Adaptively Weighted Data Augmentation Consistency Regularization for Robust Optimization under Concept Shift

Yijun Dong [*1] **Yuege Xie** [*2] **Rachel Ward** [1 3]

Abstract

Concept shift is a prevailing problem in natural tasks like medical image segmentation where samples usually come from different subpopulations with variant correlations between features and labels. One common type of concept shift in medical image segmentation is the "information imbalance" between *label-sparse* samples with few (if any) segmentation labels and *label-dense* samples with plentiful labeled pixels. Existing distributionally robust algorithms have focused on adaptively truncating/down-weighting the "less informative" (*i.e.*, label-sparse in our context) samples. To exploit data features of label-sparse samples more efficiently, we propose an adaptively weighted online optimization algorithm — *AdaWAC*— to incorporate data augmentation consistency regularization in sample reweighting. Our method introduces a set of trainable weights to balance the supervised loss and unsupervised consistency regularization of each sample separately. At the saddle point of the underlying objective, the weights assign label-dense samples to the supervised loss and label-sparse samples to the unsupervised consistency regularization. We provide a convergence guarantee by recasting the optimization as online mirror descent on a saddle point problem. Our empirical results demonstrate that *AdaWAC* not only enhances the segmentation performance and sample efficiency but also improves the robustness to concept shift on various medical image segmentation tasks with different UNet-style backbones.

[*]Equal contribution [1]Oden Institute, University of Texas at Austin, TX, USA. [2]Snap Inc., CA, USA. (Work done at University of Texas at Austin). [3]Department of Mathematics, University of Texas at Austin, TX, USA. Correspondence to: Yijun Dong <ydong@utexas.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

1 Introduction

Modern machine learning is revolutionizing the field of medical imaging, especially in computer-aided diagnosis with computed tomography (CT) and magnetic resonance imaging (MRI) scans. However, classical learning objectives like empirical risk minimization (ERM) generally assume that training samples are independently and identically (*i.i.d.*) distributed, whereas real-world medical image data rarely satisfy this assumption. Figure 1 instantiates a common observation in medical image segmentation where the segmentation labels corresponding to different cross-sections of the human body tend to have distinct proportions of labeled (*i.e.*, non-background) pixels, which is accurately reflected by the evaluation of supervised cross-entropy loss during training. We refer to this as the "information imbalance" among samples, as opposed to the well-studied "class imbalance" (Wong et al., 2018; Taghanaki et al., 2019; Yeung et al., 2022) among the numbers of segmentation labels in different classes. Such information imbalance induces distinct difficulty/paces of learning with the cross-entropy loss for different samples (Wang et al., 2021b; Tullis & Benjamin, 2011; Tang et al., 2018; Hacohen & Weinshall, 2019). Specifically, we say a sample is *label-sparse* when it contains very few (if any) segmentation labels; in contrast, a sample is *label-dense* when its segmentation labels are prolific. Motivated by the information imbalance among samples, we explore the following questions:

What is the effect of separation between sparse and dense labels on segmentation?
Can we leverage such information imbalance to improve the segmentation accuracy?

We formulate the mixture of label-sparse and label-dense samples as a concept shift — a type of distribution shift in the conditional distribution of labels given features $P(\mathbf{y}|\mathbf{x})$. Coping with concept shifts, prior works have focused on adaptively truncating (hard-thresholding) the empirical loss associated with label-sparse samples. These include the Trimmed Loss Estimator (Shen & Sanghavi, 2019), MKL-SGD (Shah et al., 2020), Ordered SGD (Kawaguchi & Lu, 2020), and the quantile-based Kacmarz algorithm (Haddock et al., 2022). Alternatively, another line of works (Wang

et al., 2018; Sagawa et al., 2020) proposes to relax the hard-thresholding operation to soft-thresholding by down-weighting instead of truncating the less informative samples. However, diminishing sample weights reduces the importance of both the features and the labels simultaneously, which is still not ideal as the potentially valuable information in the features of the label-sparse samples may not be fully used.

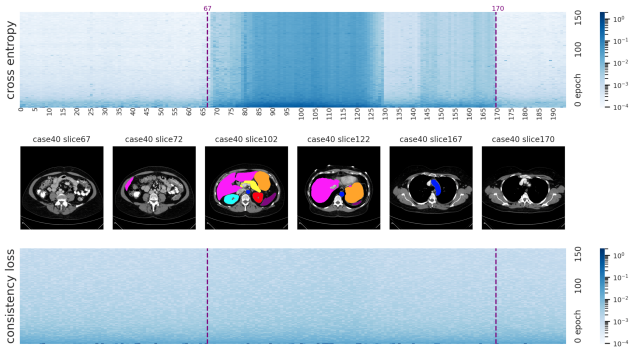

Figure 1: Evolution of cross-entropy losses versus consistency regularization terms for slices at different cross-sections of the human body in the Synapse dataset (described in Section 5) during training.

For further exploitation of the feature of training samples, we propose the incorporation of *data augmentation consistency regularization* on label-sparse samples. As a prevalent strategy for utilizing unlabeled data, consistency regularization (Bachman et al., 2014; Laine & Aila, 2016; Sohn et al., 2020) encourages data augmentations of the same samples to lie in the vicinity of each other on a proper manifold. For medical imaging segmentation, consistency regularization has been extensively studied in the semi-supervised learning setting (Bortsova et al., 2019; Zhao et al., 2019; Li et al., 2020; Wang et al., 2021a; Zhang et al., 2021; Zhou et al., 2021; Basak et al., 2022) as a strategy for overcoming label scarcity. Nevertheless, unlike general vision tasks, for medical image segmentation, the scantiness of unlabeled image data can also be a problem due to regulations and privacy considerations (Karimi et al., 2020), which makes it worthwhile to reminisce the more classical supervised learning setting. In contrast to the aforementioned semi-supervised strategies, we explore the potency of consistency regularization in the *supervised learning* setting by leveraging the information in the features of label-sparse samples via data augmentation consistency regularization.

To naturally distinguish the label-sparse and label-dense samples, we make a key observation that the unsupervised consistency regularization on encoder layer outputs (of a UNet-style architecture) is much more uniform across different subpopulations than the supervised cross-entropy loss (as exemplified in Figure 1). Since the consistency regularization is characterized by the marginal distribution of features $P(\mathbf{x})$ but not labels, and therefore is less affected by the concept shift in $P(\mathbf{y}|\mathbf{x})$, it serves as a natural reference for separating the label-sparse and label-dense samples. In light of this observation, we present the *weighted data augmentation consistency (WAC) regularization* — a minimax formulation that reweights the cross-entropy loss versus the consistency regularization associated with each sample via a set of trainable weights. At the saddle point of this minimax formulation, the WAC regularization automatically separates samples from different subpopulations by assigning all weights to the consistency regularization for label-sparse samples, and all weights to the cross-entropy terms for label-dense samples.

We further introduce an adaptively weighted online optimization algorithm — *AdaWAC*— for solving the minimax problem posed by the WAC regularization, which is inspired by a mirror-descent-based algorithm for distributionally robust optimization (Sagawa et al., 2020). By adaptively learning the weights between the cross-entropy loss and consistency regularization of different samples, *AdaWAC* comes with both a convergence guarantee and empirical success.

The main contributions are summarized as follows:

- We introduce the *WAC regularization* that leverages the consistency regularization on the encoder layer outputs (of a UNet-style architecture) as a natural reference to distinguish the label-sparse and label-dense samples (Section 3).
- We propose an adaptively weighted online optimization algorithm — *AdaWAC*— for solving the WAC regularization problem with a convergence guarantee (Section 4).
- Through extensive experiments on different medical image segmentation tasks with different UNet-style backbone architectures, we demonstrate the effectiveness of *AdaWAC* not only for enhancing the segmentation performance and sample efficiency but also for improving the robustness to concept shift (Section 5).

1.1 Related Work

Sample reweighting. Sample reweighting is a popular strategy for dealing with distribution/subpopulation shifts in training data where different weights are assigned to samples from different subpopulations. In particular, the distributionally-robust optimization (DRO) framework (Ben-Tal et al., 2013; Duchi et al., 2021; Duchi & Namkoong, 2021; Sagawa et al., 2020) considers a collection of training sample groups from different distributions. With the explicit grouping of samples, the goal is to minimize the worst-case loss over the groups. Without prior knowledge of sample grouping, importance sampling (Needell et al., 2014; Zhao & Zhang, 2015; Alain et al., 2015; Loshchilov & Hutter, 2015; Gopal, 2016; Katharopoulos & Fleuret,

2018), iterative trimming (Kawaguchi & Lu, 2020; Shen & Sanghavi, 2019), and empirical-loss-based reweighting (Wu et al., 2022) are commonly incorporated in the stochastic optimization process for adaptive reweighting and separation of samples from different subpopulations.

Data augmentation consistency regularization. As a popular way of exploiting data augmentations, consistency regularization encourages models to learn the vicinity among augmentations of the same sample based on the assumption that data augmentations generally preserve the semantic information in data and therefore lie closely on proper manifolds. Beyond being a powerful building block in semi-supervised (Bachman et al., 2014; Sajjadi et al., 2016; Laine & Aila, 2016; Sohn et al., 2020; Berthelot et al., 2019) and self-supervised (Wu et al., 2018; He et al., 2020; Chen et al., 2020; Grill et al., 2020) learning, the incorporation of data augmentation and consistency regularization also provably improves generalization and feature learning even in the supervised learning setting (Yang et al., 2023; Shen et al., 2022).

For medical imaging, data augmentation consistency regularization is generally leveraged as a semi-supervised learning tool (Bortsova et al., 2019; Zhao et al., 2019; Li et al., 2020; Wang et al., 2021a; Zhang et al., 2021; Zhou et al., 2021; Basak et al., 2022). In efforts to incorporate consistency regularization in segmentation tasks with augmentation-sensitive labels, (Li et al., 2020) encourages transformation consistency between predictions with augmentations applied to the image inputs and the segmentation outputs. (Basak et al., 2022) penalizes inconsistent segmentation outputs between teacher-student models, with MixUp (Zhang et al., 2017) applied to image inputs of the teacher model and segmentation outputs of the student model. Instead of enforcing consistency in the segmentation output space as above, we leverage the insensitivity of sparse labels to augmentations and encourage consistent encodings (in the latent space of encoder outputs) on label-sparse samples.

2 Problem Setup

Notation. For any $K \in \mathbb{N}$, we denote $[K] = \{1, \ldots, K\}$. We represent the elements and subtensors of an arbitrary tensor by adapting the syntax for Python slicing on the subscript (except counting from 1). For example, $\mathbf{x}_{[i,j]}$ denotes the (i,j)-entry of the two-dimensional tensor $\mathbf{x}$, and $\mathbf{x}_{[i,:]}$ denotes the i-th row. Let $\mathbb{I}$ be a function onto $\{0,1\}$ such that, for any event e, $\mathbb{I}\{e\} = 1$ if e is true and 0 otherwise. For any $k \in \mathbb{N}$, let $\Delta_k \triangleq \left\{\mathbf{q} \in [0,1]^k \;\middle|\; \|\mathbf{q}\|_1 = 1\right\}$ be the k-dimensional probability simplex. For any distribution P and $n \in \mathbb{N}$, we let P^n denote the joint distribution of n samples drawn *i.i.d.* from P. Finally, we say that an event happens with high probability (*w.h.p.*) if the event takes place with probability $1 - \Omega\left(\text{poly}\left(n\right)\right)^{-1}$.

2.1 Pixel-wise Classification with Sparse and Dense Labels

We consider medical image segmentation as a pixel-wise multi-class classification problem where we aim to learn a pixel-wise classifier $h : \mathcal{X} \to [K]^d$ that serves as a good approximation to the ground truth $h^* : \mathcal{X} \to [K]^d$.

Recall the separation of cross-entropy losses between samples with different proportions of background pixels from Figure 1. We refer to a sample $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times [K]^d$ as *label-sparse* if most pixels in $\mathbf{y}$ are labeled as background; for these samples, the cross-entropy loss on $(\mathbf{x}, \mathbf{y})$ converges rapidly in the early stage of training. Otherwise, we say that $(\mathbf{x}, \mathbf{y})$ is *label-dense*. Formally, we describe such variation as a concept shift in the data distribution.

Definition 2.1 (Mixture of label-sparse and label-dense subpopulations)**.** We assume that *label-sparse and label-dense samples* are drawn from P_0 and P_1 with distinct conditional distributions $P_0\left(\mathbf{y}|\mathbf{x}\right)$ and $P_1\left(\mathbf{y}|\mathbf{x}\right)$ but common marginal distribution $P\left(\mathbf{x}\right)$ such that $P_i\left(\mathbf{x}, \mathbf{y}\right) = P_i\left(\mathbf{y}|\mathbf{x}\right) P\left(\mathbf{x}\right)$ $(i = 0, 1)$. For $\xi \in [0,1]$, we define P_ξ as a data distribution where $(\mathbf{x}, \mathbf{y}) \sim P_\xi$ is drawn either from P_1 with probability ξ or from P_0 with probability $1 - \xi$.

We aim to learn a pixel-wise classifier from a function class $\mathcal{H} \ni h_\theta = \text{argmax}_{k \in [K]} f_\theta\left(\mathbf{x}\right)_{[j,:]}$ for all $j \in [d]$ where the underlying function $f_\theta \in \mathcal{F}$, parameterized by some $\theta \in \mathcal{F}_\theta$, admits an encoder-decoder structure:

$$\left\{f_\theta = \phi_\theta \circ \psi_\theta \;\middle|\; \phi_\theta : \mathcal{X} \to \mathcal{Z}, \psi_\theta : \mathcal{Z} \to [0,1]^{d \times K}\right\}. \tag{1}$$

Here ϕ_θ, ψ_θ correspond to the encoder and decoder functions, respectively. The parameter space $\mathcal{F}_\theta$ is equipped with the norm $\|\cdot\|_{\mathcal{F}}$ and its dual norm $\|\cdot\|_{\mathcal{F},*}$[1]. $(\mathcal{Z}, \varrho)$ is a latent metric space.

To learn from segmentation labels, we consider the *averaged cross-entropy loss*:

$$\begin{aligned}\ell_{CE}\left(\theta; (\mathbf{x}, \mathbf{y})\right) &= -\frac{1}{d}\sum_{j=1}^{d}\sum_{k=1}^{K}\mathbb{I}\left\{\mathbf{y}_{[j]} = k\right\} \cdot \log\left(f_\theta\left(\mathbf{x}\right)_{[j,k]}\right) \\ &= -\frac{1}{d}\sum_{j=1}^{d}\log\left(f_\theta\left(\mathbf{x}\right)_{[j,\mathbf{y}_{[j]}]}\right).\end{aligned} \tag{2}$$

We assume the proper learning setting: there exists $\theta^* \in \bigcap_{\xi \in [0,1]} \text{argmin}_{\theta \in \mathcal{F}_\theta} \mathbb{E}_{(\mathbf{x},\mathbf{y}) \sim P_\xi}\left[\ell_{CE}\left(\theta; (\mathbf{x}, \mathbf{y})\right)\right]$, which is invariant with respect to ξ.[2]

[1]For *AdaWAC* (Proposition 4.1 in Section 4), $\mathcal{F}_\theta$ is simply a subspace in the Euclidean space with dimension equal to the total number of parameters for each $\theta \in \mathcal{F}_\theta$, with $\|\cdot\|_{\mathcal{F}}$ and $\|\cdot\|_{\mathcal{F},*}$ both being the ℓ_2-norm.

[2]We assume proper learning only to (i) highlight the invariance

2.2 Augmentation Consistency Regularization

Despite the invariance of f_{θ^*} to P_ξ on the population loss, with a finite number of training samples in practice, the predominance of label-sparse samples would be problematic. As an extreme scenario for the pixel-wise classifier with encoder-decoder structure (Equation (1)), when the label-sparse samples are predominant ($\xi \ll 1$), a decoder function ψ_θ that predicts every pixel as background can achieve near-optimal cross-entropy loss, regardless of the encoder function ϕ_θ, considerably compromising the test performance (*cf.* Table 1). To encourage legit encoding even in the absence of sufficient dense labels, we leverage the unsupervised consistency regularization on the *encoder function* ϕ_θ based on data augmentations.

Let $\mathcal{A}$ be a distribution over transformations on $\mathcal{X}$ where for any $\mathbf{x} \in \mathcal{X}$, each $A \sim \mathcal{A}$ ($A : \mathcal{X} \to \mathcal{X}$) induces an augmentation $A(\mathbf{x})$ of $\mathbf{x}$ that perturbs low-level information in $\mathbf{x}$. We aim to learn an encoder function $\phi_\theta : \mathcal{X} \to \mathcal{Z}$ that is capable of filtering out low-level information from $\mathbf{x}$ and therefore provides similar encodings for augmentations of the same sample. Recalling the metric ϱ (*e.g.*, the Euclidean distance) on $\mathcal{Z}$, for a given scaling hyperparameter $\lambda_{AC} > 0$, we measure the similarity between augmentations with a consistency regularization term on $\phi_\theta(\cdot)$: for any $A_1, A_2 \sim \mathcal{A}^2$,

$$\ell_{AC}(\theta; \mathbf{x}, A_1, A_2) \triangleq \lambda_{AC} \cdot \varrho\Big(\phi_\theta(A_1(\mathbf{x})), \phi_\theta(A_2(\mathbf{x}))\Big). \tag{3}$$

For the n training samples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i\in[n]} \sim P_\xi^n$, we consider n pairs of data augmentation transformations $\{(A_{i,1}, A_{i,2})\}_{i\in[n]} \sim \mathcal{A}^{2n}$. In the basic version, we encourage the similar encoding $\phi_\theta(\cdot)$ of the augmentation pairs $(A_{i,1}(\mathbf{x}_i), A_{i,2}(\mathbf{x}_i))$ for all $i \in [n]$ via consistency regularization:

$$\min_{\theta\in\mathcal{F}_{\theta^*}(\gamma)} \frac{1}{n}\sum_{i=1}^{n} \ell_{CE}(\theta; (\mathbf{x}_i, \mathbf{y}_i)) + \ell_{AC}(\theta; \mathbf{x}_i, A_{i,1}, A_{i,2}). \tag{4}$$

We enforce consistency on $\phi_\theta(\cdot)$ in light of the encoder-decoder architecture: the encoder is generally designed to abstract essential information and filters out low-level non-semantic perturbations (*e.g.*, those introduced by augmentations), while the decoder recovers the low-level information for the pixel-wise classification. Therefore, with different $A_1, A_2 \sim \mathcal{A}$, the encoder output $\phi_\theta(\cdot)$ tends to be more consistent than the other intermediate layers, especially for label-dense samples.

3 Weighted Augmentation Consistency (WAC) Regularization

As the motivation, we begin with a key observation about the averaged cross-entropy:

Remark 3.1 (Separation of averaged cross-entropy loss on P_0 and P_1). As demonstrated in Figure 1, the sparse labels from P_0 tend to be much easier to learn than the dense ones from P_1, leading to considerable separation of averaged cross-entropy losses on the sparse and dense labels after a sufficient number of training epochs. In other words, $\ell_{CE}(\theta; (\mathbf{x}, \mathbf{y})) \ll \ell_{CE}(\theta; (\mathbf{x}', \mathbf{y}'))$ for label-sparse samples $(\mathbf{x}, \mathbf{y}) \sim P_0$ and label-dense samples $(\mathbf{x}', \mathbf{y}') \sim P_1$ with high probability.

Although Equation (4) with consistency regularization alone can boost the segmentation accuracy during testing (*cf.* Table 4), it does not take the separation between label-sparse and label-dense samples into account. In Section 5, we will empirically demonstrate that proper exploitation of such separation, like the formulation introduced below, can lead to improved classification performance.

We formalize the notion of separation between P_0 and P_1 with the consistency regularization (Equation (3)) as a reference in the following assumption [3].

Assumption 3.2 (n-separation between P_0 and P_1)**.** Given a sufficiently small $\gamma > 0$, let $\mathcal{F}_{\theta^*}(\gamma) = \{\theta \in \mathcal{F}_\theta \mid \|\theta - \theta^*\|_{\mathcal{F}} \le \gamma\}$ be a compact and convex neighborhood of well-trained pixel-wise classifiers[4]. We say that P_0 *and* P_1 *are* n*-separated over* $\mathcal{F}_{\theta^*}(\gamma)$ if there exists $\omega > 0$ such that with probability $1 - \Omega\left(n^{1+\omega}\right)^{-1}$ over $((\mathbf{x}, \mathbf{y}), (A_1, A_2)) \sim P_\xi \times \mathcal{A}^2$, the following hold:

(i) $\ell_{CE}(\theta; (\mathbf{x}, \mathbf{y})) < \ell_{AC}(\theta; \mathbf{x}, A_1, A_2)$ for all $\theta \in \mathcal{F}_{\theta^*}(\gamma)$ given $(\mathbf{x}, \mathbf{y}) \sim P_0$;

(ii) $\ell_{CE}(\theta; (\mathbf{x}, \mathbf{y})) > \ell_{AC}(\theta; \mathbf{x}, A_1, A_2)$ for all $\theta \in \mathcal{F}_{\theta^*}(\gamma)$ given $(\mathbf{x}, \mathbf{y}) \sim P_1$.

This assumption is motivated by the empirical observation that the perturbation in $\phi_\theta(\cdot)$ induced by $\mathcal{A}$ is more uniform across P_0 and P_1 than the averaged cross-entropy losses, as instantiated in Figure 3.

of the desired ground truth to ξ that can be challenging to learn with finite samples in practice and (ii) provide a natural pivot for the convex and compact neighborhood $\mathcal{F}_{\theta^*}(\gamma)$ of ground truth θ^* in Assumption 3.2 granted by the pretrained initialization, where θ^* can also be replaced with the pretrained initialization weights $\theta_0 \in \mathcal{F}_{\theta^*}(\gamma)$. In particular, neither our theory nor the *AdaWAC* algorithm requires the function class $\mathcal{F}$ to be expressive enough to truly contain such θ^*.

[3] Although Assumption 3.2 may seem to be rather strong, it is only required for the separation guarantee of label-sparse and label-dense samples with high probability in Proposition 3.3, but not for the adaptive weighting algorithm introduced in Section 4 or in practice for the experiments.

[4] With pretrained initialization, we assume that the optimization algorithm is always probing in $\mathcal{F}_{\theta^*}(\gamma)$.

Under Assumption 3.2, up to a proper scaling hyperparameter λ_{AC}, the consistency regularization (Equation (3)) can separate the averaged cross-entropy loss (Equation (2)) on n label-sparse and label-dense samples with probability $1-\Omega\left(n^{\omega}\right)^{-1}$ (as explained formally in Appendix A). In particular, the larger n corresponds to the stronger separation between P_0 and P_1.

With Assumption 3.2, we introduce a minimax formulation that incentivizes the separation of label-sparse and label-dense samples automatically by introducing a flexible weight $\boldsymbol{\beta}_{[i]} \in [0,1]$ that balances $\ell_{CE}\left(\theta;\left(\mathbf{x}_i,\mathbf{y}_i\right)\right)$ and $\ell_{AC}\left(\theta;\mathbf{x}_i,A_{i,1},A_{i,2}\right)$ for each of the n samples.

$$\begin{aligned}\widehat{\theta}^{WAC},\widehat{\boldsymbol{\beta}} &\in \operatorname*{argmin}_{\theta\in\mathcal{F}_{\theta^*}(\gamma)} \operatorname*{argmax}_{\boldsymbol{\beta}\in[0,1]^n} \widehat{L}^{WAC}\left(\theta,\boldsymbol{\beta}\right)\\ \widehat{L}^{WAC}\left(\theta,\boldsymbol{\beta}\right) &\triangleq \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\beta}_{[i]}\cdot\ell_{CE}\left(\theta;\left(\mathbf{x}_i,\mathbf{y}_i\right)\right)\\ &\quad+\left(1-\boldsymbol{\beta}_{[i]}\right)\cdot\ell_{AC}\left(\theta;\mathbf{x}_i,A_{i,1},A_{i,2}\right).\end{aligned} \tag{5}$$

With convex and continuous loss and regularization terms (formally in Proposition 3.3), Equation (5) admits a saddle point corresponding to $\widehat{\boldsymbol{\beta}}$ which separates the label-sparse and label-dense samples under Assumption 3.2.

Proposition 3.3 (Formal proof in Appendix A)**.** *Assume that $\ell_{CE}\left(\theta;\left(\mathbf{x},\mathbf{y}\right)\right)$ and $\ell_{AC}\left(\theta;\mathbf{x},A_1,A_2\right)$ are convex and continuous in θ for all $\left(\mathbf{x},\mathbf{y}\right)\in\mathcal{X}\times[K]^d$ and $A_1,A_2\sim\mathcal{A}^2$; $\mathcal{F}_{\theta^*}\left(\gamma\right)\subset\mathcal{F}_\theta$ is compact and convex. If P_0 and P_1 are n-separated (Assumption 3.2), then there exists $\widehat{\boldsymbol{\beta}}\in\{0,1\}^n$ and $\widehat{\theta}^{WAC}\in\operatorname{argmin}_{\theta\in\mathcal{F}_{\theta^*}(\gamma)}\widehat{L}^{WAC}\left(\theta,\widehat{\boldsymbol{\beta}}\right)$ such that*

$$\begin{aligned}\min_{\theta\in\mathcal{F}_{\theta^*}(\gamma)}\widehat{L}^{WAC}\left(\theta,\widehat{\boldsymbol{\beta}}\right) &= \widehat{L}^{WAC}\left(\widehat{\theta}^{WAC},\widehat{\boldsymbol{\beta}}\right)\\ &= \max_{\boldsymbol{\beta}\in[0,1]^n}\widehat{L}^{WAC}\left(\widehat{\theta}^{WAC},\boldsymbol{\beta}\right).\end{aligned} \tag{6}$$

*Further, $\widehat{\boldsymbol{\beta}}$ separates the label-sparse and label-dense samples—$\widehat{\boldsymbol{\beta}}_{[i]}=\mathbb{I}\left\{\left(\mathbf{x}_i,\mathbf{y}_i\right)\sim P_1\right\}$—*w.h.p..

That is, for n samples drawn from a mixture of n-separated P_0 and P_1, the saddle point of $L_i^{WAC}\left(\theta,\boldsymbol{\beta}\right)$ in Equation (5) corresponds to $\boldsymbol{\beta}_{[i]}=0$ on label-sparse samples (*i.e.*, learning from the unsupervised consistency regularization), and $\boldsymbol{\beta}_{[i]}=1$ on label-dense samples (*i.e.*, learning from the supervised averaged cross-entropy loss).

Remark 3.4 (Connection to hard-thresholding algorithms). The saddle point of Equation (5) is closely related to hard-thresholding algorithms like Ordered SGD (Kawaguchi & Lu, 2020) and iterative trimmed loss (Shen & Sanghavi, 2019). In each iteration, these algorithms update the model only on a proper subset of training samples based on the (ranking of) current empirical risks. Compared to hard-thresholding algorithms, (i) Equation (5) additionally leverages the unused samples (*e.g.*, label-sparse samples) for unsupervised consistency regularization on data augmentations; (ii) meanwhile, it does not require prior knowledge of the sample subpopulations (*e.g.*, ξ for P_ξ) which is essential for hard-thresholding algorithms.

Equation (5) further facilitates the more flexible optimization process. As we will empirically show in Table 2, despite the close relation between Equation (5) and the hard-thresholding algorithms (Remark 3.4), such updating strategies may be suboptimal for solving Equation (5).

4 Adaptively Weighted Augmentation Consistency (*AdaWAC*)

Inspired by the breakthrough made by (Sagawa et al., 2020) in the distributionally-robust optimization (DRO) setting where gradient updating on weights is shown to enjoy better convergence guarantees than hard thresholding, we introduce an adaptively weighted online optimization algorithm (Algorithm 1) for solving Equation (5) based on online mirror descent.

In contrast to the commonly used stochastic gradient descent (SGD), the flexibility of online mirror descent in choosing the associated norm space not only allows gradient updates on sample weights but also grants distinct learning dynamics to sample weights $\boldsymbol{\beta}_t$ and model parameters θ_t, which leads to the following convergence guarantee.

Proposition 4.1 (Formally in Proposition B.1, proof in Appendix B, assumptions instantiated in Example 1)**.** *Assume that $\ell_{CE}\left(\theta;\left(\mathbf{x},\mathbf{y}\right)\right)$ and $\ell_{AC}\left(\theta;\mathbf{x},A_1,A_2\right)$ are convex and continuous in θ for all $\left(\mathbf{x},\mathbf{y}\right)\in\mathcal{X}\times[K]^d$ and $A_1,A_2\sim\mathcal{A}^2$. Assume moreover that $\mathcal{F}_{\theta^*}\left(\gamma\right)\subset\mathcal{F}_\theta$ is convex and compact. If there exist* [5] *$C_{\theta,*}>0$ and $C_{\boldsymbol{\beta},*}>0$ such that*

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla_\theta\widehat{L}_i^{WAC}\left(\theta,\boldsymbol{\beta}\right)\right\|_{\mathcal{F},*}^2\le C_{\theta,*}^2$$

$$\frac{1}{n}\sum_{i=1}^{n}\max\left\{\ell_{CE}\left(\theta;\left(\mathbf{x}_i,\mathbf{y}_i\right)\right),\ell_{AC}\left(\theta;\mathbf{x}_i,A_{i,1},A_{i,2}\right)\right\}^2\le C_{\boldsymbol{\beta},*}^2$$

for all $\theta\in\mathcal{F}_{\theta^*}\left(\gamma\right)$, $\boldsymbol{\beta}\in[0,1]^n$, then with $\eta_\theta=\eta_{\boldsymbol{\beta}}=\frac{2}{\sqrt{5T\left(\gamma^2C_{\theta,*}^2+2nC_{\boldsymbol{\beta},*}^2\right)}}$, Algorithm 1 provides

$$\begin{aligned}&\mathbb{E}\left[\max_{\boldsymbol{\beta}\in[0,1]^n}\widehat{L}^{WAC}\left(\overline{\theta}_T,\boldsymbol{\beta}\right)-\min_{\theta\in\mathcal{F}_{\theta^*}(\gamma)}\widehat{L}^{WAC}\left(\theta,\overline{\boldsymbol{\beta}}_T\right)\right]\\ &\qquad\le 2\sqrt{5\left(\gamma^2C_{\theta,*}^2+2nC_{\boldsymbol{\beta},*}^2\right)/T}\end{aligned}$$

where $\overline{\theta}_T=\frac{1}{T}\sum_{t=1}^{T}\theta_t$ and $\overline{\boldsymbol{\beta}}_T=\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\beta}_t$.

[5] Following the convention, we use $*$ in subscription to denote the dual spaces. For instance, recalling the parameter space $\mathcal{F}_\theta$ characterized by the norm $\|\cdot\|_{\mathcal{F}}$ from Section 2.1, we use $\|\cdot\|_{\mathcal{F},*}$ to denote its dual norm; while $C_{\theta,*},C_{\boldsymbol{\beta},*}$ upper bound the dual norms of the gradients with respect to θ and $\boldsymbol{\beta}$.

Algorithm 1 Adaptively Weighted Augmentation Consistency (*AdaWAC*)

Input: Training samples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i\in[n]} \sim P_\xi^n$, augmentations $\{(A_{i,1}, A_{i,2})\}_{i\in[n]} \sim \mathcal{A}^{2n}$, maximum number of iterations $T \in \mathbb{N}$, learning rates $\eta_\theta, \eta_{\boldsymbol{\beta}} > 0$, pretrained initialization for the pixel-wise classifier $\theta_0 \in \mathcal{F}_{\theta^*}(\gamma)$.
Initialize the sample weights $\boldsymbol{\beta}_0 = \mathbf{1}/2 \in [0,1]^n$.
for $t = 1, \ldots, T$ **do**
  Sample $i_t \sim [n]$ uniformly
  $\mathbf{b} \leftarrow \left[(\boldsymbol{\beta}_{t-1})_{[i_t]}, 1 - (\boldsymbol{\beta}_{t-1})_{[i_t]}\right]$
  $\mathbf{b}_{[1]} \leftarrow \mathbf{b}_{[1]} \cdot \exp\left(\eta_{\boldsymbol{\beta}} \cdot \ell_{CE}\left(\theta_{t-1}; (\mathbf{x}_{i_t}, \mathbf{y}_{i_t})\right)\right)$
  $\mathbf{b}_{[2]} \leftarrow \mathbf{b}_{[2]} \cdot \exp\left(\eta_{\boldsymbol{\beta}} \cdot \ell_{AC}\left(\theta_{t-1}; \mathbf{x}_{i_t}, A_{i_t,1}, A_{i_t,2}\right)\right)$
  $\boldsymbol{\beta}_t \leftarrow \boldsymbol{\beta}_{t-1}, (\boldsymbol{\beta}_t)_{[i_t]} \leftarrow \mathbf{b}_{[1]} / \|\mathbf{b}\|_1$
  $\theta_t \leftarrow \theta_{t-1} - \eta_\theta \cdot \Big((\boldsymbol{\beta}_t)_{[i_t]} \cdot \nabla_\theta \ell_{CE}\left(\theta_{t-1}; (\mathbf{x}_{i_t}, \mathbf{y}_{i_t})\right)$
  $+ \left(1 - (\boldsymbol{\beta}_t)_{[i_t]}\right) \cdot \nabla_\theta \ell_{AC}\left(\theta_{t-1}; \mathbf{x}_{i_t}, A_{i_t,1}, A_{i_t,2}\right) \Big)$
end for

In addition to the convergence guarantee, Algorithm 1 also demonstrates superior performance over hard-thresholding algorithms for segmentation problems in practice (Table 2). An intuitive explanation is that instead of filtering out all the label-sparse samples via hard thresholding, the adaptive weighting allows the model to learn from some sparse labels at the early epochs, while smoothly down-weighting ℓ_{CE} of these samples since learning sparse labels tends to be easier (Remark 3.1). With the learned model tested on a mixture of label-sparse and label-dense samples, learning sparse labels at the early stage is crucial for accurate segmentation.

5 Experiments

In this section, we investigate the proposed *AdaWAC* algorithm (Algorithm 1) on different medical image segmentation tasks with different UNet-style architectures. We first demonstrate the performance improvements brought by *AdaWAC* in terms of sample efficiency and robustness to concept shift (Table 1). Then, we verify the empirical advantage of *AdaWAC* compared to the closely related hard-thresholding algorithms as discussed in Remark 3.4 (Table 2). Our ablation study (Table 4) further illustrates the indispensability of both sample reweighting and consistency regularization, the deliberate combination of which leads to the superior performance of *AdaWAC*[6].

Experiment setup. We conduct experiments on two medical image segmentation tasks: abdominal CT segmentation for Synapse multi-organ dataset (Synapse)[7] and cine-MRI segmentation for Automated cardiac diagnosis challenge dataset (ACDC)[8], with two UNet-like architectures: TransUNet (Chen et al., 2021) and UNet (Ronneberger et al., 2015) (deferred to Appendix E.2). For the main experiments with TransUNet in Section 5, we follow the official implementation in (Chen et al., 2021) and use ERM+SGD as the baseline. We evaluate segmentations with two standard metrics—the average Dice-similarity coefficient (DSC) and the average 95-percentile of Hausdorff distance (HD95). Dataset and implementation details are deferred to Appendix D. Given the sensitivity of medical image semantics to perturbations, our experiments only involve simple augmentations (*i.e.*, rotation and mirroring) adapted from (Chen et al., 2021).

It is worth highlighting that, in addition to the information imbalance among samples caused by the concept shift discussed in this work, the pixel-wise class imbalance (*e.g.*, the predominance of background pixels) is another well-investigated challenge for medical image segmentation, where coupling the dice loss (Wong et al., 2018; Taghanaki et al., 2019; Yeung et al., 2022) in the objective is a common remedy used in many state-of-the-art methods (Chen et al., 2021; Cao et al., 2023). The implementation of *AdaWAC* also leverages the dice loss to alleviate pixel-wise class imbalance. We defer the detailed discussion to Appendix C.

5.1 Segmentation Performance of *AdaWAC* with TransUNet

Segmentation on Synapse. Figure 2 visualizes the segmentation predictions on 6 Synapse test slices given by models trained via *AdaWAC* (ours) and via the baseline (ERM+SGD) with TransUNet (Chen et al., 2021). We observe that *AdaWAC* provides more accurate predictions on the segmentation boundaries and captures small organs better than the baseline.

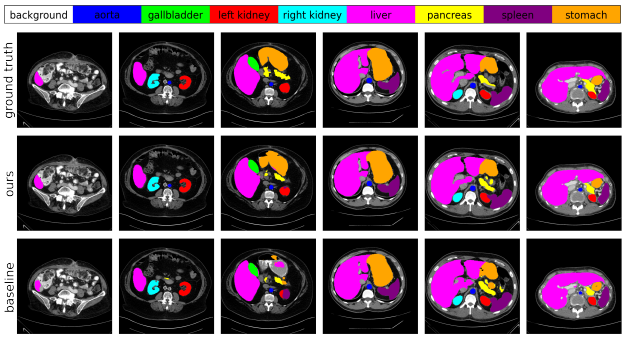

Figure 2: Visualization of segmentation predictions with TransUNet (Chen et al., 2021) on Synapse. Top to bottom: ground truth, ours (*AdaWAC*), baseline.

[6]We release our code at https://github.com/gail-yxie/adawac.
[7]https://www.synapse.org/#!Synapse:syn3193805/wiki/217789
[8]https://www.creatis.insa-lyon.fr/Challenge/acdc/

Visualization of *AdaWAC*. As shown in Figure 3, with $\ell_{CE}(\theta_t; (\mathbf{x}_i, \mathbf{y}_i))$ (Equation (2)) of label-sparse versus label-dense slices weakly separated in the early epochs, the model further learns to distinguish $\ell_{CE}(\theta_t; (\mathbf{x}_i, \mathbf{y}_i))$ of label-sparse/label-dense slices during training. By contrast, $\ell_{AC}(\theta_t; \mathbf{x}_i, A_{i,1}, A_{i,2})$ (Equation (3)) remains mixed for all slices throughout the entire training process. As a result, the CE weights of label-sparse slices are much smaller than those of label-dense ones, pushing *AdaWAC* to learn more image representations but fewer pixel classifications for slices with sparse labels and learn more pixel classifications for slices with dense labels.

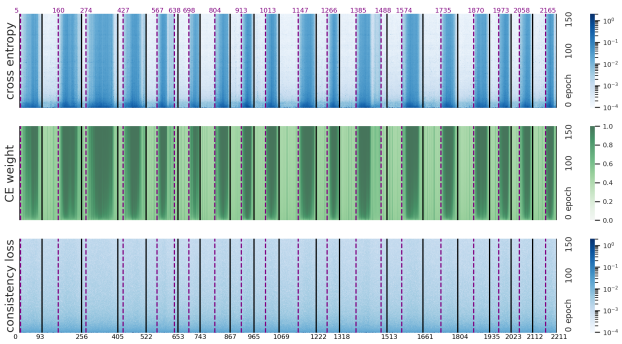

Figure 3: $\ell_{CE}(\theta_t; (\mathbf{x}_i, \mathbf{y}_i))$ (top), CE weights $\boldsymbol{\beta}_t$ (middle), and $\ell_{AC}(\theta_t; \mathbf{x}_i, A_{i,1}, A_{i,2})$ (bottom) of the entire Synapse training process. The x-axis indexes slices 0–2211. The y-axis enumerates epochs 0–150. Individual cases (patients) are partitioned by black lines, while purple lines separate slices with/without non-background pixels.

Sample efficiency and robustness. We first demonstrate the *sample efficiency* of *AdaWAC* in comparison to the baseline (ERM+SGD) when training only on different subsets of the full Synapse training set ("**full**" in Table 1). Specifically, (i) **half-slice** contains slices with even indices only in each case (patient)[9]; (ii) **half-vol** consists of 9 cases uniformly sampled from the total 18 cases in **full** where different cases tend to have distinct ξs (*i.e.*, ratios of label–dense samples); (iii) **half-sparse** takes the first half slices in each case, most of which tend to be label-sparse (*i.e.*, ξs are made to be small). As shown in Table 1, the model trained with *AdaWAC* on **half-slice** generalizes as well as a baseline model trained on **full**, if not better. Moreover, the **half-vol** and **half-sparse** experiments illustrate the *robustness* of *AdaWAC* to concept shift. Furthermore, such sample efficiency and distributional robustness of *AdaWAC* extend to the more widely used UNet architecture. We defer the detailed results and discussions on UNet to Appendix E.2.

[9]Such sampling is equivalent to doubling the time interval between two consecutive scans or halving the scanning frequency in practice, resulting in the halving of sample size.

Comparison with hard-thresholding algorithms. Table 2 illustrates the empirical advantage of *AdaWAC* over the hard-thresholding algorithms, as suggested in Remark 3.4. In particular, we consider the following hard-thresholding algorithms: (i) **trim-train** learns only from slices with at least one non-background pixel and trims the rest in each iteration on the fly; (ii) **trim-ratio** ranks the cross-entropy loss $\ell_{CE}(\theta_t; (\mathbf{x}_i, \mathbf{y}_i))$ in each iteration (mini-batch) and trims samples with the lowest cross-entropy losses at a fixed ratio – the ratio of all-background slices in the full training set ($1 - \frac{1280}{2211} \approx 0.42$); (iii) **ACR** further incorporates the data augmentation consistency regularization directly via the addition of $\ell_{AC}(\theta_t; \mathbf{x}_i, A_{i,1}, A_{i,2})$ without reweighting; (iv) **pseudo-*AdaWAC*** simulates the sample weights $\boldsymbol{\beta}$ at the saddle point and learns via $\ell_{CE}(\theta_t; (\mathbf{x}_i, \mathbf{y}_i))$ on slices with at least one non-background pixel while via $\ell_{AC}(\theta_t; \mathbf{x}_i, A_{i,1}, A_{i,2})$ otherwise. We see that naive incorporation of **ACR** brings less observable boosts to the hard-thresholding methods. Therefore, the deliberate combination via reweighting in *AdaWAC* is essential for performance improvement.

Segmentation on ACDC. Performance improvements granted by *AdaWAC* are also observed on the ACDC dataset (Table 3). We defer detailed visualization of ACDC segmentation to Appendix E.

5.2 Ablation Study

On the influence of consistency regularization. To illustrate the role of consistency regularization in *AdaWAC*, we consider the **reweight-only** scenario with $\lambda_{AC} = 0$ and $\boldsymbol{\beta} \in \Delta_n$ (*cf.* $\boldsymbol{\beta} \in [0, 1]^n$ in Equation (5)) such that Equation (5) is reduced to a standard DRO formulation. As an alternative pure sample reweighting strategy, we also examine the loss percentile minimization algorithm (**reweight-EM**) proposed in (Fidon et al., 2021), which can be interpreted as **reweight-only** with an additional entropy maximization regularization term (refer to Appendix E.1). We observe that, with zero consistency regularization in *AdaWAC*, reweighting alone brings little improvement (Table 4).

On the influence of sample reweighting. We then investigate the effect of sample reweighting under different reweighting learning rates $\eta_{\boldsymbol{\beta}}$ (recall Algorithm 1): (i) **ACR-only** for $\eta_{\boldsymbol{\beta}} = 0$ (equivalent to the naive addition of $\ell_{AC}(\theta_t; \mathbf{x}_i, A_{i,1}, A_{i,2})$), (ii) ***AdaWAC*-0.01** for $\eta_{\boldsymbol{\beta}} = 0.01$, and (iii) ***AdaWAC*-1.0** for $\eta_{\boldsymbol{\beta}} = 1.0$. As Table 4 implies, when removing reweighting from *AdaWAC*, augmentation consistency regularization alone improves DSC slightly from 76.28 (baseline) to 77.89 (ACR-only), whereas *AdaWAC* boosts DSC to 79.12 (*AdaWAC*-1.0) with a proper choice of $\eta_{\boldsymbol{\beta}}$.

Table 1: *AdaWAC* with TransUNet trained on the full Synapse and its subsets.

Training	Method	DSC ↑	HD95 ↓	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
full	baseline	76.66 ± 0.88	29.23 ± 1.90	87.06	55.90	81.95	75.58	94.29	56.30	86.05	76.17
	AdaWAC	$\mathbf{79.04} \pm \mathbf{0.21}$	$\mathbf{27.39} \pm \mathbf{1.91}$	87.53	56.57	83.23	81.12	94.04	62.05	89.51	78.32
half-slice	baseline	74.62 ± 0.78	31.62 ± 8.37	86.14	44.23	79.09	78.46	93.50	55.78	84.54	75.24
	AdaWAC	$\mathbf{77.37} \pm \mathbf{0.40}$	$\mathbf{29.56} \pm \mathbf{1.09}$	86.89	55.96	82.15	78.63	94.34	57.36	86.60	77.05
half-vol	baseline	71.08 ± 0.90	46.83 ± 2.91	84.38	46.71	78.19	74.55	92.02	48.03	76.28	68.47
	AdaWAC	$\mathbf{73.81} \pm \mathbf{0.94}$	$\mathbf{35.33} \pm \mathbf{0.92}$	84.37	48.14	80.32	77.39	93.23	52.78	83.50	70.79
half-sparse	baseline	31.74 ± 2.78	69.72 ± 1.37	65.71	8.33	59.46	51.59	51.18	10.72	6.92	0.00
	AdaWAC	$\mathbf{41.03} \pm \mathbf{2.12}$	$\mathbf{59.04} \pm \mathbf{12.32}$	71.27	8.33	69.14	63.09	64.29	17.74	30.77	3.57

Table 2: *AdaWAC* versus hard-thresholding algorithms with TransUNet on Synapse.

Method	baseline	trim-train		trim-ratio		pseudo-*AdaWAC*	*AdaWAC*
			+ACR		+ACR		
DSC ↑	76.66 ± 0.88	76.80 ± 1.13	78.42 ± 0.17	76.49 ± 0.16	77.71 ± 0.56	77.72 ± 0.65	$\mathbf{79.04} \pm \mathbf{0.21}$
HD95 ↓	29.23 ± 1.90	32.05 ± 2.34	27.84 ± 1.16	31.96 ± 2.60	28.51 ± 2.66	28.45 ± 1.18	$\mathbf{27.39} \pm \mathbf{1.91}$

Table 3: *AdaWAC* with TransUNet trained on ACDC.

Method	DSC ↑	HD95 ↓	RV	Myo	LV
TransUNet	89.40 ± 0.22	2.55 ± 0.37	89.17	83.24	95.78
AdaWAC (ours)	$\mathbf{90.67} \pm \mathbf{0.27}$	$\mathbf{1.45} \pm \mathbf{0.55}$	90.00	85.94	96.06

Table 4: Ablation study of *AdaWAC* with TransUNet trained on Synapse.

Method	DSC ↑	HD95 ↓	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
baseline	76.66 ± 0.88	29.23 ± 1.90	87.06	55.90	81.95	75.58	94.29	56.30	86.05	76.17
reweight-only	76.27 ± 0.42	32.66 ± 3.48	87.30	52.56	81.21	75.77	94.13	58.96	84.69	75.52
reweight-EM	76.83 ± 0.62	31.95 ± 2.64	87.33	54.16	82.20	76.00	93.84	58.59	86.35	76.16
ACR-only	78.01 ± 0.62	27.78 ± 2.80	87.51	58.79	83.39	79.26	94.70	58.99	86.02	75.43
AdaWAC-0.01	77.75 ± 0.23	28.02 ± 3.50	87.33	56.68	83.35	78.53	94.45	57.02	87.72	76.94
AdaWAC-1.0	$\mathbf{79.04} \pm \mathbf{0.21}$	$\mathbf{27.39} \pm \mathbf{1.91}$	87.53	56.57	83.23	81.12	94.04	62.05	89.51	78.32

6 Discussion

In this paper, we explore the information imbalance commonly observed in medical image segmentation and exploit the information in features of label-sparse samples via *AdaWAC*, an adaptively weighted online optimization algorithm. *AdaWAC* can be viewed as a careful combination of adaptive sample reweighting and data augmentation consistency regularization. By casting the information imbalance among samples as a concept shift in the data distribution, we leverage the unsupervised data augmentation consistency regularization on the encoder layer outputs (of UNet-style architectures) as a natural reference for distinguishing the label-sparse and label-dense samples via the comparisons against the supervised average cross-entropy loss. We formulate such comparisons as a weighted augmentation consistency (WAC) regularization problem and propose *AdaWAC* for iterative and smooth separation of samples from different subpopulations with a convergence guarantee. Our experiments on various medical image segmentation tasks with different UNet-style architectures empirically demonstrate the effectiveness of *AdaWAC* not only in improving the segmentation performance and sample efficiency but also in enhancing the distributional robustness to concept shifts.

Limitations and future directions. From an algorithmic perspective, a limitation of this work is the utilization of the encoder layer outputs $\phi_\theta(\cdot)$ for data augmentation consistency regularization, which resulted in *AdaWAC* being tailored to encoder-decoder architectures at the current stage. However, our method can be generalized to other architectures in principle by selecting a representation extractor in the network that (i) well characterizes the marginal distribution of features $P(\mathbf{x})$ (ii) while being robust to the concept shift in $P(\mathbf{y}|\mathbf{x})$. For other (non-segmentation) applications where encoder-decoder architectures do not apply, further investigation into such generalizations is a promising avenue for future research.

Meanwhile, noticing the prevalence of concept shifts in natural data, especially for dense prediction tasks like segmentation and detection, we hope to extend the application/idea of *AdaWAC* beyond medical image segmentation as a potential future direction.

Acknowledgement

R. Ward was partially supported by AFOSR MURI FA9550-19-1-0005, NSF DMS 1952735, NSF HDR1934932, and NSF 2019844. Y. Dong was supported by NSF DMS 1952735. Y. Xie was supported by NSF 2019844. The authors wish to thank Qi Lei and Xiaoxia Wu for valuable discussions and Jieneng Chen for generously providing preprocessed medical image segmentation datasets.

References

Alain, G., Lamb, A., Sankar, C., Courville, A., and Bengio, Y. Variance reduction in sgd by distributed importance sampling. *arXiv preprint arXiv:1511.06481*, 2015.

Asgari Taghanaki, S., Abhishek, K., Cohen, J. P., Cohen-Adad, J., and Hamarneh, G. Deep semantic segmentation of natural and medical images: a review. *Artificial Intelligence Review*, 54:137–178, 2021.

Bachman, P., Alsharif, O., and Precup, D. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27:3365–3373, 2014.

Basak, H., Bhattacharya, R., Hussain, R., and Chatterjee, A. An embarrassingly simple consistency regularization method for semi-supervised medical image segmentation. *arXiv preprint arXiv:2202.00677*, 2022.

Ben-Tal, A., den Hertog, D., Waegenaere, A. D., Melenberg, B., and Rennen, G. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013. ISSN 00251909, 15265501.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. A. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

Bertsekas, D. *Convex optimization theory*, volume 1. Athena Scientific, 2009.

Bortsova, G., Dubost, F., Hogeweg, L., Katramados, I., and Bruijne, M. d. Semi-supervised medical image segmentation via learning consistency under transformations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 810–818. Springer, 2019.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., and Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pp. 205–218. Springer, 2023.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., and Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Duchi, J. C. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

Duchi, J. C., Glynn, P. W., and Namkoong, H. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3): 946–969, 2021.

Fidon, L., Aertsen, M., Mufti, N., Deprest, T., Emam, D., Guffens, F., Schwartz, E., Ebner, M., Prayer, D., Kasprian, G., et al. Distributionally robust segmentation of abnormal fetal brain 3d mri. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3*, pp. 263–273. Springer, 2021.

Gopal, S. Adaptive sampling for sgd by exploiting side information. In *International Conference on Machine Learning*, pp. 364–372. PMLR, 2016.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Hacohen, G. and Weinshall, D. On the power of curriculum learning in training deep networks. In *International Conference on Machine Learning*, pp. 2535–2544. PMLR, 2019.

Haddock, J., Needell, D., Rebrova, E., and Swartworth, W. Quantile-based iterative methods for corrupted systems of linear equations. *SIAM Journal on Matrix Analysis and Applications*, 43(2):605–637, 2022.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Iakubovskii, P. Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch, 2019.

Karimi, D., Dou, H., Warfield, S. K., and Gholipour, A. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759, 2020.

Katharopoulos, A. and Fleuret, F. Not all samples are created equal: Deep learning with importance sampling. In *International conference on machine learning*, pp. 2525–2534. PMLR, 2018.

Kawaguchi, K. and Lu, H. Ordered sgd: A new stochastic optimization framework for empirical risk minimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 669–679. PMLR, 2020.

Laine, S. and Aila, T. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

Li, X., Yu, L., Chen, H., Fu, C.-W., Xing, L., and Heng, P.-A. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):523–534, 2020.

Loshchilov, I. and Hutter, F. Online batch selection for faster training of neural networks. *arXiv preprint arXiv:1511.06343*, 2015.

Milletari, F., Navab, N., and Ahmadi, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, 2016.

Needell, D., Ward, R., and Srebro, N. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Advances in neural information processing systems*, 27, 2014.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Russell, S. J. and Norvig, P. Artificial intelligence: a modern approach. malaysia, 2016.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

Sajjadi, M., Javanmardi, M., and Tasdizen, T. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29:1163–1171, 2016.

Shah, V., Wu, X., and Sanghavi, S. Choosing the sample with lowest loss makes sgd robust. In *International Conference on Artificial Intelligence and Statistics*, pp. 2120–2130. PMLR, 2020.

Shen, R., Bubeck, S., and Gunasekar, S. Data augmentation as feature manipulation. In *International Conference on Machine Learning*, pp. 19773–19808. PMLR, 2022.

Shen, Y. and Sanghavi, S. Learning with bad training data via iterative trimmed loss minimization. In *International Conference on Machine Learning*, pp. 5739–5748. PMLR, 2019.

Sion, M. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171 – 176, 1958.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., and Li, C.-L. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Taghanaki, S. A., Zheng, Y., Zhou, S. K., Georgescu, B., Sharma, P., Xu, D., Comaniciu, D., and Hamarneh, G. Combo loss: Handling input and output imbalance in multi-organ segmentation. *Computerized Medical Imaging and Graphics*, 75:24–33, 2019.

Tang, Y., Wang, X., Harrison, A. P., Lu, L., Xiao, J., and Summers, R. M. Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs. In *International Workshop on Machine Learning in Medical Imaging*, pp. 249–258. Springer, 2018.

Tullis, J. G. and Benjamin, A. S. On the effectiveness of self-paced learning. *Journal of memory and language*, 64 (2):109–118, 2011.

Wang, X., Chen, H., Xiang, H., Lin, H., Lin, X., and Heng, P.-A. Deep virtual adversarial self-training with consistency regularization for semi-supervised medical image classification. *Medical image analysis*, 70:102010, 2021a.

Wang, X., Chen, Y., and Zhu, W. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021b.

Wang, Y., Liu, W., Ma, X., Bailey, J., Zha, H., Song, L., and Xia, S.-T. Iterative learning with open-set noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8688–8696, 2018.

Wong, K. C., Moradi, M., Tang, H., and Syeda-Mahmood, T. 3d segmentation with exponential logarithmic loss for highly unbalanced object sizes. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III 11*, pp. 612–619. Springer, 2018.

Wu, X., Xie, Y., Du, S. S., and Ward, R. Adaloss: A computationally-efficient and provably convergent adaptive gradient method. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8):8691–8699, Jun. 2022.

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

Yang, S., Dong, Y., Ward, R., Dhillon, I. S., Sanghavi, S., and Lei, Q. Sample efficiency of data augmentation consistency regularization. In *International Conference on Artificial Intelligence and Statistics*, pp. 3825–3853. PMLR, 2023.

Yeung, M., Sala, E., Schönlieb, C.-B., and Rundo, L. Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Computerized Medical Imaging and Graphics*, 95: 102026, 2022. ISSN 0895-6111.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Zhang, Y., Zhou, B., Chen, L., Wu, Y., and Zhou, H. Multi-transformation consistency regularization for semi-supervised medical image segmentation. In *2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pp. 485–489. IEEE, 2021.

Zhao, A., Balakrishnan, G., Durand, F., Guttag, J. V., and Dalca, A. V. Data augmentation using learned transformations for one-shot medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8543–8553, 2019.

Zhao, P. and Zhang, T. Stochastic optimization with importance sampling for regularized loss minimization. In *international conference on machine learning*, pp. 1–9. PMLR, 2015.

Zhou, H.-Y., Wang, C., Li, H., Wang, G., Zhang, S., Li, W., and Yu, Y. Ssmd: semi-supervised medical image detection with adaptive consistency and heterogeneous perturbation. *Medical Image Analysis*, 72:102117, 2021.

A Separation of Label-sparse and Label-dense Samples

Proof of Proposition 3.3. We first observe that, since $\ell_{CE}(\theta; (\mathbf{x}, \mathbf{y}))$ and $\ell_{AC}(\theta; \mathbf{x}, A_1, A_2)$ are convex and continuous in θ for all $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$ and $A_1, A_2 \sim \mathcal{A}^2$, for all $i \in [n]$, $\widehat{L}_i^{WAC}(\theta, \boldsymbol{\beta})$ is continuous, convex in θ, and affine (thus concave) in $\boldsymbol{\beta}$; and therefore so is $\widehat{L}^{WAC}(\theta, \boldsymbol{\beta})$. Then with the compact and convex domains $\theta \in \mathcal{F}_{\theta^*}(\gamma)$ and $\boldsymbol{\beta} \in [0,1]^n$, Sion's minimax theorem (Sion, 1958) suggests the minimax equality,

$$\min_{\theta \in \mathcal{F}_{\theta^*}(\gamma)} \max_{\boldsymbol{\beta} \in [0,1]^n} \widehat{L}^{WAC}(\theta, \boldsymbol{\beta}) = \max_{\boldsymbol{\beta} \in [0,1]^n} \min_{\theta \in \mathcal{F}_{\theta^*}(\gamma)} \widehat{L}^{WAC}(\theta, \boldsymbol{\beta}), \tag{7}$$

where $\inf, \sup$ can be replaced by $\min, \max$ respectively due to compactness of the domains.

Further, by the continuity and convexity-concavity of $\widehat{L}^{WAC}(\theta, \boldsymbol{\beta})$, the pointwise maximum $\max_{\boldsymbol{\beta} \in [0,1]^n} \widehat{L}^{WAC}(\theta, \boldsymbol{\beta})$ is lower semi-continuous and convex in θ while the pointwise minimum $\min_{\theta \in \mathcal{F}_{\theta^*}(\gamma)} \widehat{L}^{WAC}(\theta, \boldsymbol{\beta})$ is upper semi-continuous and concave in $\boldsymbol{\beta}$. Then via Weierstrass' theorem ((Bertsekas, 2009), Proposition 3.2.1), there exist $\widehat{\theta}^{WAC} \in \mathcal{F}_{\theta^*}(\gamma)$ and $\widehat{\boldsymbol{\beta}} \in [0,1]^n$ that achieve the minimax optimal by minimizing $\max_{\boldsymbol{\beta} \in [0,1]^n} \widehat{L}^{WAC}(\theta, \boldsymbol{\beta})$ and maximizing $\min_{\theta \in \mathcal{F}_{\theta^*}(\gamma)} \widehat{L}^{WAC}(\theta, \boldsymbol{\beta})$. Along with Equation (7), such $\left(\widehat{\theta}^{WAC}, \widehat{\boldsymbol{\beta}}\right)$ provides a saddle point for Equation (5) ((Bertsekas, 2009), Proposition 3.4.1).

Next, we show via contradiction that there exists a saddle point with $\widehat{\boldsymbol{\beta}}$ attained on a vertex $\widehat{\boldsymbol{\beta}} \in \{0,1\}^n$. Suppose the opposite, then for any saddle point $\left(\widehat{\theta}^{WAC}, \widehat{\boldsymbol{\beta}}\right)$, there must be an $i \in [n]$ with $\widehat{\boldsymbol{\beta}}_{[i]} \in (0,1)$, where we have the following contradictions:

(i) If $\ell_{CE}\left(\widehat{\theta}^{WAC}; (\mathbf{x}_i, \mathbf{y}_i)\right) < \ell_{AC}\left(\widehat{\theta}^{WAC}; \mathbf{x}_i, A_{i,1}, A_{i,2}\right)$, decreasing $\widehat{\boldsymbol{\beta}}_{[i]} > 0$ to $\widehat{\boldsymbol{\beta}}'_{[i]} = 0$ leads to $\widehat{L}^{WAC}\left(\widehat{\theta}^{WAC}, \widehat{\boldsymbol{\beta}}'\right) > \widehat{L}^{WAC}\left(\widehat{\theta}^{WAC}, \widehat{\boldsymbol{\beta}}\right)$, contradicting Equation (6).

(ii) If $\ell_{CE}\left(\widehat{\theta}^{WAC}; (\mathbf{x}_i, \mathbf{y}_i)\right) > \ell_{AC}\left(\widehat{\theta}^{WAC}; \mathbf{x}_i, A_{i,1}, A_{i,2}\right)$, increasing $\widehat{\boldsymbol{\beta}}_{[i]} < 1$ to $\widehat{\boldsymbol{\beta}}'_{[i]} = 1$ again leads to $\widehat{L}^{WAC}\left(\widehat{\theta}^{WAC}, \widehat{\boldsymbol{\beta}}'\right) > \widehat{L}^{WAC}\left(\widehat{\theta}^{WAC}, \widehat{\boldsymbol{\beta}}\right)$, contradicting Equation (6).

(iii) If $\ell_{CE}\left(\widehat{\theta}^{WAC}; (\mathbf{x}_i, \mathbf{y}_i)\right) = \ell_{AC}\left(\widehat{\theta}^{WAC}; \mathbf{x}_i, A_{i,1}, A_{i,2}\right)$, $\widehat{\boldsymbol{\beta}}_{[i]}$ can be replaced with any value in $[0,1]$, including $0, 1$.

Therefore, there must be a saddle point $\left(\widehat{\theta}^{WAC}, \widehat{\boldsymbol{\beta}}\right)$ with $\widehat{\boldsymbol{\beta}} \in \{0,1\}^n$ such that

$$\boldsymbol{\beta}_{[i]} = \mathbb{I}\left\{\ell_{CE}\left(\widehat{\theta}^{WAC}; (\mathbf{x}_i, \mathbf{y}_i)\right) > \ell_{AC}\left(\widehat{\theta}^{WAC}; \mathbf{x}_i, A_{i,1}, A_{i,2}\right)\right\}.$$

Finally, it remains to show that *w.h.p.* over $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i \in [n]} \sim P_\xi^n$ and $\{(A_{i,1}, A_{i,2})\}_{i \in [n]} \sim \mathcal{A}^{2n}$,

(i) $\ell_{CE}\left(\widehat{\theta}^{WAC}; (\mathbf{x}_i, \mathbf{y}_i)\right) \leq \ell_{AC}\left(\widehat{\theta}^{WAC}; \mathbf{x}_i, A_{i,1}, A_{i,2}\right)$ for all $(\mathbf{x}_i, \mathbf{y}_i) \sim P_0$; and

(ii) $\ell_{CE}\left(\widehat{\theta}^{WAC}; (\mathbf{x}_i, \mathbf{y}_i)\right) > \ell_{AC}\left(\widehat{\theta}^{WAC}; \mathbf{x}_i, A_{i,1}, A_{i,2}\right)$ for all $(\mathbf{x}_i, \mathbf{y}_i) \sim P_1$;

which leads to $\boldsymbol{\beta}_{[i]} = \mathbb{I}\{(\mathbf{x}_i, \mathbf{y}_i) \sim P_1\}$ *w.h.p.* as desired. To illustrate this, we begin by observing that when P_0 and P_1 are n-separated (Assumption 3.2), since $\widehat{\theta}^{WAC} \in \mathcal{F}_{\theta^*}(\gamma)$, there exists some $\omega > 0$ such that for each $i \in [n]$,

$$\mathbb{P}\left[\ell_{CE}\left(\widehat{\theta}^{WAC}; (\mathbf{x}_i, \mathbf{y}_i)\right) < \ell_{AC}\left(\widehat{\theta}^{WAC}; \mathbf{x}_i, A_{i,1}, A_{i,2}\right) \;\middle|\; (\mathbf{x}_i, \mathbf{y}_i) \sim P_0\right] \geq 1 - \frac{1}{\Omega\left(n^{1+\omega}\right)},$$

and

$$\mathbb{P}\left[\ell_{CE}\left(\widehat{\theta}^{WAC}; (\mathbf{x}_i, \mathbf{y}_i)\right) > \ell_{AC}\left(\widehat{\theta}^{WAC}; \mathbf{x}_i, A_{i,1}, A_{i,2}\right) \;\middle|\; (\mathbf{x}_i, \mathbf{y}_i) \sim P_1\right] \geq 1 - \frac{1}{\Omega\left(n^{1+\omega}\right)}.$$

Therefore by the union bound over the set of n samples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i \in [n]} \sim P_\xi^n$,

$$\mathbb{P}\left[\ell_{CE}\left(\widehat{\theta}^{WAC}; (\mathbf{x}_i, \mathbf{y}_i)\right) < \ell_{AC}\left(\widehat{\theta}^{WAC}; \mathbf{x}_i, A_{i,1}, A_{i,2}\right) \;\forall\; (\mathbf{x}_i, \mathbf{y}_i) \sim P_0\right] \geq 1 - \frac{1}{\Omega\left(n^{\omega}\right)}, \tag{8}$$

and

$$\mathbb{P}\left[\ell_{CE}\left(\widehat{\theta}^{WAC};(\mathbf{x}_i,\mathbf{y}_i)\right) > \ell_{AC}\left(\widehat{\theta}^{WAC};\mathbf{x}_i,A_{i,1},A_{i,2}\right)\ \forall\ (\mathbf{x}_i,\mathbf{y}_i)\sim P_1\right] \geq 1-\frac{1}{\Omega\left(n^{\omega}\right)}. \tag{9}$$

Applying the union bound again on Equation (8) and Equation (9), we have the desired condition holds with probability $1-\Omega\left(n^{\omega}\right)^{-1}$, *i.e.*, *w.h.p.*. □

B Convergence of *AdaWAC*

Recall the underlying function class $\mathcal{F}\ni f_\theta$ parameterized by some $\theta\in\mathcal{F}_\theta$ that we aim to learn for the pixel-wise classifier $h_\theta=\operatorname{argmax}_{k\in[K]} f_\theta\left(\mathbf{x}\right)_{[j,:]}$, $j\in[d]$:

$$\mathcal{F}=\left\{f_\theta=\phi_\theta\circ\psi_\theta \;\middle|\; \phi_\theta:\mathcal{X}\to\mathcal{Z},\ \psi_\theta:\mathcal{Z}\to[0,1]^{d\times K}\right\}, \tag{10}$$

where ϕ_θ,ψ_θ correspond to the encoder and decoder functions. Formally, we consider an inner product space of parameters $(\mathcal{F}_\theta,\langle\cdot,\cdot\rangle_{\mathcal{F}})$ with the induced norm $\|\cdot\|_{\mathcal{F}}$ and dual norm $\|\cdot\|_{\mathcal{F},*}$.

For any $d\in\mathbb{N}$, let $\Delta_d^n\triangleq\left\{[\boldsymbol{\beta}_1;\ldots;\boldsymbol{\beta}_n]\in[0,1]^{n\times d}\;\middle|\;\|\boldsymbol{\beta}_i\|_1=1\ \forall\, i\in[n]\right\}$. Then Equation (5) can be reformulated as:

$$\begin{aligned}\widehat{\theta}^{WAC},\widehat{\boldsymbol{\beta}}&=\operatorname*{argmin}_{\theta\in\mathcal{F}_{\theta^*}(\gamma)}\ \operatorname*{argmax}_{\mathbf{B}\in\Delta_2^n}\left\{\widehat{L}^{WAC}\left(\theta,\mathbf{B}\right)\triangleq\frac{1}{n}\sum_{i=1}^{n}\widehat{L}_i^{WAC}\left(\theta,\mathbf{B}\right)\right\},\\ \widehat{L}_i^{WAC}\left(\theta,\mathbf{B}\right)&\triangleq\mathbf{B}_{[i,1]}\cdot\ell_{CE}\left(\theta;(\mathbf{x}_i,\mathbf{y}_i)\right)+\mathbf{B}_{[i,2]}\cdot\ell_{AC}\left(\theta;\mathbf{x}_i,A_{i,1},A_{i,2}\right).\end{aligned} \tag{11}$$

Proposition B.1 (Convergence (formal restatement of Proposition 4.1))**.** *Assume that $\ell_{CE}\left(\theta;(\mathbf{x},\mathbf{y})\right)$ and $\ell_{AC}\left(\theta;\mathbf{x},A_1,A_2\right)$ are convex and continuous in θ for all $(\mathbf{x},\mathbf{y})\in\mathcal{X}\times\mathcal{Y}$ and $A_1,A_2\sim\mathcal{A}^2$, and that $\mathcal{F}_{\theta^*}\left(\gamma\right)\subset\mathcal{F}_\theta$ is convex and compact. If there exist*

(i) $C_{\theta,*}>0$ such that $\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla_\theta\widehat{L}_i^{WAC}\left(\theta,\mathbf{B}\right)\right\|_{\mathcal{F},*}^2\leq C_{\theta,*}^2$ for all $\theta\in\mathcal{F}_{\theta^*}\left(\gamma\right)$, $\mathbf{B}\in\Delta_2^n$ and

(ii) $C_{\mathbf{B},*}>0$ such that $\frac{1}{n}\sum_{i=1}^{n}\max\left\{\ell_{CE}\left(\theta;(\mathbf{x}_i,\mathbf{y}_i)\right),\ell_{AC}\left(\theta;\mathbf{x}_i,A_{i,1},A_{i,2}\right)\right\}^2\leq C_{\mathbf{B},*}^2$ for all $\theta\in\mathcal{F}_{\theta^*}\left(\gamma\right)$,

then with $\eta_\theta=\eta_{\mathbf{B}}=2\Big/\sqrt{5T\left(\gamma^2C_{\theta,*}^2+2nC_{\mathbf{B},*}^2\right)}$, Algorithm 1 provides the convergence guarantee for the duality gap $\mathcal{E}\left(\overline{\theta}_T,\overline{\mathbf{B}}_T\right)\triangleq\max_{\mathbf{B}\in\Delta_2^n}\widehat{L}^{WAC}\left(\overline{\theta}_T,\mathbf{B}\right)-\min_{\theta\in\mathcal{F}_{\theta^*}(\gamma)}\widehat{L}^{WAC}\left(\theta,\overline{\mathbf{B}}_T\right)$:

$$\mathbb{E}\left[\mathcal{E}\left(\overline{\theta}_T,\overline{\mathbf{B}}_T\right)\right]\leq 2\sqrt{\frac{5\left(\gamma^2C_{\theta,*}^2+2nC_{\mathbf{B},*}^2\right)}{T}},$$

where $\overline{\theta}_T=\frac{1}{T}\sum_{t=1}^{T}\theta_t$ and $\overline{\mathbf{B}}_T=\frac{1}{T}\sum_{t=1}^{T}\mathbf{B}_t$.

Proof of Proposition B.1. The proof is an application of the standard convergence guarantee for the online mirror descent on saddle point problems, as recapitulated in Lemma B.4.

Specifically, for $\mathbf{B}\in\Delta_2^n$, we use the norm $\|\mathbf{B}\|_{1,2}\triangleq\sqrt{\sum_{i=1}^{n}\left(\sum_{j=1}^{2}\left|\mathbf{B}_{[i,j]}\right|\right)^2}$ with its dual norm $\|\mathbf{B}\|_{1,2,*}\triangleq\sqrt{\sum_{i=1}^{n}\left(\max_{j\in[2]}\left|\mathbf{B}_{[i,j]}\right|\right)^2}$. We consider a mirror map $\varphi_{\mathbf{B}}:[0,1]^{n\times 2}\to\mathbb{R}$ such that $\varphi_{\mathbf{B}}\left(\mathbf{B}\right)=\sum_{i=1}^{n}\sum_{j=1}^{2}\mathbf{B}_{[i,j]}\log\mathbf{B}_{[i,j]}$. We observe that, since $\mathbf{B}_{[i,:]},\mathbf{B'}_{[i,:]}\in\Delta_2$ for all $i\in[n]$,

$$D_{\varphi_{\mathbf{B}}}\left(\mathbf{B},\mathbf{B'}\right)=\sum_{i=1}^{n}\sum_{j=1}^{2}\mathbf{B}_{[i,j]}\log\frac{\mathbf{B}_{[i,j]}}{\mathbf{B'}_{[i,j]}}\geq\frac{1}{2}\sum_{i=1}^{n}\left(\sum_{j=1}^{2}\left|\mathbf{B}_{[i,j]}-\mathbf{B'}_{[i,j]}\right|\right)^2=\frac{1}{2}\left\|\mathbf{B}-\mathbf{B'}\right\|_{1,2}^2,$$

and therefore $\varphi_{\mathbf{B}}$ is 1-strongly convex with respect to $\|\cdot\|_{1,2}$. With such $\varphi_{\mathbf{B}}$, we have the associated Fenchel dual $\varphi_{\mathbf{B}}^*(\mathbf{G}) = \sum_{i=1}^n \log\left(\sum_{j=1}^2 \exp\left(\mathbf{G}_{[i,j]}\right)\right)$, along with the gradients

$$\nabla\varphi_{\mathbf{B}}(\mathbf{B})_{[i,j]} = 1 + \log \mathbf{B}_{[i,j]}, \quad \nabla\varphi_{\mathbf{B}}^*(\mathbf{G})_{[i,j]} = \frac{\exp\left(\mathbf{G}_{[i,j]}\right)}{\sum_{j=1}^2 \exp\left(\mathbf{G}_{[i,j]}\right)},$$

such that the mirror descent update on $\mathbf{B}$ is given by

$$\begin{aligned}(\mathbf{B}_{t+1})_{[i,j]} =& \nabla\varphi_{\mathbf{B}}^*\left(\nabla\varphi_{\mathbf{B}}(\mathbf{B}_t) - \eta_{\mathbf{B}} \cdot \nabla_{\mathbf{B}} \widehat{L}_{i_t}^{WAC}(\theta_t, \mathbf{B}_t)\right) \\ =& \frac{(\mathbf{B}_t)_{[i,j]} \exp\left(\eta_{\mathbf{B}} \cdot \left(\nabla_{\mathbf{B}} \widehat{L}_{i_t}^{WAC}(\theta_t, \mathbf{B}_t)\right)_{[i,j]}\right)}{\sum_{j=1}^2 (\mathbf{B}_t)_{[i,j]} \exp\left(\eta_{\mathbf{B}} \cdot \left(\nabla_{\mathbf{B}} \widehat{L}_{i_t}^{WAC}(\theta_t, \mathbf{B}_t)\right)_{[i,j]}\right)}.\end{aligned}$$

For $i_t \sim [n]$ uniformly, the stochastic gradient with respect to $\mathbf{B}$ satisfies

$$\begin{aligned}&\mathbb{E}_{i_t\sim[n]}\left[\left\|\nabla_{\mathbf{B}} \widehat{L}_{i_t}^{WAC}(\theta_t, \mathbf{B}_t)\right\|_{1,2,*}^2\right] \\ =& \frac{1}{n}\sum_{i_t=1}^n \max\left\{\ell_{CE}(\theta_t; (\mathbf{x}_{i_t}, \mathbf{y}_{i_t})), \ell_{AC}(\theta_t; \mathbf{x}_{i_t}, A_{i_t,1}, A_{i_t,2})\right\}^2 \leq C_{\mathbf{B},*}^2.\end{aligned}$$

Further, in the distance induced by $\varphi_{\mathbf{B}}$, we have

$$R_{\Delta_2^n}^2 \triangleq \max_{\mathbf{B}\in\Delta_2^n} \varphi_{\mathbf{B}}(\mathbf{B}) - \min_{\mathbf{B}\in\Delta_2^n} \varphi_{\mathbf{B}}(\mathbf{B}) = 0 - \sum_{i=1}^n \sum_{j=1}^2 \frac{1}{2}\log\frac{1}{2} = n.$$

Meanwhile, for $\theta \in \mathcal{F}_{\theta^*}(\gamma)$, we consider the norm $\|\theta\|_{\mathcal{F}} \triangleq \sqrt{\langle\theta,\theta\rangle_{\mathcal{F}}}$ induced by the inner product that characterizes $\mathcal{F}_\theta$, with the associated dual norm $\|\cdot\|_{\mathcal{F},*}$. We use a mirror map $\varphi_\theta : \mathcal{F}_\theta \to \mathbb{R}$ such that $\varphi_\theta(\theta) = \frac{1}{2}\|\theta - \theta^*\|_{\mathcal{F}}^2$. By observing that

$$D_{\varphi_\theta}(\theta, \theta') = \frac{1}{2}\|\theta - \theta'\|_{\mathcal{F}}^2 \quad \forall\, \theta, \theta' \in \mathcal{F}.$$

we have φ_θ being 1-strongly convex with respect to $\|\cdot\|_{\mathcal{F}}$. With the gradient of φ_θ, $\nabla\varphi_\theta(\theta) = \theta - \theta^*$, and that of its Fenchel dual $\nabla\varphi_\theta^*(g) = g + \theta^*$, at the $(t+1)$-th iteration, we have

$$\theta_{t+1} = \nabla\varphi_\theta^*\left(\nabla\varphi_\theta(\theta_t) - \eta_\theta \cdot \nabla_\theta \widehat{L}_{i_t}^{WAC}(\theta_t, \mathbf{B}_{t+1})\right) = \theta_t - \eta_\theta \cdot \nabla_\theta \widehat{L}_{i_t}^{WAC}(\theta_t, \mathbf{B}_{t+1}).$$

For $i_t \sim [n]$ uniformly, the stochastic gradient with respect to f satisfies that

$$\mathbb{E}_{i_t\sim[n]}\left[\left\|\nabla_\theta \widehat{L}_{i_t}^{WAC}(\theta_t, \mathbf{B}_{t+1})\right\|_{\mathcal{F},*}^2\right] = \frac{1}{n}\sum_{i_t=1}^n \left\|\nabla_\theta \widehat{L}_{i_t}^{WAC}(\theta_t, \mathbf{B}_{t+1})\right\|_{\mathcal{F},*}^2 \leq C_{\theta,*}^2.$$

Further, in light of the definition of $\mathcal{F}_{\theta^*}(\gamma)$, since $\theta^* \in \mathcal{F}_{\theta^*}(\gamma)$, with $\theta^* = \operatorname{argmin}_{\theta\in\mathcal{F}_{\theta^*}(\gamma)} \varphi_\theta(\theta)$ and $\theta' = \operatorname{argmax}_{\theta\in\mathcal{F}_{\theta^*}(\gamma)} \varphi_\theta(\theta)$, we have

$$R_{\mathcal{F}_{\theta^*}(\gamma)}^2 \triangleq \max_{\theta\in\mathcal{F}_{\theta^*}(\gamma)} \varphi_\theta(\theta) - \min_{\theta\in\mathcal{F}_{\theta^*}(\gamma)} \varphi_\theta(\theta) = \frac{1}{2}\|\theta' - \theta^*\|_{\mathcal{F}}^2 \leq \frac{\gamma^2}{2}.$$

Finally, leveraging Lemma B.4 completes the proof. □

We recall the standard convergence guarantee for online mirror descent on saddle point problems. In general, we consider a stochastic function $F : \mathcal{U} \times \mathcal{V} \times \mathcal{I} \to \mathbb{R}$ with the randomness of $F(u, v; i)$ on $i \in \mathcal{I}$. Overloading notation $\mathcal{I}$ both as the distribution of i and as the support, we are interested in solving the saddle point problem on the expectation function

$$\min_{u \in \mathcal{U}} \max_{v \in \mathcal{V}} f(u, v) \quad \text{where} \quad f(u, v) \triangleq \mathbb{E}_{i \sim \mathcal{I}} \left[F(u, v; i) \right]. \tag{12}$$

Assumption B.2. Assume that the stochastic objective satisfies the following:

(i) For every $i \in \mathcal{I}$, $F(\cdot, v, i)$ is convex for all $v \in \mathcal{V}$ and $F(u, \cdot, i)$ is concave for all $u \in \mathcal{U}$.

(ii) The stochastic subgradients $G_u(u, v; i) \in \partial_u F(u, v; i)$ and $G_v(u, v; i) \in \partial_v F(u, v; i)$ with respect to u and v evaluated at any $(u, v) \in \mathcal{U} \times \mathcal{V}$ provide unbiased estimators for some respective subgradients of the expectation function: for any $(u, v) \in \mathcal{U} \times \mathcal{V}$, there exist some $g_u(u, v) \triangleq \mathbb{E}_{i \sim \mathcal{I}} \left[G_u(u, v; i) \right] \in \partial_u f(u, v)$ and $g_v(u, v) \triangleq \mathbb{E}_{i \sim \mathcal{I}} \left[G_v(u, v; i) \right] \in \partial_v f(u, v)$.

(iii) Let $\|\cdot\|_{\mathcal{U}}$ and $\|\cdot\|_{\mathcal{V}}$ be arbitrary norms that are well-defined on $\mathcal{U}$ and $\mathcal{V}$, while $\|\cdot\|_{\mathcal{U},*}$ and $\|\cdot\|_{\mathcal{V},*}$ be their respective dual norms. There exist constants $C_{u,*}, C_{v,*} > 0$ such that

$$\mathbb{E}_{i \sim \mathcal{I}} \left[\|G_u(u, v; i)\|_{\mathcal{U},*}^2 \right] \leq C_{u,*}^2 \quad \text{and} \quad \mathbb{E}_{i \sim \mathcal{I}} \left[\|G_v(u, v; i)\|_{\mathcal{V},*}^2 \right] \leq C_{v,*}^2 \quad \forall\ (u, v) \in \mathcal{U} \times \mathcal{V}.$$

For online mirror descent, we further introduce two mirror maps that induce distances on $\mathcal{U}$ and $\mathcal{V}$, respectively.

Assumption B.3. Let $\varphi_u : \mathcal{D}_u \to \mathbb{R}$ and $\varphi_v : \mathcal{D}_v \to \mathbb{R}$ satisfy the following:

(i) $\mathcal{U} \subseteq \mathcal{D}_u \cup \partial \mathcal{D}_u, \mathcal{U} \cap \mathcal{D}_u \neq \emptyset$ and $\mathcal{V} \subseteq \mathcal{D}_v \cup \partial \mathcal{D}_v, \mathcal{V} \cap \mathcal{D}_v \neq \emptyset$.

(ii) φ_u is ρ_u-strongly convex with respect to $\|\cdot\|_{\mathcal{U}}$; φ_v is ρ_v-strongly convex with respect to $\|\cdot\|_{\mathcal{V}}$.

(iii) $\lim_{u \to \partial \mathcal{D}_u} \|\nabla \varphi_u(u)\|_{\mathcal{U},*} = \lim_{v \to \partial \mathcal{D}_v} \|\nabla \varphi_v(v)\|_{\mathcal{V},*} = +\infty$.

Given the learning rates η_u, η_v, in each iteration $t = 1, \ldots, T$, the online mirror descent samples $i_t \sim \mathcal{I}$ and updates

$$\begin{aligned} v_{t+1} &= \operatorname*{argmin}_{v \in \mathcal{V}} \ -\eta_v \cdot G_v(u_t, v_t; i_t)^\top v + D_{\varphi_v}(v, v_t), \\ u_{t+1} &= \operatorname*{argmin}_{u \in \mathcal{U}} \ \eta_u \cdot G_u(u_t, v_{t+1}; i_t)^\top u + D_{\varphi_u}(u, u_t), \end{aligned} \tag{13}$$

where $D_\varphi(w, w') = \varphi(w) - \varphi(w') - \nabla \varphi(w')^\top (w - w')$ denotes the Bregman divergence.

We measure the convergence of the saddle point problem in the duality gap:

$$\mathcal{E}(\overline{u}_T, \overline{v}_T) \triangleq \max_{v \in \mathcal{V}} f(\overline{u}_T, v) - \min_{u \in \mathcal{U}} f(u, \overline{v}_T)$$

such that, with

$$R_{\mathcal{U}} \triangleq \sqrt{\max_{u \in \mathcal{U} \cap \mathcal{D}_u} \varphi_u(u) - \min_{u \in \mathcal{U} \cap \mathcal{D}_u} \varphi_u(u)} \quad \text{and} \quad R_{\mathcal{V}} \triangleq \sqrt{\max_{v \in \mathcal{V} \cap \mathcal{D}_v} \varphi_v(v) - \min_{v \in \mathcal{V} \cap \mathcal{D}_v} \varphi_v(v)},$$

the online mirror descent converges as follows.

Lemma B.4 ((Nemirovski et al., 2009) (3.11))**.** *Under Assumption B.2 and Assumption B.3, when taking constant learning rates* $\eta_u = \eta_v = 2 \Big/ \sqrt{5T \left(\frac{2R_{\mathcal{U}}^2}{\rho_u} C_{u,*}^2 + \frac{2R_{\mathcal{V}}^2}{\rho_v} C_{v,*}^2 \right)}$*, with* $\overline{u}_T = \frac{1}{T} \sum_{t=1}^T u_t$ *and* $\overline{v}_T = \frac{1}{T} \sum_{t=1}^T v_t$*,*

$$\mathbb{E}\left[\mathcal{E}(\overline{u}_T, \overline{v}_T) \right] \leq 2 \sqrt{\frac{10 \left(\rho_v R_{\mathcal{U}}^2 C_{u,*}^2 + \rho_u R_{\mathcal{V}}^2 C_{v,*}^2 \right)}{\rho_u \rho_v \cdot T}}.$$

Example 1 (Binary linear pixel-wise classifiers with convex and continuous objectives). We consider a pixel-wise binary classification problem with $\mathcal{X} = [0, 1]^d$, augmentations $A : \mathcal{X} \to \mathcal{X}$ for all $A \sim \mathcal{A}$, and a class of linear "UNets",

$$\mathcal{F} = \left\{ f_\theta : \mathcal{X} \to [0, 1]^d \ \middle| \ f_\theta(\mathbf{x}) = \sigma\left(\boldsymbol{\theta}_d \boldsymbol{\theta}_e^\top \mathbf{x} \right) = \psi_\theta\left(\phi_\theta(\mathbf{x}) \right),\ \phi_\theta(\mathbf{x}) = \frac{1}{\sqrt{d}} \boldsymbol{\theta}_e^\top \mathbf{x} \right\},$$

where the parameter space $\theta = (\boldsymbol{\theta}_e, \boldsymbol{\theta}_d) \in \mathcal{F}_\theta = \mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$ is equipped with the ℓ_2 norm $\|\theta\|_{\mathcal{F}} = \left(\|\boldsymbol{\theta}_e\|_2^2 + \|\boldsymbol{\theta}_d\|_2^2\right)^{1/2}$; $\sigma : \mathbb{R}^d \to [0,1]^d$ denotes entry-wise application of the sigmoid function $\sigma(z) = (1 + e^{-z})^{-1}$; and the latent space of encoder outputs $(\mathcal{Z}, \varrho)$ is simply the real line. Given the data distribution P_ξ, we recall that $\theta^* = \operatorname{argmin}_{\theta \in \mathcal{F}_\theta} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim P_\xi}\left[\ell_{CE}\left(\theta; (\mathbf{x}, \mathbf{y})\right)\right]$ for all $\xi \in [0,1]$ and let $\mathcal{F}_{\theta^*}(\gamma) = \{\theta \in \mathcal{F}_\theta \mid \|\theta - \theta^*\|_{\mathcal{F}} \leq \gamma\}$ for some $\gamma = O\left(1/\sqrt{d}\right)$. We assume that $\left|\mathbf{x}^\top \boldsymbol{\theta}_e^*\right| = O(1)$ for all $\mathbf{x} \in \mathcal{X}$. Then, $\ell_{CE}\left(\theta; (\mathbf{x}, \mathbf{y})\right)$ and $\ell_{AC}\left(\theta; \mathbf{x}, A_1, A_2\right)$ are convex and continuous in θ for all $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times [K]^d$, $A_1, A_2 \sim \mathcal{A}^2$; while $C_{\theta,*} \leq \max\left(2\sqrt{2}, 2\lambda_{AC}\right)$ and $C_{\boldsymbol{\beta},*} \leq \max\left(O(1), 2\lambda_{AC}\right)$.

Rationale for Example 1. Let $\mathbf{y}_k = \mathbb{I}\{\mathbf{y} = k\}$ entry-wise for $k = 0, 1$. We would like to show that, for any given $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times [K]^d$, $A_1, A_2 \sim \mathcal{A}^2$,

$$\ell_{CE}\left(\theta\right) = -\frac{1}{d}\left(\mathbf{y}_1^\top \log \sigma\left(\boldsymbol{\theta}_d \boldsymbol{\theta}_e^\top \mathbf{x}\right) + \mathbf{y}_0^\top \log \sigma\left(-\boldsymbol{\theta}_d \boldsymbol{\theta}_e^\top \mathbf{x}\right)\right),$$
$$\ell_{AC}\left(\theta\right) = \frac{\lambda_{AC}}{\sqrt{d}} \cdot \left(A_1(\mathbf{x}) - A_2(\mathbf{x})\right)^\top \boldsymbol{\theta}_e$$

are convex and continuous in $\theta = (\boldsymbol{\theta}_e, \boldsymbol{\theta}_d)$.

First, we observe that $\ell_{AC}\left(\theta\right)$ is linear (and therefore convex and continuous) in θ for all $\mathbf{x} \in \mathcal{X}$, $A_1, A_2 \sim \mathcal{A}^2$, with

$$\nabla_{\boldsymbol{\theta}_e} \ell_{AC}\left(\theta\right) = \frac{\lambda_{AC}}{\sqrt{d}} \cdot \left(A_1(\mathbf{x}) - A_2(\mathbf{x})\right), \quad \nabla_{\boldsymbol{\theta}_d} \ell_{AC}\left(\theta\right) = \mathbf{0}$$

such that $\|\nabla_\theta \ell_{AC}\left(\theta\right)\|_{\mathcal{F},*} \leq 2\lambda_{AC}$.

Meanwhile, with $\mathbf{z}\left(\theta\right) = \boldsymbol{\theta}_d \boldsymbol{\theta}_e^\top \mathbf{x}$, we have $\ell_{CE}\left(\theta\right) = -\frac{1}{d}\left(\mathbf{y}_1^\top \log \sigma\left(\mathbf{z}\left(\theta\right)\right) + \mathbf{y}_0^\top \log \sigma\left(-\mathbf{z}\left(\theta\right)\right)\right)$ being convex and continuous in $\mathbf{z}\left(\theta\right)$:

$$\nabla_{\mathbf{z}}^2 \ell_{CE}\left(\theta\right) = \frac{1}{d} \operatorname{diag}\left(\sigma\left(\mathbf{z}\left(\theta\right)\right)\right) \operatorname{diag}\left(1 - \sigma\left(\mathbf{z}\left(\theta\right)\right)\right) \succcurlyeq 0.$$

Therefore, $\ell_{CE}\left(\theta\right)$ is convex and continuous in θ for all $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times [K]^d$:

$$\underbrace{\nabla_\theta^2 \ell_{CE}\left(\theta\right)}_{2d \times 2d} = \begin{bmatrix} \mathbf{x}\boldsymbol{\theta}_d^\top \\ \left(\boldsymbol{\theta}_e^\top \mathbf{x}\right) \mathbf{I}_d \end{bmatrix} \left(\frac{1}{d} \operatorname{diag}\left(\sigma\left(\mathbf{z}\left(\theta\right)\right)\right) \operatorname{diag}\left(1 - \sigma\left(\mathbf{z}\left(\theta\right)\right)\right)\right) \begin{bmatrix} \mathbf{x}\boldsymbol{\theta}_d^\top & \left(\boldsymbol{\theta}_e^\top \mathbf{x}\right) \mathbf{I}_d \end{bmatrix} \succcurlyeq 0,$$

where $\mathbf{I}_d$ denotes the $d \times d$ identity matrix. Further, from the derivation, we have

$$\nabla_{\boldsymbol{\theta}_e} \ell_{CE}\left(\theta\right) = \frac{1}{d} \boldsymbol{\theta}_d^\top \left(\sigma\left(\boldsymbol{\theta}_d \boldsymbol{\theta}_e^\top \mathbf{x}\right) - \mathbf{y}\right) \mathbf{x}, \quad \nabla_{\boldsymbol{\theta}_d} \ell_{CE}\left(\theta\right) = \frac{\boldsymbol{\theta}_e^\top \mathbf{x}}{d} \left(\sigma\left(\boldsymbol{\theta}_d \boldsymbol{\theta}_e^\top \mathbf{x}\right) - \mathbf{y}\right)$$

such that $\|\nabla_\theta \ell_{CE}\left(\theta\right)\|_{\mathcal{F},*} = \sqrt{\|\nabla_{\boldsymbol{\theta}_e} \ell_{CE}\left(\theta\right)\|_2^2 + \|\nabla_{\boldsymbol{\theta}_d} \ell_{CE}\left(\theta\right)\|_2^2} \leq 2\sqrt{2}$.

Finally, knowing $\|\nabla_\theta \ell_{CE}\left(\theta\right)\|_{\mathcal{F},*} \leq 2\sqrt{2}$ and $\|\nabla_\theta \ell_{AC}\left(\theta\right)\|_{\mathcal{F},*} \leq 2\lambda_{AC}$, we have

$$\left\|\nabla_\theta \widehat{L}_i^{WAC}\left(\theta, \boldsymbol{\beta}\right)\right\|_{\mathcal{F},*} \leq \boldsymbol{\beta}_{[i]} \|\nabla_\theta \ell_{CE}\left(\theta\right)\|_{\mathcal{F},*} + \left(1 - \boldsymbol{\beta}_{[i]}\right) \|\nabla_\theta \ell_{AC}\left(\theta\right)\|_{\mathcal{F},*} \leq \max\left(2\sqrt{2}, 2\lambda_{AC}\right)$$

for all $i \in [n]$, and therefore,

$$C_{\theta,*} \leq \max\left(2\sqrt{2}, 2\lambda_{AC}\right).$$

Besides, with

$$\ell_{AC}\left(\theta\right) \leq \frac{\lambda_{AC}}{\sqrt{d}} \|A_1(\mathbf{x}) - A_2(\mathbf{x})\|_2 \|\boldsymbol{\theta}_e\|_2 \leq 2\lambda_{AC},$$

and since

$$\begin{aligned}\left(\boldsymbol{\theta}_d\boldsymbol{\theta}_e^\top\mathbf{x}\right)_{[j]} \leq & \left|\mathbf{x}^\top\boldsymbol{\theta}_e\right| \leq \left|\mathbf{x}^\top\left(\boldsymbol{\theta}_e-\boldsymbol{\theta}_e^*\right)\right| + \left|\mathbf{x}^\top\boldsymbol{\theta}_e^*\right| \leq \|\mathbf{x}\|_2 \left\|\boldsymbol{\theta}_e-\boldsymbol{\theta}_e^*\right\|_2 + O(1) \\ \leq & \gamma\sqrt{d} + O(1) = O(1)\end{aligned}$$

for all $j \in [d]$, $\ell_{CE}(\theta) \leq \log\left(1+e^{O(1)}\right) = O(1)$, we have

$$C_{\boldsymbol{\beta},*} \leq \max\left(O(1), 2\lambda_{AC}\right).$$

□

C Dice Loss for Pixel-wise Class Imbalance

With finite samples in practice, since the averaged cross-entropy loss (Equation (2)) weights each pixel in the image label equally, the pixel-wise class imbalance can become a problem. For example, the background pixels can be dominant in most of the segmentation labels, making the classifier prone to predict pixels as background.

It is worth highlighting that despite the similar terminology "imbalance", the "class imbalance" (*i.e.*, discrepancies among numbers of labeled pixels *in different classes*) is a fundamentally different problem from the "information imbalance" studied in this work that is caused by the concept shift in segmentation labels *across different samples*.

To cope with "class imbalance", (Chen et al., 2021; Cao et al., 2023; Wong et al., 2018; Taghanaki et al., 2019; Yeung et al., 2022) propose to combine the cross-entropy loss with the *dice loss*—a popular segmentation loss based on the overlap between true labels and their corresponding predictions in each class:

$$\ell_{DICE}\left(\theta;(\mathbf{x},\mathbf{y})\right) = 1 - \frac{1}{K}\sum_{k=1}^{K} DSC\left(f_\theta\left(\mathbf{x}\right)_{[:,k]}, \mathbb{I}\left\{\mathbf{y}=k\right\}\right), \tag{14}$$

where for any $\mathbf{p} \in [0,1]^d$, $\mathbf{q} \in \{0,1\}^d$, $DSC(\mathbf{p},\mathbf{q}) = \frac{2\mathbf{p}^\top\mathbf{q}}{\|\mathbf{p}\|_1+\|\mathbf{q}\|_1} \in [0,1]$ denotes the dice coefficient (Milletari et al., 2016; Asgari Taghanaki et al., 2021). Notice that by measuring the bounded dice coefficient for each of the K classes individually, the dice loss tends to be robust to class imbalance.

(Taghanaki et al., 2019) merges both dice and averaged cross-entropy losses via a convex combination. It is also a common practice to add a smoothing term in both the nominator and denominator of the DSC (Russell & Norvig, 2016).

Combining the dice loss (Equation (14)) with the weighted augmentation consistency regularization formulation (Equation (5)), in practice, we solve

$$\begin{aligned}&\widehat{\theta}^{WAC}, \widehat{\boldsymbol{\beta}} \in \underset{\theta\in\mathcal{F}_{\theta^*}(\gamma)}{\operatorname{argmin}} \underset{\boldsymbol{\beta}\in[0,1]^n}{\operatorname{argmax}} \left\{\widehat{L}^{WAC}\left(\theta,\boldsymbol{\beta}\right) \triangleq \frac{1}{n}\sum_{i=1}^{n}\widehat{L}_i^{WAC}\left(\theta,\boldsymbol{\beta}\right)\right\} \\ &\widehat{L}_i^{WAC}\left(\theta,\boldsymbol{\beta}\right) \triangleq \ell_{DICE}\left(\theta;(\mathbf{x}_i,\mathbf{y}_i)\right) + \boldsymbol{\beta}_{[i]}\cdot\ell_{CE}\left(\theta;(\mathbf{x}_i,\mathbf{y}_i)\right) + \left(1-\boldsymbol{\beta}_{[i]}\right)\cdot\ell_{AC}\left(\theta;\mathbf{x}_i,A_{i,1},A_{i,2}\right)\end{aligned} \tag{15}$$

with a slight modification in Algorithm 1 line 9:

$$\begin{aligned}\theta_t \leftarrow \theta_{t-1} - \eta_\theta\cdot\Big(&\nabla_\theta\ell_{DICE}\left(\theta_{t-1};(\mathbf{x}_{i_t},\mathbf{y}_{i_t})\right) + \left(\boldsymbol{\beta}_t\right)_{[i_t]}\cdot\nabla_\theta\ell_{CE}\left(\theta_{t-1};(\mathbf{x}_{i_t},\mathbf{y}_{i_t})\right) \\ &+ \left(1-\left(\boldsymbol{\beta}_t\right)_{[i_t]}\right)\cdot\nabla_\theta\ell_{AC}\left(\theta_{t-1};\mathbf{x}_{i_t},A_{i_t,1},A_{i_t,2}\right)\Big).\end{aligned}$$

On the influence of incorporating dice loss in experiments. We note that, in the experiments, the dice loss ℓ_{DICE} is treated independently of *AdaWAC* in Algorithm 1 via standard stochastic gradient descent. In particular for the comparison with hard-thresholding algorithms in Table 2, we keep the updating on ℓ_{DICE} of the original untrimmed batch intact for both **trim-train** and **trim-ratio** to exclude the potential effect of ℓ_{DICE} that is not involved in reweighting.

D Implementation Details and Datasets

We follow the official implementation of TransUNet[10] for model training. We use the same optimizer (SGD with learning rate 0.01, momentum 0.9, and weight decay 1e-4). For the Synapse dataset, we train TransUNet for 150 epochs on the training dataset and evaluate the last-iteration model on the test dataset. For the ACDC dataset, we train TransUNet for 360 epochs in total, while validating models on the ACDC validation dataset for every 10 epochs and testing on the best model selected by the validation. The total number of training iterations (*i.e.*, total number of batches) is set to be the same as that in the vanilla TransUNet (Chen et al., 2021) experiments. In particular, the results in Table 1 are averages (and standard deviations) over 3 arbitrary random seeds. The results in Table 2, Table 3, and Table 4 are given by the original random seed used in the TransUNet experiments.

Synapse multi-organ segmentation dataset (Synapse). The Synapse dataset[11] is multi-organ abdominal CT scans for medical image segmentation in the MICCAI 2015 Multi-Atlas Abdomen Labelling Challenge (Chen et al., 2021). There are 30 cases of CT scans with variable sizes $(512 \times 512 \times 85 - 512 \times 512 \times 198)$, and slice thickness ranges from 2.5mm to 5.0mm. We use the pre-processed data provided by (Chen et al., 2021) and follow their train/test split to use 18 cases for training and 12 cases for testing on 8 abdominal organs—aorta, gallbladder, left kidney (L), right kidney (R), liver, pancreas, spleen, and stomach. The abdominal organs were labeled by experience undergraduates and verified by a radiologist using MIPAV software according to the information from Synapse wiki page.

Automated cardiac diagnosis challenge dataset (ACDC). The ACDC dataset[12] is cine-MRI scans in the MICCAI 2017 Automated Cardiac Diagnosis Challenge. There are 200 scans from 100 patients, and each patient has two frames with slice thickness from 5mm to 8mm. We use the pre-processed data also provided by (Chen et al., 2021) and follow their train/validate/test split to use 70 patients' scans for training, 10 patients' scans for validation, and 20 patients' scans for testing on three cardiac structures—left ventricle (LV), myocardium (MYO), and right ventricle (RV). The data were labeled by one clinical expert according to the description on ACDC dataset website.

E Additional Experimental Details and Results

E.1 Pure Sample-reweighting Algorithms

Here, we provide a zoomed view of the connection and differences between the two pure sample-reweighting methods based on DRO in the ablation study (Section 5.2) — "reweight-only" (standard DRO) and "reweight-EM" (DRO with entropy maximization) — from both formulation and optimization perspectives.

In particular, the "reweight-only" (standard DRO) method takes the formulation

$$\min_{\theta \in \mathcal{F}_{\theta^*}(\gamma)} \max_{\boldsymbol{\beta} \in \Delta_n} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\beta}_{[i]} \cdot \ell_{CE}\left(\theta; (\mathbf{x}_i, \mathbf{y}_i)\right); \tag{16}$$

while the "reweight-EM" (DRO with entropy maximization) formulation (Fidon et al., 2021) can be expressed as[13]

$$\min_{\theta \in \mathcal{F}_{\theta^*}(\gamma)} \max_{\boldsymbol{\beta} \in \Delta_n} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\beta}_{[i]} \cdot \ell_{CE}\left(\theta; (\mathbf{x}_i, \mathbf{y}_i)\right) - \lambda_{EM} \cdot \sum_{i=1}^{n} \boldsymbol{\beta}_{[i]} \log\left(\boldsymbol{\beta}_{[i]}\right), \tag{17}$$

where we follow the default hyper-parameter setting in (Fidon et al., 2021) and set $\lambda_{EM} = 0.01$.

Compared with the WAC regularization formulation in Equation (5), apart from the apparent difference in (missing) the consistency regularization term, the key discrepancy between Equation (16), Equation (17) and Equation (5) lies in $\boldsymbol{\beta}$. That is, for the DRO-based formulations (Equation (16) and Equation (17)), $\boldsymbol{\beta} \in \Delta_n$ is a distribution over the n samples where samples are reweighted solely based on their respective cross-entropy losses; whereas for Equation (5) with $\boldsymbol{\beta} \in [0,1]^n$, each $\left(\boldsymbol{\beta}_{[i]}, 1 - \boldsymbol{\beta}_{[i]}\right)$ is a probability distribution that quantifies the relative importance of the cross-entropy loss versus the consistency regularization for the i-th sample.

[10] https://github.com/Beckschen/TransUNet

[11] See detailed description at https://www.synapse.org/#!Synapse:syn3193805/wiki/217789

[12] See detailed description at https://www.creatis.insa-lyon.fr/Challenge/acdc/

[13] The original formulation in (Fidon et al., 2021) uses the KL-divergence between $\boldsymbol{\beta}$ and the uniform distribution $\mathbf{1}/n$ to describe the entropy term: $-D_{KL}\left(\boldsymbol{\beta} \| \mathbf{1}/n\right) = -\log(n) - \sum_{i=1}^{n} \boldsymbol{\beta}_{[i]} \log\left(\boldsymbol{\beta}_{[i]}\right)$.

From the optimization perspective, with mirror descent via a mirror map $\varphi_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{\beta}_{[i]} \log\left(\boldsymbol{\beta}_{[i]}\right)$ on $\boldsymbol{\beta}$ and standard stochastic gradient descent on θ, the "reweight-only" (standard DRO) method (Sagawa et al., 2020) updates the sample weights $\boldsymbol{\beta}$ via (*cf.* Algorithm 1)

$$\boldsymbol{\beta}' \leftarrow \boldsymbol{\beta}_{t-1}, \quad \boldsymbol{\beta}'_{[i_t]} \leftarrow \boldsymbol{\beta}'_{[i_t]} \cdot \exp\left(\eta_{\boldsymbol{\beta}} \cdot \ell_{CE}\left(\theta_{t-1}; (\mathbf{x}_{i_t}, \mathbf{y}_{i_t})\right)\right), \quad \boldsymbol{\beta}_t \leftarrow \boldsymbol{\beta}' / \left\|\boldsymbol{\beta}'\right\|_1;$$

while the "reweight-EM" (DRO with entropy maximization) method (Fidon et al., 2021) updates the sample weights $\boldsymbol{\beta}$ with an additional tunable exponent on $\boldsymbol{\beta}$, characterized by the hyper-parameter λ_{EM}:

$$\boldsymbol{\beta}' \leftarrow \boldsymbol{\beta}_{t-1}, \quad \boldsymbol{\beta}'_{[i_t]} \leftarrow \left(\boldsymbol{\beta}'_{[i_t]}\right)^{1-\eta_{\boldsymbol{\beta}}\lambda_{EM}} \cdot \exp\left(\eta_{\boldsymbol{\beta}} \cdot \ell_{CE}\left(\theta_{t-1}; (\mathbf{x}_{i_t}, \mathbf{y}_{i_t})\right)\right), \quad \boldsymbol{\beta}_t \leftarrow \boldsymbol{\beta}' / \left\|\boldsymbol{\beta}'\right\|_1.$$

E.2 Sample Efficiency and Robustness of *AdaWAC* with UNet

In addition to the empirical evidence on TransUNet presented in Table 1, here, we demonstrate that the sample efficiency and distributional robustness of *AdaWAC* extend to the more widely used UNet architecture. In Table 5, analogous to Table 1, the experiments on the **full** and **half-slice** datasets provide evidence for the *sample efficiency* of *AdaWAC* compared to the baseline (ERM+SGD) on UNet. Meanwhile, the *distributional robustness* of *AdaWAC* with UNet is well illustrated by the **half-vol** and **half-sparse** experiments.

Table 5: *AdaWAC* with UNet trained on the full Synapse and its subsets

Training	Method	DSC ↑	HD95 ↓	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
full	baseline	74.04 ± 1.52	36.65 ± 0.33	84.93	55.59	77.59	70.92	92.21	55.01	82.87	73.21
	AdaWAC	$\mathbf{76.71 \pm 0.62}$	$\mathbf{30.67 \pm 2.85}$	85.68	55.19	80.15	75.45	94.11	56.19	87.54	81.39
half-slice	baseline	73.09 ± 0.10	40.05 ± 4.99	83.23	53.18	74.69	71.51	92.74	52.81	83.85	72.71
	AdaWAC	$\mathbf{75.12 \pm 0.78}$	$\mathbf{29.26 \pm 2.16}$	85.15	55.77	79.29	72.47	93.71	54.93	86.09	73.53
half-vol	baseline	63.21 ± 2.53	64.20 ± 4.46	79.46	45.79	55.79	54.91	88.65	41.61	71.68	67.77
	AdaWAC	$\mathbf{71.09 \pm 1.14}$	$\mathbf{39.95 \pm 7.76}$	83.15	49.14	75.74	70.33	90.47	44.81	82.34	72.75
half-sparse	baseline	37.30 ± 1.32	69.67 ± 2.89	61.57	8.33	57.45	50.44	60.28	23.51	17.83	18.99
	AdaWAC	$\mathbf{44.85 \pm 1.03}$	$\mathbf{62.40 \pm 5.17}$	71.56	8.40	65.42	62.73	74.02	24.16	36.65	15.88

Implementation details of UNet experiments. For the backbone architecture of experiments in Table 5, we use a UNet with a ResNet-34 encoder initialized with ImageNet pre-trained weights. We leverage the implementation of UNet and load the pre-trained model via the PyTorch API for segmentation models (Iakubovskii, 2019). For training, we use the same optimizer (SGD with learning rate 0.01, momentum 0.9, and weight decay 1e-4) and the total number of epochs (150 epochs on Synapse training set) as the TransUNet experiments, evaluating the last-iteration model on the test dataset. As before, the results in Table 5 are averages (and standard deviations) over 3 arbitrary random seeds.

E.3 Visualization of Segmentation on ACDC dataset

As shown in Figure 4, the model trained by *AdaWAC* segments cardiac structures with more accurate shapes (column 1), identifies organs missed by baseline TransUNet (column 2-3) and circumvents the false-positive pixel classifications (*i.e.*, fake predictions of background pixels as organs) suffered by the TransUNet baseline (column 4-6).

E.4 Visualization of Segmentation on Synapse with Distributional Shift

Figure 5 visualizes the segmentation predictions on 6 Synapse test slices made by models trained via *AdaWAC* (ours) and via the baseline (ERM+SGD) with TransUNet (Chen et al., 2021) on the **half-sparse** subset of the Synapse training set. We observe that, although the segmentation performances of both the baseline and *AdaWAC* are compromised by the extreme scarcity of label-dense samples and the severe distributional shift, *AdaWAC* provides more accurate predictions on the relative positions of organs, as well as less misclassification of organs (*e.g.*, the baseline tends to misclassify other organs and the background as the left kidney). Nevertheless, due to the scarcity of labels, both the model trained with *AdaWAC* and that trained with the baseline fail to make good predictions on the segmentation boundaries.

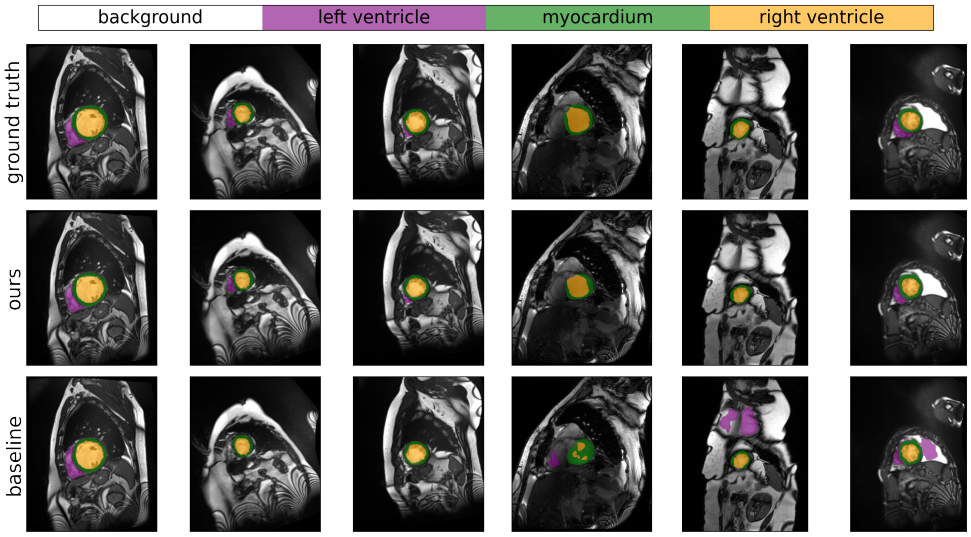

Figure 4: Visualization of segmentation results on ACDC dataset. From top to bottom: ground truth, ours, and baseline method.

E.5 Experimental Results on Previous Metrics

In this section, we include the results of experiments on Synapse[14] dataset with metrics defined in TransUNet (Chen et al., 2021) for reference. In TransUNet (Chen et al., 2021), DSC is 1 when the sum of ground truth labels is zero (i.e., gt.sum() == 0) while the sum of predicted labels is nonzero (i.e., pred.sum() > 0). However, according to the definition of dice scores, $DSC = 2|A \cap B|/(|A| + |B|), \forall A, B$, the DSC for the above case should be 0 since the intersection is 0 and the denominator is non-zero. In our evaluation, we change the special condition for DSC as 1 to pred.sum == 0 and gt.sum() == 0 instead, in which case the denominator is 0.

Table 6: *AdaWAC* with TransUNet trained on the full Synapse and its subsets, measured by metrics in TransUNet (Chen et al., 2021).

Training	Method	DSC ↑	HD95 ↓	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
full	baseline	77.32	29.23	87.46	63.54	82.06	77.76	94.10	54.06	85.07	74.54
	AdaWAC	80.16	25.79	87.23	63.27	84.58	81.69	94.62	58.29	90.63	81.01
half-slice	baseline	76.24	24.66	86.26	57.61	79.32	76.55	94.34	54.04	86.20	75.57
	AdaWAC	78.14	29.75	86.66	62.28	81.36	78.84	94.60	57.95	85.38	78.01
half-vol	baseline	72.65	35.86	83.29	43.70	78.25	77.25	92.92	51.32	83.80	70.66
	AdaWAC	75.93	34.95	84.45	60.40	79.59	76.06	93.19	54.46	84.91	74.37
half-sparse	baseline	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	AdaWAC	39.68	80.93	76.59	0.00	66.53	62.11	49.69	31.09	12.30	19.11

[14]Note that the numbers of correct metrics and metrics in TransUNet (Chen et al., 2021) on ACDC dataset are the same.

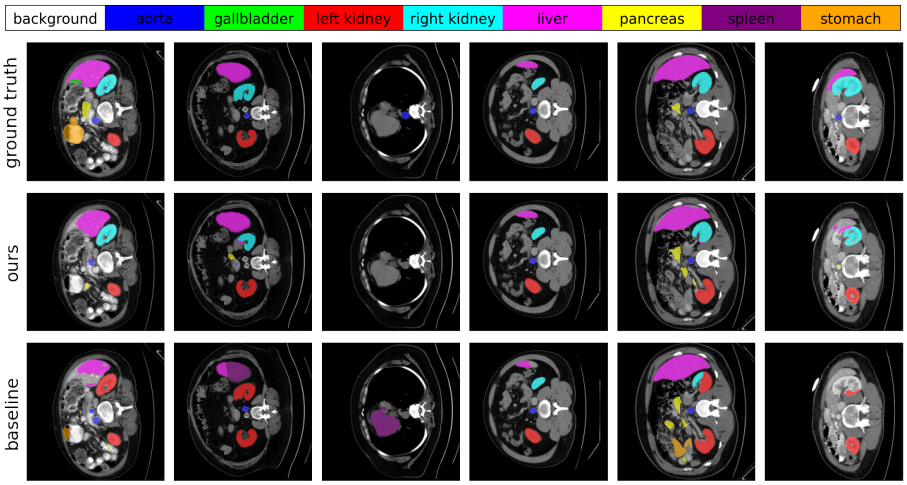

Figure 5: Visualization of segmentation predictions made by models trained via *AdaWAC* (ours) and via the baseline (ERM+SGD) with TransUNet (Chen et al., 2021) on the **half-sparse** subset of the Synapse training set. Top to bottom: ground truth, ours (*AdaWAC*), baseline.

Table 7: AdaWAC versus hard-thresholding algorithms with TransUNet on Synapse, measured by metrics in TransUNet (Chen et al., 2021).

Method	DSC ↑	HD95↓	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
baseline	77.32	29.23	87.46	63.54	82.06	77.76	94.10	54.06	85.07	74.54
trim-train	77.05	26.94	86.70	60.65	80.02	76.64	94.25	54.20	86.44	77.49
trim-ratio	75.30	28.59	87.35	57.29	78.70	72.22	94.18	52.32	86.31	74.03
trim-train+ACR	76.70	35.06	87.11	62.22	74.19	75.25	92.19	57.16	88.21	77.30
trim-ratio+ACR	79.02	33.59	86.82	61.67	83.52	81.22	94.07	59.06	88.08	77.71
AdaWAC (ours)	80.16	25.79	87.23	63.27	84.58	81.69	94.62	58.29	90.63	81.01

Table 8: Ablation study of AdaWAC with TransUNet trained on Synapse, measured by metrics in TransUNet (Chen et al., 2021).

Method	DSC ↑	HD95↓	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
baseline	77.32	29.23	87.46	63.54	82.06	77.76	94.10	54.06	85.07	74.54
reweight-only	77.72	29.24	86.15	62.31	82.96	80.28	93.42	55.86	85.29	75.49
ACR-only	78.93	31.65	87.96	62.67	81.79	80.21	94.52	60.41	88.07	75.83
AdaWAC-0.01	78.98	27.81	87.58	61.09	82.29	80.22	94.90	55.92	91.63	78.23
AdaWAC-1.0	80.16	25.79	87.23	63.27	84.58	81.69	94.62	58.29	90.63	81.01