sparsity: 14.29%							
L2C (Ours)	42.23	37.90	39.94	**100.00**	**100.00**	**100.00**	4
DICE-Random	**66.25**	30.93	**42.15**	**100.00**	**100.00**	85.30	412
DICE-Genetic	23.05	**47.54**	31.04	92.91	**100.00**	66.69	6,171
MCCE	17.39	22.98	19.51	43.79	84.60	79.11	**1**
Student Performance (Logistic Regression) - Min Sparsity: 38.57%							
L2C (Ours)	55.32	29.54	38.51	**100.00**	**100.00**	**100.00**	6
DICE-Random	**87.60**	13.64	23.60	**100.00**	**100.00**	98.99	2,518
DICE-Genetic	39.20	**39.88**	**38.54**	84.83	**100.00**	60.77	3,406
COPA	50.45	25.28	33.68	67.26	67.26	95.32	18,774
MCCE	25.97	24.97	25.46	60.98	93.10	67.70	**1**

protection of the sensitive attributes but does greatly compromise the *validity* of the output explanations.

The purpose of k-anonymization is to ensure when attacked, an individual remains indistinguishable from at least $k-1$ others. However, notice that k-anonymization does not prevent `k-Map` against which the number of successful attacks is still high for DICE and MCCE. The threat is thus no less severe when the attacker is interested in the re-identifiability of both datasets. If the released data contains the sensitive information that is missing from the 'attack' dataset, and given the fact that 1-diversity remains well above zero for all methods, the attacker could easily infer private information of every linked record. To this end, generalizing the data as done in L2C is proved to be useful to prevent such an inference attack. It is also observed that combining CF-K with L2C in particular significantly improves the anonymity of our counterfactual data. We therefore believe that the integration of L2C with other privacy techniques in the cybersecurity area would yield a more effective safeguard.

5.3 Discretization

Discretization is an important pre-processing step in data analysis in which the problem of optimal discretization with a minimum number of splits is proved to be NP-Hard [6, 35]. We here adopt the unsupervised *Equal-frequency* discretizer, which splits a continuous attribute into buckets with the same number of instances[4]. More concretely in this experiment, features are quantized using Python function `qcut`[5], which requires specifying the maximum number of buckets/levels and later adjusts it depending on the input data distribution. We set the maximum buckets to be 4 such that every bucket averagely has 25% of total observations. Table **??** reports how

[4]Except for the features Capital gain and Capital loss from Adult Income which we convert into binary variables to accurately reflect the semantics of their data.
[5]https://pandas.pydata.org/docs/reference/api/pandas.qcut.html

Table 4: Successful attacks on counterfactual explanation methods. Bold / Underline indicates the lowest / second-lowest privacy risk for each dataset. `1-Diversity` is evaluated on 2 most sensitive attributes as shown in the columns 2 & 3.

Method	1-Anoy.↓	1-Diversity↓		1-Map↓
German Credit				
L2C (Ours)	62.15%	67.09%	**71.60%**	**0.21%**
DICE-Random	55.15%	82.75%	89.96%	23.67%
MCCE	62.83%	**65.40%**	76.95%	26.64%
DICE-Genetic	**15.36%**	90.19%	95.80%	N/A
COPA	87.61%	89.41%	89.24%	N/A
Adult Income				
L2C (Ours)	**5.07%**	**14.90%**	**39.58%**	**0.00%**
DICE-Random	72.21%	91.28%	93.07%	3.31%
MCCE	15.19%	66.42%	72.36%	1.77%
DICE-Genetic	15.11%	65.32%	87.54%	N/A

the numerical features of each dataset are discretized. We argue that having very few buckets is likely to cause under-fitting since there are very few useful combinatorial patterns that can counter the original label. Whereas we need diversity for effective learning, too many buckets are undesirable since it can hurt generalization due to some following reasons : (1) each bucket would contain too little data and the chosen middle value may not represent the bucket well, and (2) the model has more combinations of features to explore, thus can converge to sub-optimal combinations that cannot generalize well on unseen test points. In this regard, we decide to split data into equal-sized buckets in the hope of balancing the trade-off.

Various discretization methods exist [39]. Table 5 analyzes the performance of our method L2C under 3 other discretization strategies. The first one is *Minimal entropy partitioning (MDP)* [13]. It is an old-school supervised approach and one of the most widely used. MDP determines the binary discretization for a value range

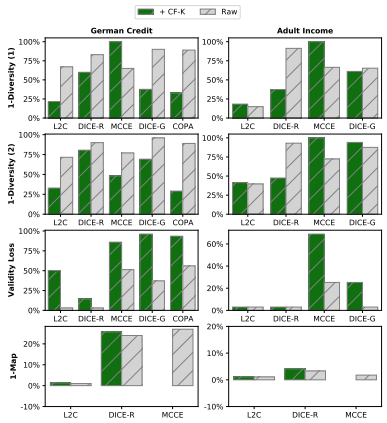


Figure 2: Privacy risk comparison between the raw output data and the data subject to 2-anonymization under the strategy of CF-K. For all metrics, lower is better. `1-Diversity` is evaluated on 2 most sensitive attributes.

by selecting the cut point that minimizes the class entropy. The algorithm can be applied recursively on sub-partitions induced by the initial cut point and the paper proposes using Minimum Description Length Principle[6] [42] as a stopping condition. Another strategy is to apply *Domain knowledge* where feature values can be grouped based on common demographic or social characteristics. For example, Age could be translated into different age groups (e.g., Teenagers, Young Adults, etc.), or TOEFL scores are divided into proficiency levels. While the aforementioned methods are univariate, we also implement a multivariate approach using Decision tree. Following the motivation of MDP, we run *CART* [4] to search for the splits that minimize class information entropy. Note that we here consider the predicted labels from the black-box models as the target variable. The goal is to each sure the prediction on each combination of features is stable as possible. To avoid fine-grained intervals, we set the minimum number of samples for a split to be 30. Our experiment shows that we are still able to achieve 100% of `Validity` and `Coverage` in roughly the same amount of time. We therefore only present the remaining metrics in Table 5, which here demonstrates a comparable quality of explanations among different discretization strategies. Given that L2C performance is relatively insensitive to the choice of discretizers, we therefore suggest using *Equal-frequency* for which no labels or external knowledge is required.

5.4 Feature Correlational Constraints

While the treatment of unary constraints is straightforward for heterogeneous datasets, we argue that this is not the case for binary constraints. For example, suggesting that a person get a Master's degree at precisely the age of 34 is unrealistically rigid. This issue indeed stems from the presence of continuous features. A direct solution is to allow for more flexible suggestions through discretization (e.g., suggesting an age range from $30 - 40$ instead of an exact value at 34). This indeed aligns with the generative mechanism

[6]https://orange3.readthedocs.io/

Table 5: Desirability of L2C counterfactual explanations under various discretization strategies. *Proposed method.

Strategy	Sparsity (%)↑	Diversity (%)↑	Harmonic Mean (%)↑	Unary (%)↑
German Credit - Min Sparsity: 20.00%				
Equal Freq.*	61.35	37.31	46.39	99.06
MDP	61.58	35.00	44.63	100.00
CART	60.92	39.85	48.18	100.00
Domain Know.	61.73	37.56	46.70	100.00
Graduate Admission - Min Sparsity: 14.29%				
Equal Freq.*	42.23	37.90	39.94	100.00
MDP	58.20	41.34	41.56	100.00
CART	41.80	41.34	41.56	100.00
Domain Know.	42.17	42.30	42.22	100.00
Student Performance - Min Sparsity: 38.57%				
Equal Freq.*	55.32	29.54	38.51	100.00
MDP	55.57	28.50	37.67	100.00
CART	55.07	27.79	36.94	100.00
Domain Know.	55.47	31.57	40.23	100.00

of our L2C, which sets us apart from existing works. However, discretization is currently treated as a subroutine of internal processing, meaning that in the output examples, the values for the continuous features are still returned in the numerical format for the sake of consistency. Therefore, the best strategy would be to have the machine learning classifiers trained in the discretized feature space accordingly. Practitioners could then ignore the one-hot decoding stage and deploying L2C for this purpose would be effortless. In Appendix B, we demonstrate how L2C effectively addresses binary constraints in this scenario with success rates of 91.54% and 100.00% respectively on German Credit and Adult Income.

6 CONCLUSION AND FUTURE WORK

In this paper, we study the challenges facing algorithmic recourse approaches in generating diverse counterfactual explanations: how *diversity* can be tackled without compromising the other desiderata of an explanation while preserving privacy against linkage attacks. We analyze how existing engines fail to resolve the trade-offs among counterfactual constraints and fill the research gap with our novel framework **L2C**. Here we target a broad class of differentiable machine learning classifiers. To fit non-differentiable models in our framework, one could use policy gradient [49] or attempt to approximate such models as decision trees or random forests with a differentiable version [27, 55]. L2C is currently proposed to deal with re-identification risks of the released examples. Defense against *model stealing* and *membership inference attacks* however remains exigent. Integrating differential privacy in the framework of L2C is one interesting research avenue, which we leave for future works to explore.

7 ACKNOWLEDGEMENT

Dinh Phung and Trung Le gratefully acknowledge the support by the US Airforce FA2386-21-1-4049 grant and the Australian Research Council ARC DP230101176 project. This does not imply endorsement by the funding agency of the research findings or conclusions. Any errors or misinterpretations in this paper are the sole responsibility of the authors.

REFERENCES

[1] Mohan S Acharya, Asfia Armaan, and Aneeta S Antony. 2019. A comparison of regression models for prediction of graduate admissions. In *2019 international*

Flat Seeking Bayesian Neural Networks

Van-Anh Nguyen[1] **Tung-Long Vuong**[1,2] **Hoang Phan**[2,3] **Thanh-Toan Do**[1]
Dinh Phung [1,2] **Trung Le** [1]
[1]Department of Data Science and AI, Monash University, Australia
[2]VinAI, Vietnam
[3]New York University, United States
{van-anh.nguyen, tung-long.vuong, toan.do, dinh.phung, trunglm}@monash.edu
hvp2011@nyu.edu

Abstract

Bayesian Neural Networks (BNNs) provide a probabilistic interpretation for deep learning models by imposing a prior distribution over model parameters and inferring a posterior distribution based on observed data. The model sampled from the posterior distribution can be used for providing ensemble predictions and quantifying prediction uncertainty. It is well-known that deep learning models with lower sharpness have better generalization ability. However, existing posterior inferences are not aware of sharpness/flatness in terms of formulation, possibly leading to high sharpness for the models sampled from them. In this paper, we develop theories, the Bayesian setting, and the variational inference approach for the sharpness-aware posterior. Specifically, the models sampled from our sharpness-aware posterior, and the optimal approximate posterior estimating this sharpness-aware posterior, have better flatness, hence possibly possessing higher generalization ability. We conduct experiments by leveraging the sharpness-aware posterior with state-of-the-art Bayesian Neural Networks, showing that the flat-seeking counterparts outperform their baselines in all metrics of interest.

1 Introduction

Bayesian Neural Networks (BNNs) provide a way to interpret deep learning models probabilistically. This is done by setting a prior distribution over model parameters and then inferring a posterior distribution over model parameters based on observed data. This allows us to not only make predictions, but also quantify prediction uncertainty, which is useful for many real-world applications. To sample deep learning models from complex and complicated posterior distributions, advanced particle-sampling approaches such as Hamiltonian Monte Carlo (HMC) [41], Stochastic Gradient HMC (SGHMC) [10], Stochastic Gradient Langevin dynamics (SGLD) [58], and Stein Variational Gradient Descent (SVGD) [36] are often used. However, these methods can be computationally expensive, particularly when many models need to be sampled for better ensembles.

To alleviate this computational burden and enable the sampling of multiple deep learning models from posterior distributions, variational inference approaches employ approximate posteriors to estimate the true posterior. These methods utilize approximate posteriors that belong to sufficiently rich families, which are both economical and convenient to sample from. However, the pioneering works in variational inference, such as [21, 5, 33], assume approximate posteriors to be fully factorized distributions, also known as mean-field variational inference. This approach fails to account for the strong statistical dependencies among random weights of neural networks, limiting its ability to capture the complex structure of the true posterior and estimate the true model uncertainty. To overcome this issue, latter works have attempted to provide posterior approximations with richer

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

expressiveness [61, 52, 53, 54, 20, 45, 55, 30, 48]. These approaches aim to improve the accuracy of the posterior approximation and enable more effective uncertainty quantification.

In the context of standard deep network training, it has been observed that flat minimizers can enhance the generalization capability of models. This is achieved by enabling them to locate wider local minima that are more robust to shifts between train and test sets. Several studies, including [27, 47, 15], have shown evidence to support this principle. However, the posteriors used in existing Bayesian neural networks (BNNs) do not account for the sharpness/flatness of the models derived from them in terms of model formulation. As a result, the sampled models can be located in regions of high sharpness and low flatness, leading to poor generalization ability. Moreover, in variational inference methods, using approximate posteriors to estimate these non-sharpness-aware posteriors can result in sampled models from the corresponding optimal approximate posterior lacking awareness of sharpness/flatness, hence causing them to suffer from poor generalization ability.

In this paper, our objective is to propose a sharpness-aware posterior for learning BNNs, which samples models with high flatness for better generalization ability. To achieve this, we devise both a Bayesian setting and a variational inference approach for the proposed posterior. By estimating the optimal approximate posteriors, we can generate flatter models that improve the generalization ability. Our approach is as follows: In Theorem 3.1, we show that the standard posterior is the optimal solution to an optimization problem that balances the empirical loss induced by models sampled from an approximate posterior for fitting a training set with a Kullback-Leibler (KL) divergence, which encourages a simple approximate posterior. Based on this insight, we replace the empirical loss induced by the approximate posterior with the general loss over the entire data-label distribution in Theorem 3.2 to improve the generalization ability. Inspired by sharpness-aware minimization [16], we develop an upper-bound of the general loss in Theorem 3.2, leading us to formulate the sharpness-aware posterior in Theorem 3.3. Finally, we devise the Bayesian setting and variational approach for the sharpness-aware posterior. Overall, our contributions in this paper can be summarized as follows:

- We propose and develop theories, the Bayesian setting, and the variational inference approach for the sharpness-aware posterior. This posterior enables us to sample a set of flat models that improve the model generalization ability. We note that SAM [16] only considers the sharpness for a single model, while ours is the first work studying the concept and theory of the sharpness for a distribution $\mathbb{Q}$ over models. Additionally, the proof of Theorem 3.2 is very challenging, elegant, and complicated because of the infinite number of models in the support of $\mathbb{Q}$.
- We conduct extensive experiments by leveraging our sharpness-aware posterior with the state-of-the-art and well-known BNNs, including *BNNs with an approximate Gaussian distribution [33]*, *BNNs with stochastic gradient Langevin dynamics (SGLD) [58]*, *MC-Dropout [18]*, *Bayesian deep ensemble [35]*, and *SWAG [39]* to demonstrate that the flat-seeking counterparts consistently outperform the corresponding approaches in all metrics of interest, including the ensemble accuracy, expected calibration error (ECE), and negative log-likelihood (NLL).

2 Related Work

2.1 Bayesian Neural Networks

Markov chain Monte Carlo (MCMC): This approach allows us to sample multiple models from the posterior distribution and was well-known for inference with neural networks through the Hamiltonian Monte Carlo (HMC) [41]. However, HMC requires the estimation of full gradients, which is computationally expensive for neural networks. To make the HMC framework practical, Stochastic Gradient HMC (SGHMC) [10] enables stochastic gradients to be used in Bayesian inference, crucial for both scalability and exploring a space of solutions. Alternatively, stochastic gradient Langevin dynamics (SGLD) [58] employs first-order Langevin dynamics in the stochastic gradient setting. Additionally, Stein Variational Gradient Descent (SVGD) [36] maintains a set of particles to gradually approach a posterior distribution. Theoretically, all SGHMC, SGLD, and SVGD asymptotically sample from the posterior in the limit of infinitely small step sizes.

Variational Inference: This approach uses an approximate posterior distribution in a family to estimate the true posterior distribution by maximizing a variational lower bound. [21] suggests fitting

a Gaussian variational posterior approximation over the weights of neural networks, which was generalized in [32, 33, 5], using the reparameterization trick for training deep latent variable models. To provide posterior approximations with richer expressiveness, many extensive studies have been proposed. Notably, [38] treats the weight matrix as a whole via a matrix variate Gaussian [22] and approximates the posterior based on this parameterization. Several later works have inspected this distribution to examine different structured representations for the variational Gaussian posterior, such as Kronecker-factored [59, 52, 53], k-tied distribution [54], non-centered or rank-1 parameterization [20, 14]. Another recipe to represent the true covariance matrix of Gaussian posterior is through the low-rank approximation [45, 55, 30, 39].

Dropout Variational Inference: This approach utilizes dropout to characterize approximate posteriors. Typically, [18] and [33] use this principle to propose Bayesian Dropout inference methods such as MC Dropout and Variational Dropout. Concrete dropout [19] extends this idea to optimize the dropout probabilities. Variational Structured Dropout [43] employs Householder transformation to learn a structured representation for multiplicative Gaussian noise in the Variational Dropout method.

2.2 Flat Minima

Flat minimizers have been found to improve the generalization ability of neural networks. This is because they enable models to find wider local minima, which makes them more robust against shifts between train and test sets [27, 47, 15, 44]. The relationship between generalization ability and the width of minima has been investigated theoretically and empirically in many studies, notably [23, 42, 12, 17]. Moreover, various methods seeking flat minima have been proposed in [46, 9, 29, 25, 16, 44]. Typically, [29, 26, 57] investigate the impacts of different training factors such as batch size, learning rate, covariance of gradient, and dropout on the flatness of found minima. Additionally, several approaches pursue wide local minima by adding regularization terms to the loss function [46, 61, 60, 9]. Examples of such regularization terms include softmax output's low entropy penalty [46] and distillation losses [61, 60].

SAM, a method that aims to minimize the worst-case loss around the current model by seeking flat regions, has recently gained attention due to its scalability and effectiveness compared to previous methods [16, 56]. SAM has been widely applied in various domains and tasks, such as meta-learning bi-level optimization [1], federated learning [51], multi-task learning [50], where it achieved tighter convergence rates and proposed generalization bounds. SAM has also demonstrated its generalization ability in vision models [11], language models [3], domain generalization [8], and multi-task learning [50]. Some researchers have attempted to improve SAM by exploiting its geometry [34, 31], additionally minimizing the surrogate gap [62], and speeding up its training time [13, 37]. Regarding the behavior of SAM, [28] empirically studied the difference in sharpness obtained by SAM [16] and SWA [24], [40] showed that SAM is an optimal Bayes relaxation of the standard Bayesian inference with a normal posterior, while [44] proved that distribution robustness [4, 49] is a probabilistic extension of SAM.

3 Proposed Framework

In what follows, we present the technicality of our proposed sharpness-aware posterior. Particularly, Section 3.1 introduces the problem setting and motivation for our sharpness-aware posterior. Section 3.2 is dedicated to our theory development, while Section 3.3 is used to describe the Bayesian setting and variational inference approach for our sharpness-aware posterior.

3.1 Problem Setting and Motivation

We aim to develop Sharpness-Aware Bayesian Neural Networks (SA-BNN). Consider a family of neural networks $f_\theta(x)$ with $\theta \in \Theta$ and a training set $\mathcal{S} = \{(x_1, y_1), ..., (x_n, y_n)\}$ where $(x_i, y_i) \sim \mathcal{D}$. We wish to learn a posterior distribution $\mathbb{Q}_S^{SA}$ with the density function $q^{SA}(\theta|\mathcal{S})$ such that any model $\theta \sim \mathbb{Q}_S^{SA}$ is aware of the sharpness when predicting over the training set $\mathcal{S}$.

We depart with the standard posterior

$$q(\theta \mid \mathcal{S}) \propto \prod_{i=1}^{n} p(y_i \mid x_i, \mathcal{S}, \theta) p(\theta),$$

where the prior distribution $\mathbb{P}$ has the density function $p(\theta)$ and the likelihood has the form

$$p\left(y \mid x, \mathcal{S}, \theta\right) \propto \exp\left\{-\frac{\lambda}{|\mathcal{S}|}\ell\left(f_\theta(x), y\right)\right\} = \exp\left\{-\frac{\lambda}{n}\ell\left(f_\theta(x), y\right)\right\}$$

with the loss function ℓ. The standard posterior $\mathbb{Q}_{\mathcal{S}}$ has the density function defined as

$$q(\theta \mid \mathcal{S}) \propto \exp\left\{-\frac{\lambda}{n}\sum_{i=1}^{n}\ell\left(f_\theta\left(x_i\right), y_i\right)\right\} p(\theta), \tag{1}$$

where $\lambda \geq 0$ is a regularization parameter.

We define the general and empirical losses as follows:

$$\mathcal{L}_{\mathcal{D}}\left(\theta\right) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell\left(f_\theta\left(x\right), y\right)\right].$$

$$\mathcal{L}_{\mathcal{S}}\left(\theta\right) = \mathbb{E}_{(x,y)\sim\mathcal{S}}\left[\ell\left(f_\theta\left(x\right), y\right)\right] = \frac{1}{n}\sum_{i=1}^{n}\ell\left(f_\theta\left(x_i\right), y_i\right).$$

Basically, the general loss is defined as the expected loss over the entire data-label distribution $\mathcal{D}$, while the empirical loss is defined as the empirical loss over a specific training set $\mathcal{S}$.

The standard posterior in Eq. (1) can be rewritten as

$$q(\theta \mid \mathcal{S}) \propto \exp\left\{-\lambda\mathcal{L}_{\mathcal{S}}\left(\theta\right)\right\} p(\theta). \tag{2}$$

Given a distribution $\mathbb{Q}$ with the density function $q\left(\theta\right)$ over the model parameters $\theta \in \Theta$, we define the empirical and general losses over this model distribution $\mathbb{Q}$ as

$$\mathcal{L}_{\mathcal{S}}\left(\mathbb{Q}\right) = \int_{\Theta}\mathcal{L}_{\mathcal{S}}\left(\theta\right)d\mathbb{Q}\left(\theta\right) = \int_{\Theta}\mathcal{L}_{\mathcal{S}}\left(\theta\right)q\left(\theta\right)d\theta.$$

$$\mathcal{L}_{\mathcal{D}}\left(\mathbb{Q}\right) = \int_{\Theta}\mathcal{L}_{\mathcal{D}}\left(\theta\right)d\mathbb{Q}\left(\theta\right) = \int_{\Theta}\mathcal{L}_{\mathcal{D}}\left(\theta\right)q\left(\theta\right)d\theta.$$

Specifically, the general loss over the model distribution $\mathbb{Q}$ is defined as the expectation of the general losses incurred by the models sampled from this distribution, while the empirical loss over the model distribution $\mathbb{Q}$ is defined as the expectation of the empirical losses incurred by the models sampled from this distribution.

3.2 Our Theory Development

We now present the theory development for the sharpness-aware posterior whose proofs can be found in the supplementary material. Inspired by the Gibbs form of the standard posterior $\mathbb{Q}_{\mathcal{S}}$ in Eq. (2), we establish the following theorem to connect the standard posterior $\mathbb{Q}_{\mathcal{S}}$ with the density $q(\theta \mid \mathcal{S})$ and the empirical loss $\mathcal{L}_{\mathcal{S}}\left(\mathbb{Q}\right)$ [7, 2].

Theorem 3.1. *Consider the following optimization problem*

$$\min_{\mathbb{Q}<<\mathbb{P}}\left\{\lambda\mathcal{L}_S\left(\mathbb{Q}\right) + KL\left(\mathbb{Q}, \mathbb{P}\right)\right\}, \tag{3}$$

where we search over $\mathbb{Q}$ absolutely continuous w.r.t. $\mathbb{P}$ and $KL\left(\cdot,\cdot\right)$ is the Kullback-Leibler divergence. This optimization has a closed-form optimal solution $\mathbb{Q}^*$ with the density

$$q^*\left(\theta\right) \propto \exp\left\{-\lambda\mathcal{L}_{\mathcal{S}}\left(\theta\right)\right\} p(\theta),$$

which is exactly the standard posterior $\mathbb{Q}_{\mathcal{S}}$ with the density $q(\theta \mid \mathcal{S})$.

Theorem 3.1 reveals that we need to find the posterior $\mathbb{Q}_{\mathcal{S}}$ balancing between optimizing its empirical loss $\mathcal{L}_S\left(\mathbb{Q}\right)$ and simplicity via $KL\left(\mathbb{Q}, \mathbb{P}\right)$. However, minimizing the empirical loss $\mathcal{L}_S\left(\mathbb{Q}\right)$ only ensures the correct predictions for the training examples in $\mathcal{S}$, hence possibly encountering overfitting. Therefore, it is desirable to replace the empirical loss by the general loss to combat overfitting.

To mitigate overfitting, in (3), we replace the empirical loss by the general loss and solve the following optimization problem (OP):

$$\min_{\mathbb{Q}<<\mathbb{P}}\left\{\lambda\mathcal{L}_{\mathcal{D}}\left(\mathbb{Q}\right) + KL\left(\mathbb{Q}, \mathbb{P}\right)\right\}. \tag{4}$$

Notably, solving the optimization problem (OP) in (4) is generally intractable. To make it tractable, we find its upper-bound which is relevant to the sharpness of a distribution $\mathbb{Q}$ over models as shown in the following theorem.

Theorem 3.2. *Assume that Θ is a compact set. Under some mild conditions, given any $\delta \in [0;1]$, with the probability at least $1-\delta$ over the choice of $\mathcal{S} \sim \mathcal{D}^n$, for any distribution $\mathbb{Q}$, we have*

$$\mathcal{L}_{\mathcal{D}}(\mathbb{Q}) \leq \mathbb{E}_{\theta\sim\mathbb{Q}}\left[\max_{\theta':\|\theta'-\theta\|\leq\rho} \mathcal{L}_{\mathcal{S}}(\theta')\right] + f\left(\max_{\theta\in\Theta}\|\theta\|^2, n\right),$$

where f is a non-decreasing function w.r.t. the first variable and approaches 0 when the training size n approaches ∞.

We note that the proof of Theorem 3.2 is not a trivial extension of sharpness-aware minimization because we need to tackle the general and empirical losses over a distribution $\mathbb{Q}$. To make explicit our sharpness over a distribution $\mathbb{Q}$ on models, we rewrite the upper-bound of the inequality as

$$\mathbb{E}_{\theta\sim\mathbb{Q}}\left[\max_{\theta':\|\theta'-\theta\|\leq\rho} \mathcal{L}_{\mathcal{S}}(\theta') - \mathcal{L}_{\mathcal{S}}(\theta)\right] + \mathcal{L}_{\mathcal{S}}(\mathbb{Q}) + f\left(\max_{\theta\in\Theta}\|\theta\|^2, n\right),$$

where the first term $\mathbb{E}_{\theta\sim\mathbb{Q}}\left[\max_{\theta':\|\theta'-\theta\|\leq\rho} \mathcal{L}_{\mathcal{S}}(\theta') - \mathcal{L}_{\mathcal{S}}(\theta)\right]$ can be regarded as *the sharpness over the distribution $\mathbb{Q}$ on the model space* and the last term $f\left(\max_{\theta\in\Theta}\|\theta\|^2, n\right)$ is a constant.

Moreover, inspired by Theorem 3.2, we propose solving the following OP which forms an upper-bound of the desirable OP in (4)

$$\min_{\mathbb{Q}<<\mathbb{P}}\left\{\lambda\mathbb{E}_{\theta\sim\mathbb{Q}}\left[\max_{\theta':\|\theta'-\theta\|\leq\rho} \mathcal{L}_{\mathcal{S}}(\theta')\right] + KL(\mathbb{Q},\mathbb{P})\right\}. \tag{5}$$

The following theorem characterizes the optimal solution of the OP in (5).

Theorem 3.3. *The optimal solution the OP in (5) is the sharpness-aware posterior distribution $\mathbb{Q}_S^{SA}$ with the density function $q^{SA}(\theta|\mathcal{S})$:*

$$q^{SA}(\theta|\mathcal{S}) \propto \exp\left\{-\lambda \max_{\theta':\|\theta'-\theta\|\leq\rho} \mathcal{L}_{\mathcal{S}}(\theta')\right\} p(\theta) = \exp\left\{-\lambda\mathcal{L}_{\mathcal{S}}(s(\theta))\right\} p(\theta),$$

where we have defined $s(\theta) = \underset{\theta':\|\theta'-\theta\|\leq\rho}{argmax}\ \mathcal{L}_{\mathcal{S}}(\theta')$.

Theorem 3.3 describes the close form of the sharpness-aware posterior distribution $\mathbb{Q}_S^{SA}$ with the density function $q^{SA}(\theta|\mathcal{S})$. Based on this characterization, in what follows, we introduce the SA Bayesian setting that sheds lights on its variational approach.

3.3 Sharpness-Aware Bayesian Setting and Its Variational Approach

Bayesian Setting: To promote the Bayesian setting for sharpness-aware posterior distribution $\mathbb{Q}_S^{SA}$, we examine the sharpness-aware likelihood

$$p^{SA}(y \mid x, \mathcal{S}, \theta) \propto \exp\left\{-\frac{\lambda}{|\mathcal{S}|}\ell\left(f_{s(\theta)}(x), y\right)\right\} = \exp\left\{-\frac{\lambda}{n}\ell\left(f_{s(\theta)}(x), y\right)\right\},$$

where $s(\theta) = \underset{\theta':\|\theta'-\theta\|\leq\rho}{\text{argmax}}\ \mathcal{L}_{\mathcal{S}}(\theta')$.

With this predefined sharpness-aware likelihood, we can recover the sharpness-aware posterior distribution $\mathbb{Q}_S^{SA}$ with the density function $q^{SA}(\theta|\mathcal{S})$:

$$q^{SA}(\theta|\mathcal{S}) \propto \prod_{i=1}^{n} p^{SA}(y_i \mid x_i, \mathcal{S}, \theta)\, p(\theta).$$

Variational inference for the sharpness-aware posterior distribution: We now develop the variational inference for the sharpness-aware posterior distribution. Let denote $X = [x_1, ..., x_n]$ and

$Y = [y_1, ..., y_n]$. Considering an approximate posterior family $\{q_\phi(\theta) : \phi \in \Phi\}$, we have

$$\begin{aligned}
&\log p^{SA}(Y \mid X, \mathcal{S}) = \int_\Theta q_\phi(\theta) \log p^{SA}(Y \mid X, \mathcal{S})\, d\theta \\
&= \int_\Theta q_\phi(\theta) \log \frac{p^{SA}(Y \mid \theta, X, \mathcal{S})\, p(\theta)}{q_\phi(\theta)} \frac{q_\phi(\theta)}{q^{SA}(\theta|\mathcal{S})} d\theta \\
&= \mathbb{E}_{q_\phi(\theta)} \left[\sum_{i=1}^n \log p^{SA}(y_i \mid x_i, \mathcal{S}, \theta) \right] - KL(q_\phi, p) + KL\left(q_\phi, q^{SA}\right).
\end{aligned}$$

It is obvious that we need to maximize the following lower bound for maximally reducing the gap $KL\left(q_\phi, q^{SA}\right)$:

$$\max_{q_\phi} \left\{ \mathbb{E}_{q_\phi(\theta)} \left[\sum_{i=1}^n \log p^{SA}(y_i \mid x_i, \mathcal{S}, \theta) \right] - KL(q_\phi, p) \right\},$$

which can be equivalently rewritten as

$$\begin{aligned}
&\min_{q_\phi} \left\{ \lambda \mathbb{E}_{q_\phi(\theta)} \left[\mathcal{L}_{\mathcal{S}}(s(\theta)) \right] + KL(q_\phi, p) \right\} \text{ or} \\
&\min_{q_\phi} \left\{ \lambda \mathbb{E}_{q_\phi(\theta)} \left[\max_{\theta' : \|\theta' - \theta\| \le \rho} \mathcal{L}_{\mathcal{S}}(\theta') \right] + KL(q_\phi, p) \right\}. \qquad (6)
\end{aligned}$$

Derivation for Variational Approach with A Gaussian Approximate Posterior: Inspired by the geometry-based SAM approaches [34, 31], we incorporate the geometry to the SA variational approach via the distance to define the ball for the sharpness as $\|\theta' - \theta\|_{\text{diag}(T_\theta)} = \sqrt{(\theta' - \theta)^T \text{diag}(T_\theta)^{-1} (\theta' - \theta)}$ as

$$\min_{q_\phi} \left\{ \lambda \mathbb{E}_{q_\phi(\theta)} \left[\max_{\theta' : \|\theta' - \theta\|_{\text{diag}(T_\theta)} \le \rho} \mathcal{L}_{\mathcal{S}}(\theta') \right] + KL(q_\phi, p) \right\}.$$

To further clarify, we consider our SA posterior distribution to Bayesian NNs, wherein we impose the Gaussian distributions to its weight matrices $W_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2 \mathbb{I}\right), i = 1, \ldots, L$[1]. The parameter ϕ consists of $\mu_i, \sigma_i, i = 1, \ldots, L$. For $\theta = W_{1:L} \sim q_\phi$, using the reparameterization trick $W_i = \mu_i + \text{diag}(\sigma_i)\epsilon_i, \epsilon_i \sim \mathcal{N}(0, \mathbb{I})$ and by searching $\theta' = W'_{1:L}$ with $W'_i = \mu'_i + \text{diag}(\sigma_i)\epsilon_i, \epsilon_i \sim \mathcal{N}(0, \mathbb{I})$, the constraint $\|\theta - \theta'\|_{\text{diag}(T_\theta)} = \|\mu - \mu'\|_{\text{diag}(T_\theta)}$ with $\mu = \mu_{1:L}$ and $\mu' = \mu'_{1:L}$. Thus, the OP in (6) reads

$$\min_{\mu, \sigma} \left\{ \lambda \mathbb{E}_\epsilon \left[\max_{\|\mu' - \mu\|_{\text{diag}(T_{\mu,\sigma})} \le \rho} \mathcal{L}_{\mathcal{S}} \left(\left[\mu'_i + \text{diag}(\sigma_i)\epsilon_i \right]_{i=1}^L \right) \right] \right\}, \qquad (7)$$

where $\sigma = \sigma_{1:L}$, $\epsilon = \epsilon_{1:L}$, and we define $\text{diag}(T_\theta) = \text{diag}(T_{\mu,\sigma})$ in the distance of the geometry.

To solve the OP in (7), we sample $\epsilon = \epsilon_{1:L}$ from the standard Gaussian distributions, employ an one-step gradient ascent to find μ', and use the gradient at μ' to update μ. Specifically, we find μ' [6] (Chapter 9) as

$$\mu' = \mu + \rho \frac{\text{diag}(T_{\mu,\sigma}) \nabla_\mu \mathcal{L}_{\mathcal{S}} \left(\left[\mu_i + \text{diag}(\sigma_i)\epsilon_i \right]_{i=1}^L \right)}{\left\| \text{diag}(T_{\mu,\sigma}) \nabla_\mu \mathcal{L}_{\mathcal{S}} \left(\left[\mu_i + \text{diag}(\sigma_i)\epsilon_i \right]_{i=1}^L \right) \right\|}.$$

The diagnose of $\text{diag}(T_{\mu,\sigma})$ specifies the importance level of the model weights, i.e., the weight with a higher importance level is encouraged to have a higher sharpness via a smaller absolute partial derivative of the loss w.r.t. this weight. We consider $\text{diag}(T_{\mu,\sigma}) = \mathbb{I}$ (i.e., *the standard SA BNN*) and $\text{diag}(T_{\mu,\sigma}) = \text{diag}\left(\frac{|\mu|}{\sigma}\right)$ (i.e., *the geometry SA BNN*). Here we note that $\frac{\cdot}{\cdot}$ represents the element-wise division.

[1]We absorb the biases to the weight matrices.

Table 1: Classification score on CIFAR-100 dataset.Each experiment is repeated three times with different random seeds and reports the mean and standard deviation.

Method	PreResNet-164 ACC ↑	PreResNet-164 NLL ↓	PreResNet-164 ECE ↓	WideResNet28x10 ACC ↑	WideResNet28x10 NLL ↓	WideResNet28x10 ECE ↓
Variational inference						
MC-Dropout	79.50 ± 0.37	0.9162 ± 0.0103	0.0993 ± 0.0033	82.30 ± 0.19	0.6500 ± 0.0049	0.0574 ± 0.0028
F-MC-Dropout	**81.06 ± 0.44**	**0.7027 ± 0.0049**	**0.0514 ± 0.0047**	**83.24 ± 0.11**	**0.6144 ± 0.0068**	**0.0250 ± 0.0027**
Deep-ens	82.08 ± 0.42	0.7189 ± 0.0108	0.0334 ± 0.0064	83.04 ± 0.15	0.6958 ± 0.0335	0.0483 ± 0.0017
F-Deep-ens	**82.54 ± 0.10**	**0.6286 ± 0.0022**	**0.0143 ± 0.0041**	**84.52 ± 0.03**	**0.5644 ± 0.0106**	**0.0191 ± 0.0039**
Markov chain Monte Carlo						
SGLD	80.13 ± 0.01	0.7604 ± 0.0010	0.1161 ± 0.0031	81.38 ± 0.10	0.7123 ± 0.0204	0.0958 ± 0.0004
F-SGLD	**80.82 ± 0.02**	**0.7276 ± 0.0012**	**0.1085 ± 0.0008**	**82.12 ± 0.16**	**0.6722 ± 0.0112**	**0.0820 ± 0.0021**
Sample						
SWAG-Diag	80.18 ± 0.50	0.6837 ± 0.0186	**0.0239 ± 0.0047**	82.40 ± 0.09	0.6150 ± 0.0029	0.0322 ± 0.0018
F-SWAG-Diag	**81.01 ± 0.29**	**0.6645 ± 0.0050**	0.0242 ± 0.0039	**83.50 ± 0.29**	**0.5763 ± 0.0120**	**0.0151 ± 0.0020**
SWAG	79.90 ± 0.50	**0.6595 ± 0.0019**	0.0587 ± 0.0048	82.23 ± 0.19	0.6078 ± 0.0006	**0.0113 ± 0.0020**
F-SWAG	**80.93 ± 0.27**	0.6704 ± 0.0049	**0.0350 ± 0.0025**	**83.57 ± 0.26**	**0.5757 ± 0.0136**	0.0196 ± 0.0015

Table 2: Classification score on CIFAR-10 dataset.Each experiment is repeated three times with different random seeds and reports the mean and standard deviation.

Method	PreResNet-164 ACC ↑	PreResNet-164 NLL ↓	PreResNet-164 ECE ↓	WideResNet28x10 ACC ↑	WideResNet28x10 NLL ↓	WideResNet28x10 ECE ↓
Variational inference						
MC-Dropout	96.18 ± 0.02	0.1270 ± 0.0030	0.0162 ± 0.0007	96.39 ± 0.09	0.1094 ± 0.0021	**0.0094 ± 0.0014**
F-MC-Dropout	**96.39 ± 0.18**	**0.1137 ± 0.0024**	**0.0118 ± 0.0006**	**97.10 ± 0.12**	**0.0966 ± 0.0047**	0.0095 ± 0.0008
Deep-ens	96.39 ± 0.09	0.1277 ± 0.0030	0.0108 ± 0.0015	96.96 ± 0.10	0.1031 ± 0.0076	0.0087 ± 0.0018
F-Deep-ens	**96.70 ± 0.04**	**0.1031 ± 0.0016**	**0.0057 ± 0.0031**	**97.11 ± 0.10**	**0.0851 ± 0.0011**	**0.0059 ± 0.0012**
Markov chain Monte Carlo						
SGLD	94.79 ± 0.10	0.2089 ± 0.0021	0.0711 ± 0.0061	95.87 ± 0.08	0.1573 ± 0.0190	0.0463 ± 0.0050
F-SGLD	**95.04 ± 0.06**	**0.1912 ± 0.0080**	**0.0601 ± 0.0002**	**96.43 ± 0.05**	**0.1336 ± 0.004**	**0.0385 ± 0.0003**
Sample						
SWAG-Diag	96.03 ± 0.10	0.1251 ± 0.0029	0.0082 ± 0.0008	96.41 ± 0.05	0.1077 ± 0.0009	0.0047 ± 0.0013
F-SWAG-Diag	**96.23 ± 0.01**	**0.1108 ± 0.0013**	**0.0043 ± 0.0005**	**97.05 ± 0.08**	**0.0888 ± 0.0052**	**0.0043 ± 0.0004**
SWAG	96.03 ± 0.02	0.1232 ± 0.0022	**0.0053 ± 0.0004**	96.32 ± 0.08	0.1122 ± 0.0009	0.0088 ± 0.0006
F-SWAG	**96.25 ± 0.03**	**0.11062 ± 0.0014**	0.0056 ± 0.0002	**97.09 ± 0.14**	**0.0883 ± 0.0004**	**0.0036 ± 0.0008**

Finally, the objective function in (6) indicates that we aim to find an approximate posterior distribution that ensures any model sampled from it is aware of the sharpness, while also preferring simpler approximate posterior distributions. This preference can be estimated based on how we equip these distributions. With the Bayesian setting and variational inference formulation, our proposed sharpness-aware posterior can be integrated into MCMC-based and variational inference-based Bayesian Neural Networks. The supplementary material contains the details on how to derive variational approaches and incorporate the sharpness-awareness into the BNNs used in our experiments including BNNs with an approximate Gaussian distribution [33], BNNs with stochastic gradient Langevin dynamics (SGLD) [58], MC-Dropout [18], Bayesian deep ensemble [35], and SWAG [39].

4 Experiments

In this section, we conduct various experiments to demonstrate the effectiveness of the sharpness-aware approach on Bayesian Neural networks, including BNNs with an approximate Gaussian distribution [33] (i.e., SGVB for model's reparameterization trick and SGVB-LRT for representation's reparameterization trick), BNNs with stochastic gradient Langevin dynamics (SGLD) [58], MC-Dropout [18], Bayesian deep ensemble [35], and SWAG [39]. The experiments are conducted on three benchmark datasets: CIFAR-10, CIFAR-100, and ImageNet ILSVRC-2012, and report accuracy, negative log-likelihood (NLL), and Expected Calibration Error (ECE) to estimate the calibration capability and uncertainty of our method against baselines. The details of the dataset and implementation are described in the supplementary material[2].

[2]The implementation is provided in https://github.com/anh-ntv/flat_bnn.git

Table 3: Classification scores of approximate the Gaussian posterior on the CIFAR datasets. Each experiment is repeated three times with different random seeds and reports the mean and standard deviation.

Method	Resnet10 ACC ↑	Resnet10 NLL ↓	Resnet10 ECE ↓	Resnet18 ACC ↑	Resnet18 NLL ↓	Resnet18 ECE ↓
Experiments on Cifar-100 dataset						
SGVB-LRT	61.75 ± 0.75	1.534 ± 0.03	0.0676 ± 0.01	68.95 ± 1.20	1.140 ± 0.21	0.063 ± 0.04
F-SGVB-LRT	62.25 ± 0.57	1.4001 ± 0.04	0.0642 ± 0.01	70.00 ± 1.42	1.127 ± 0.25	**0.022 ± 0.05**
+ Geometry	**62.54 ± 0.67**	**1.3704 ± 0.01**	**0.0301 ± 0.03**	**70.12 ± 1.02**	**1.121 ± 0.23**	0.036 ± 0.06
SGVB	54.40 ± 0.98	1.968 ± 0.05	0.214 ± 0.00	60.91 ± 2.31	1.746 ± 0.15	0.246 ± 0.03
F-SGVB	54.53 ± 0.33	1.967 ± 0.00	0.212 ± 0.00	61.54 ± 2.23	1.695 ± 0.15	0.242 ± 0.03
+ Geometry	**55.53 ± 0.65**	**1.906 ± 0.02**	**0.207 ± 0.00**	**62.58 ± 0.53**	**1.612 ± 0.03**	**0.224 ± 0.00**
Experiments on Cifar-10 dataset						
SGVB-LRT	84.98 ± 1.87	0.422 ± 0.10	0.043 ± 0.04	89.10 ± 1.32	0.344 ± 0.02	0.033 ± 0.02
F-SGVB-LRT	86.32 ± 1.34	0.409 ± 0.03	**0.017 ± 0.06**	90.00 ± 1.10	0.291 ± 0.02	0.019 ± 0.01
+ Geometry	**86.44 ± 1.12**	**0.403 ± 0.06**	0.025 ± 0.03	**90.31 ± 1.11**	**0.262 ± 0.01**	**0.014 ± 0.02**
SGVB	80.52 ± 2.10	0.781 ± 0.23	0.237 ± 0.06	86.74 ± 1.25	0.541 ± 0.01	0.181 ± 0.02
F-SGVB	80.60 ± 1.88	0.776 ± 0.13	0.223 ± 0.05	**87.01 ± 0.91**	0.534 ± 0.01	0.183 ± 0.01
+ Geometry	**82.05 ± 0.47**	**0.704 ± 0.01**	**0.206 ± 0.00**	86.80 ± 1.30	**0.531 ± 0.01**	**0.175 ± 0.01**

Table 4: Classification score on ImageNet dataset

Model	Densenet-161 ACC ↑	Densenet-161 NLL ↓	Densenet-161 ECE ↓	ResNet-152 ACC ↑	ResNet-152 NLL ↓	ResNet-152 ECE ↓
SWAG-Diag	78.59	0.8559	0.0459	78.96	0.8584	0.0566
F-SWAG-Diag	**78.71**	**0.8267**	**0.0194**	**79.20**	**0.8065**	**0.0199**
SWAG	78.59	0.8303	0.0204	79.08	0.8205	0.0279
F-SWAG	**78.70**	**0.8262**	**0.0185**	**79.17**	**0.8078**	**0.0208**
SGLD	78.50	0.8317	**0.0157**	79.00	0.8165	0.0220
F-SGLD	**78.64**	**0.8236**	0.0166	**79.16**	**0.8050**	**0.0167**

4.1 Experimental results

4.1.1 Predictive performance

Our experimental results, presented in Tables 1, 2, 3 for CIFAR-100 and CIFAR-10 dataset, and Table 4 for the ImageNet dataset, indicate a notable improvement across all experiments. It is worth noting that there is a trade-off between accuracy, negative log-likelihood, and expected calibration error. Nonetheless, our approach obtains a fine balance between these factors compared to the overall improvement.

4.2 Effectiveness of sharpness-aware posterior

Calibration of uncertainty estimates: We evaluate the ECE of each setting and compare it to baselines in Tables 1, 2, and 4. This score measures the maximum discrepancy between the accuracy

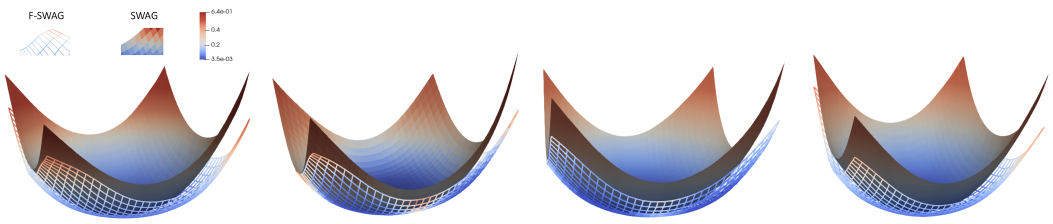


Figure 1: Comparing loss landscape of PreResNet-164 on CIFAR-100 dataset training with SWAG and F-SWAG method. For visualization purposes, we sample two models for each SWAG and F-SWAG and then plot the loss landscapes. It can be observed that the loss landscapes of our F-SWAG are flatter, supporting our argument for the flatter sampled models.

Figure 2: Reliability diagrams for PreResNet164 on CIFAR-100. The confidence is split into 20 bins and plots the gap between confidence and accuracy in each bin. The best case is the black dashed line when this gap is zeros. The plots of F-SWAG get closer to the zero lines, implying our F-SWAG can calibrate the uncertainty better.

Table 5: Classification score on CIFAR-10-C on PreResNet-164 model when training with CIFAR-10. The full result on each type of corruption is displayed in the supplemetary material.

	ECE ↓				**Accuracy** ↑			
Corruption	SWAG-D	F-SWAG-D	SWAG	F-SWAG	SWAG-D	F-SWAG-D	SWAG	F-SWAG
Noise	0.0729	0.0701	0.0958	0.0078	74.26	75.59	74.02	75.08
Blur	0.0121	0.0090	0.0202	0.0273	91.13	90.55	91.03	90.93
Weather	0.018	0.0142	0.0272	0.0240	89.47	89.18	89.42	89.11
Digital and others	0.0277	0.0229	0.0384	0.0209	87.03	86.94	86.93	87.19
Average	0.0328	**0.0290**	0.0454	**0.0200**	85.47	**85.56**	85.35	**85.58**

and confidence of the model. To further clarify it, we display the Reliability Diagrams of PreResNet-164 on CIFAR-100 to understand how well the model predicts according to the confidence threshold in Figure 2. The experiments is detailed in the supplementary material.

Out-of-distribution prediction: The effectiveness of the sharpness-aware Bayesian neural network (BNN) is demonstrated in the above experiments, particularly in comparison to non-flat methods. In this section, we extend the evaluation to an out-of-distribution setting. Specifically, we utilize the BNN models trained on the CIFAR-10 dataset to assess their performance on the CIFAR-10-C dataset. This is an extension of the CIFAR-10 designed to evaluate the robustness of machine learning models against common corruptions and perturbations in the input data. The corruptions include various forms of noise, blur, weather conditions, and digital distortions. We conduct an ensemble of 30 models sampled from the flat-posterior distribution and compared them with non-flat ones. We present the average result of each corruption group and the average result on the whole dataset in Table 5, the detailed result of each corruption form is displayed in the supplementary material. Remarkably, the flat BNN models consistently surpass their non-flat counterparts with respect to average ECE and accuracy metrics. This finding is additional evidence of the generalization ability of the sharpness-aware posterior.

4.3 Ablation studies

In Figure 1, we plot the loss-landscape of the models sampled from our proposal of sharpness-aware posterior against the non-sharpness-aware one. Particularly, we compare two methods F-SWAG and SWAG by selecting four random models sampled from the posterior distribution of each method under the same hyper-parameter settings. As observed, our method not only improves the generalization of ensemble inference, demonstrated by classification results in Section 4.1 and sharpness in Section 4.2, but also the individual sampled model is flatter itself.

We measure and visualize the sharpness of the models. To this end, we sample five models from the approximate posteriors and then take the average of the sharpness of these models. For a model θ, the sharpness is evaluated as $\max_{||\epsilon||_2 \leq \rho} \mathcal{L}_{\mathcal{S}}(\theta + \epsilon) - \mathcal{L}_{\mathcal{S}}(\theta)$ to measure the change of loss value around θ. We calculate the sharpness score of PreResNet-164 network for SWAG, and F-SWAG training on CIFAR-100 dataset and visualize them in the supplementary material. As shown there, the sharpness-aware versions produce smaller *sharpness* scores compared to the corresponding baselines, indicating that our models get into flatter regions.