Adversarially Robust Hypothesis Transfer Learning

Yunjuan Wang [1] **Raman Arora** [1]

Abstract

In this work, we explore Hypothesis Transfer Learning (HTL) under adversarial attacks. In this setting, a learner has access to a training dataset of size n from an underlying distribution $\mathcal{D}$ and a set of auxiliary hypotheses. These auxiliary hypotheses, which can be viewed as prior information originating either from expert knowledge or as pre-trained foundation models, are employed as an initialization for the learning process. Our goal is to develop an adversarially robust model for $\mathcal{D}$.

We begin by examining an adversarial variant of the regularized empirical risk minimization learning rule that we term A-RERM. Assuming a non-negative smooth loss function with a strongly convex regularizer, we establish a bound on the robust generalization error of the hypothesis returned by A-RERM in terms of the robust empirical loss and the quality of the initialization. If the initialization is good, i.e., there exists a weighted combination of auxiliary hypotheses with a small robust population loss, the bound exhibits a fast rate of $\mathcal{O}(1/n)$. Otherwise, we get the standard rate of $\mathcal{O}(1/\sqrt{n})$. Additionally, we provide a bound on the robust excess risk which is similar in nature, albeit with a slightly worse rate.

We also consider solving the problem using a practical variant, namely proximal stochastic adversarial training, and present a bound that depends on the initialization. This bound has the same dependence on the sample size as the ARERM bound, except for an additional term that depends on the size of the adversarial perturbation.

[1]Department of Computer Science, Johns Hopkins University, Baltimore, USA. Correspondence to: Yunjuan Wang <ywang509@jhu.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

Despite the incredible success of machine learning on real-world problems and its widespread adoption, several studies over the years have shown that models trained using machine learning can be highly susceptible to adversarial attacks (Goodfellow et al., 2014; Kurakin et al., 2018). These attacks involve intentionally designing imperceptible perturbations of the input data that cause the deployed (trained) model to predict unreliably. A popular defense against such inference-time attacks is adversarial training wherein the learner is presented with simulated adversarial corruptions of clean training data. Empirical studies have consistently demonstrated that the use of adversarial training (Madry et al., 2018) and its variants (Cai et al., 2018; Zhang et al., 2019; Wang et al., 2020) result in models that exhibit greater resilience to perturbations in the input space.

However, we rarely train models from scratch in real-world scenarios, irrespective of whether we use adversarial training or standard training. One of the primary reasons is the substantial increase in the size of available training data – training from scratch demands not only the storage of copious amounts of data but also significant computational expense (e.g., parameter tuning). It may also be the case that the underlying data distribution is not aligned with the distribution of training data. Finally, in many scenarios, training data might not be accessible due to privacy concerns.

A compelling solution in such large-scale, real-world settings is to consider transferring knowledge from a source domain to a target domain using an auxiliary set of hypotheses; in prior work, this is referred to as hypothesis transfer learning (HTL) (Kuzborskij & Orabona, 2013; 2017; Du et al., 2017; Aghbalou & Staerman, 2023). These auxiliary hypotheses can be viewed as prior information, originating either from expert knowledge or as pre-trained foundation models (trained on various related source tasks), and are employed as an initialization for the learning process. Given a hypothesis class, $\mathcal{H}$, we linearly combine any candidate predictor $h_{\mathrm{w}} \in \mathcal{H}$ with a weighted combination of the auxiliary hypotheses, $f_1^{\mathrm{aux}}, \ldots, f_k^{\mathrm{aux}}$, to construct the following model for the target task:

$$h_{\mathrm{w},\beta}(\cdot) := h_{\mathrm{w}}(\cdot) + \sum_{j=1}^{k} \beta_j f_j^{\mathrm{aux}}(\cdot),$$

where $\beta = [\beta_1, \ldots, \beta_k] \in \mathbb{R}^k$ can be interpreted as the *effectiveness* of the j^{th} auxiliary hypothesis towards solving the target task. Typically, we consider $\mathcal{H}$ to be a simple hypothesis class (e.g., linear predictors) or a reproducing kernel Hilbert space (RKHS), whereas the auxiliary hypotheses f_j^{aux} are fairly complex (e.g., deep neural networks). We train the weight parameters while keeping auxiliary hypotheses fixed during the training process.

The HTL setup is related, yet distinct, from popular frameworks of transfer learning and domain adaptation, where the learner has access to data from one or more source domains. Instead, in HTL, we assume that all the knowledge from the source domains has been distilled into a set of auxiliary hypotheses presented to the learner as side information. Therefore, The HTL framework is also an excellent model for studying phenomena such as fine-tuning. Indeed, with the advent of foundation models such as the Vision Transformer (ViT) (Dosovitskiy et al., 2020) and large language models (LLM) (Floridi & Chiriatti, 2020), we can view such pre-trained models as auxiliary hypotheses. This paradigm shift not only offers efficiency but also provides flexibility in adapting models to diverse tasks by building upon the foundational knowledge embedded in pre-trained models.

In this work, we study the theoretical aspects of hypothesis transfer learning while ensuring adversarial robustness. We make the following contributions.

In Section 3, we begin by exploring Adversarial Regularized Empirical Risk Minimization framework (A-RERM) for non-negative smooth loss functions with a strongly convex regularizer. We establish a data-dependent bound on robust generalization error of A-RERM that depends on the *utility* of the auxilliary hypotheses. In particular, assuming that there exists a hypothesis in the RKHS, $\mathcal{H}$, which in conjunction with a linear combination of auxiliary hypotheses can achieve a small robust loss on the target task, then the generalization error of A-RERM converges at a fast rate of $\mathcal{O}(1/n)$. Otherwise, the error decays at the standard rate of $\mathcal{O}(1/\sqrt{n})$. We also give an optimistic bound on excess robust risk with an $O(1/n)$ rate, but a worse rate of $\mathcal{O}(1/n^{1/4})$ when near robust-realizability does not hold.

In Section 4, we explore a practical algorithm based on proximal stochastic gradient descent algorithm. We show a bound of $\mathcal{O}(\frac{1}{n} + \alpha)$ on the generalization gap (w.r.t. the robust loss). Unlike prior work, nowhere in our analysis we assume convexity.

1.1. Related Work

Hypothesis Transfer Learning (HTL). In an early work, Kuzborskij & Orabona (2013) analyze the generalization ability of hypothesis transfer learning by leveraging the stability of regularized least squares regression. Their framework was subsequently extended to a metric learning setting by Perrot & Habrard (2015) wherein the auxiliary hypothesis is a PSD matrix defining a (Mahalanobis) distance on input features. For smooth non-negative loss functions, Kuzborskij & Orabona (2017) give an optimistic guarantee that exhibits a fast rate if the auxilliary hypotheses prove beneficial for the (target) task. Du et al. (2017) establish a similar fast rate for kernel smoothing and kernel ridge regression for the setting when the source and target tasks are related by a transformation. More recently, Aghbalou & Staerman (2023) study HTL for surrogate losses for the binary classification problem.

Robust Generalization Guarantees. Several works give generalization guarantees for adversarially robust empirical risk minimization using uniform convergence, i.e., by bounding the difference between the expected and the empirical errors on an i.i.d. sample, simultaneously for all hypotheses in the hypothesis class. These yield guarantees based on various complexity measures of the hypothesis class, including Rademacher complexity (Yin et al., 2019; Khim & Loh, 2018; Awasthi et al., 2020), VC dimension (Cullina et al., 2018; Montasser et al., 2020), the covering number (Balda et al., 2019; Mustafa et al., 2022; Li & Telgarsky, 2023), or utilizing PAC Bayesian analysis (Viallard et al., 2021; Xiao et al., 2023) or margin theory (Farnia et al., 2018). Another line of work focuses on analyzing robust generalization guarantees of adversarial training (Madry et al., 2018) for linear predictors (Zou et al., 2021) or shallow neural networks (Allen-Zhu & Li, 2022; Mianjy & Arora, 2023; Li & Telgarsky, 2023; Wang et al., 2024), albeit under somewhat restrictive distributional assumptions.

Algorithmic Stability Analysis. The stability-based analysis, introduced by Bousquet & Elisseeff (2002), offers an alternative approach for obtaining generalization bounds in scenarios where uniform convergence-based guarantees prove inadequate. Significant recent breakthroughs from Feldman & Vondrak (2018; 2019); Bousquet et al. (2020); Klochkov & Zhivotovskiy (2021) strengthen the nature of these guarantees by improving high-probability bounds for uniformly-stable learning algorithms. Relatedly, Hardt et al. (2016) provide stability-based analysis of stochastic gradient descent (SGD) for stochastic convex optimization with smooth loss functions. Kuzborskij & Lampert (2018) introduce a data-dependent notion of algorithmic stability to give novel generalization bounds. More recently, Zhang et al. (2022) show that the stability analysis of Hardt et al. (2016) is tight for convex and strongly convex functions while improving upon the results of Hardt et al. (2016) for non-convex loss functions and of Kuzborskij & Lampert (2018) to give a tighter bound for the data-dependent average stability of SGD for non-convex smooth loss functions. Complementing these advances, the smoothness assumption

in the stability analysis of SGD is relaxed in Lei & Ying (2020) to loss functions with Hölder continuous subgradients and Bassily et al. (2020) extend the analysis to nonsmooth loss functions. Zhou et al. (2022) characterize the stability of SGD for non-convex smooth functions in terms of on-average variance of the stochastic gradients and Lei (2023) extends the analysis to weakly convex problems with non-convex and nonsmooth objective functions.

Compared to the standard setting, there has been limited work applying stability-based analysis to study the generalization gap in adversarial learning. While Xing et al. (2021) examine the algorithmic stability of a generic adversarial training algorithm by leveraging the non-smooth nature of adversarial loss, Xiao et al. (2022b) introduce a notion of approximate smoothness to characterize adversarial loss. However, all prior works assume that the adversarial loss function is convex, which is unrealistic. In this work, we forgo such unrealistic assumptions and focus on the general setting where the adversarial loss is non-convex.

2. Problem Setup

Notation. Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y}$ denote the input feature space and the output label space, respectively; for regression, $\mathcal{Y} \subseteq [-1, 1]$ and for binary classification, $\mathcal{Y} = \{\pm 1\}$. Write $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Let $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{Y}\}$ denote a hypothesis class parameterized by some vector space $\mathcal{W}$. The uncertain relationship between inputs and outputs is modeled using an (unknown) distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. We are given a training data $\mathcal{S} = \{\mathrm{z}_i := (\mathrm{x}_i, y_i)\}_{i=1}^n$, a sample of size n drawn i.i.d. from $\mathcal{D}$. Let $\ell : \mathcal{H} \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}_+$ denote the loss function. Given a hypothesis h_{w} parameterized by $\mathrm{w} \in \mathcal{W}$ and a random example $\mathrm{z} = (\mathrm{x}, y) \sim \mathcal{D}$, the loss function can be written as $\ell(h_{\mathrm{w}}, (\mathrm{x}, y)) = \phi(y h_{\mathrm{w}}(\mathrm{x}))$, where $\phi : \mathbb{R} \to \mathbb{R}_+$. We also write $\ell(h_{\mathrm{w}}, (\mathrm{x}, y))$ as $\ell(\mathrm{w}, \mathrm{z})$. Define the population and empirical loss, respectively, as $L(h) := \mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}}[\ell(h, (\mathrm{x}, y))]$ and $\widehat{L}(h) := \frac{1}{n}\sum_{i=1}^n [\ell(h, (\mathrm{x}_i, y_i))]$.

We make the following assumptions on the loss function.

Assumption 1. For all $\mathrm{z} \sim \mathcal{D}$, we assume that the loss function satisfies the following: (1) $\ell(\cdot, \mathrm{z})$ is continuously differentiable; (2) $\ell(\cdot, \mathrm{z})$ is non-negative, monotonically decreasing, and $|\ell(\cdot, \mathrm{z})|$ is uniformly bounded by M; (3) $\phi(\cdot)$ is H-smooth.

Adversarial Attacks. We consider ℓ_p-norm-bounded adversarial attacks with a perturbation budget of $\alpha > 0$. For an input example $\mathrm{x} \in \mathcal{X}$, the set of all such perturbations is an ℓ_p-ball of size α centered at x, i.e., $\mathcal{B}_p(\mathrm{x}, \alpha) \subseteq \mathcal{X}$. Given the threat model, it is natural to consider the following robust (or, adversarial) loss function:

$$\tilde{\ell}^{\alpha}(h, (\mathrm{x}, y)) := \sup_{\tilde{\mathrm{x}} \in \mathcal{B}_p(\mathrm{x}, \alpha)} \ell(h, (\tilde{\mathrm{x}}, y)).$$

We refer to the population and empirical loss w.r.t. the robust loss as robust population loss and robust empirical loss, respectively:

$$L_{\text{adv}}^{\alpha}(h) := \mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}}\left[\sup_{\tilde{\mathrm{x}} \in \mathcal{B}_p(\mathrm{x}, \alpha)} \ell(h, (\tilde{\mathrm{x}}, y))\right],$$

$$\widehat{L}_{\text{adv}}^{\alpha}(h) := \frac{1}{n}\sum_{i=1}^n \sup_{\tilde{\mathrm{x}}_i \in \mathcal{B}_p(\mathrm{x}_i, \alpha)} \ell(h, (\tilde{\mathrm{x}}_i, y_i)).$$

To simplify notation, we often suppress the superscript α in $L_{\text{adv}}^{\alpha}, \widehat{L}_{\text{adv}}^{\alpha}, \tilde{\ell}^{\alpha}$. We emphasize that, unlike some prior works, we do not assume that the robust loss function $\tilde{\ell}$ is convex.

2.1. Robust Transfer from Auxiliary Hypotheses

In this setup, the learner has access to a set of models or hypotheses $\mathcal{F}^{\text{aux}} = \left\{f_j^{\text{aux}} : \mathcal{X} \to \mathcal{Y}\right\}_{j=1}^k$. These hypotheses, serving as prior information, are provided by experts with express domain knowledge or as a result of pre-training on source distributions possibly distinct from the underlying (target) distribution $\mathcal{D}$. The learner's ultimate goal is to identify a classifier h that has a small robust population loss, i.e., find $h^* = \operatorname{argmin}_{h\in\mathcal{H}} L_{\text{adv}}(h)$. The learner incorporates the auxiliary prior information $\mathcal{F}^{\text{aux}}$ by augmenting its hypothesis class and considering a combination classifier, denoted $h_{\mathrm{w},\beta}$, of the following form:

$$h_{\mathrm{w},\beta}(\cdot) := h_{\mathrm{w}}(\cdot) + f_{\beta}^{\text{aux}}(\cdot), \text{ with } f_{\beta}^{\text{aux}}(\cdot) = \sum_{j=1}^k \beta_j f_j^{\text{aux}}(\cdot).$$

Here, the weight β_j is a parameter that encodes the relevance of the j^{th} auxiliary hypothesis for the target task. While in practice, β would be learned on training data for the target task, for simplicity, we assume that β is fixed throughout the training process. Nonetheless, it offers a form of capacity control as we combine different, potentially very complex, hypotheses, e.g., deep neural networks. Naturally, we find that a bound on the size of β offers a useful tradeoff between the ability of the auxiliary hypotheses to fit the training data versus generalizing to the unseen data. We use $\Psi : \mathbb{R}^k \to \mathbb{R}_+$ to measure the size of weights on the auxiliary hypotheses.

Since f_j^{aux} are kept fixed throughout the training process, we can also interpret f_{β}^{aux} as a (warm) initialization for the learning algorithm which then refines the initial model akin to fine-tuning, albeit by additively incorporating a simple hypothesis h_{w} from $\mathcal{H}$ into the model. The expanded hypothesis set, which we denote as $\tilde{\mathcal{H}}$, is essentially the direct sum $\tilde{\mathcal{H}} = \mathcal{H} \bigoplus \text{span}\{f_1^{\text{aux}}, \cdots, f_k^{\text{aux}}\}$. Formally,

$$\tilde{\mathcal{H}} = \{\tilde{h} = h + \sum_{j=1}^k \beta_j f_j^{\text{aux}} \mid h \in \mathcal{H}, \beta_j \in \mathbb{R}, j = 1, \ldots, k\}.$$

We formulate learning as solving an Adversarial Regularized Empirical Risk Minimization (A-RERM) problem.

Given a regularizer function $\Omega : \mathcal{H} \mapsto \mathbb{R}_+$ and parameter $\lambda \in \mathbb{R}_+$, let $\mathcal{A}_{\text{A-RERM}} : \mathcal{Z}^n \times \mathcal{F}^{\text{aux}} \to \tilde{\mathcal{H}}$ denote the learning algorithm that given an i.i.d. sample $\mathcal{S} \sim \mathcal{D}^n$, returns $h_{\widehat{\mathrm{w}},\beta} = h_{\widehat{\mathrm{w}}} + f_\beta^{\text{aux}}$, where

$$h_{\widehat{\mathrm{w}}} = \operatorname*{argmin}_{h \in \mathcal{H}} \left(\widehat{L}_{\text{adv}}(h + f_\beta^{\text{aux}}) + \lambda \Omega(h) \right). \tag{1}$$

Note that the weights $\beta \in \mathbb{R}^k$ are fixed and we only optimize over $h \in \mathcal{H}$. For suitable choices of the regularizer, we can argue that the larger the regularization parameter λ, the closer the final model $h_{\widehat{\mathrm{w}},\beta}$ to f_β^{aux}. More concretely, we can make the following connection between a special case of Problem (1) and ERM with a biased regularizer. To that end, consider the setting where the hypothesis class is that of linear predictors, i.e., $h_{\mathrm{w}}(\mathrm{x}) = \mathrm{w}^\top \mathrm{x}$, the auxiliary model $f_\beta^{\text{aux}}(\mathrm{x}) = \mathrm{u}^\top \mathrm{x}$ is linear, and the regularizer $\Omega(\cdot) = \|\cdot\|^2$. Then, for squared loss $\phi(z) = (1-z)^2$, write the optimization Problem (1) as

$$\begin{aligned}\widehat{\mathrm{w}} &= \operatorname*{argmin}_{\mathrm{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \max_{\tilde{\mathrm{x}}_i \in \mathcal{B}_p(\mathrm{x}_i, \alpha)} \left(\mathrm{w}^\top \tilde{\mathrm{x}}_i + \mathrm{u}^\top \tilde{\mathrm{x}}_i - y_i \right)^2 + \lambda \|\mathrm{w}\|^2 \\ &= \operatorname*{argmin}_{\mathrm{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \left((\mathrm{w}+\mathrm{u})^\top (\mathrm{x}_i + \varepsilon_i) - y_i \right)^2 + \lambda \|\mathrm{w}\|_2^2, \end{aligned} \tag{2}$$

where $\varepsilon_{ij} = -\alpha y_i \text{sign}(w_j + u_j) \, |w_j + u_j|^{q-1} / \|\mathrm{w}+\mathrm{u}\|_q^{q-1}$ is the j-th component of the optimal adversarial perturbation ε_i of the training example x_i given the combined model $\mathrm{w} + \mathrm{u}$; note that q is the Hölder conjugate to p, i.e., $1/p + 1/q = 1$. Replace $\mathrm{w}' := \widehat{\mathrm{w}} + \mathrm{u}$, and let ε_i' denote the optimal adversarial perturbation of x_i given the model w'. Then, $\varepsilon_{ij}' = -\alpha y_i \text{sign}(w_j') \left| w_j' \right|^{q-1} / \|\mathrm{w}'\|_q^{q-1}$, and Problem (2) is equivalent to

$$\begin{aligned}\widehat{\mathrm{w}}' &= \operatorname*{argmin}_{\mathrm{w}' \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \left(\mathrm{w}'^\top (\mathrm{x}_i + \varepsilon_i') - y_i \right)^2 + \lambda \|\mathrm{w}' - \mathrm{u}\|^2 \\ &= \operatorname*{argmin}_{\mathrm{w}' \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \max_{\tilde{\mathrm{x}}_i \in \mathcal{B}_p(\mathrm{x}_i, \alpha)} \left(\mathrm{w}'^\top \tilde{\mathrm{x}}_i - y_i \right)^2 + \lambda \|\mathrm{w}' - \mathrm{u}\|^2. \end{aligned}$$

In the standard (non-robust) setting, several works study biased regularization for both transfer learning (Tommasi & Caputo, 2009; Tommasi et al., 2012; Balcan et al., 2019; Denevi et al., 2019; Takada & Fujisawa, 2020; Denevi et al., 2020) as well as hypothesis transfer learning Kuzborskij & Orabona (2013; 2017).

2.2. Algorithmic Stability

Following Kuzborskij (2018), we consider the following two notions of algorithmic stability.

Definition (On-Average Stability)**.** Given training data $\mathcal{S} = \{\mathrm{z}_i\}_{i=1}^n \sim \mathcal{D}^n$, let $\mathcal{S}^{(i)}$ denote a copy of $\mathcal{S}$ with the i-th example replaced by $\mathrm{z} \sim \mathcal{D}$, where $i \sim \text{Uniform}[n]$ is sampled according to uniform distribution over $\{1, \ldots, n\}$. We say that

1. An algorithm $\mathcal{A}$ is μ_1-on-average stable with respect to loss $\ell(\cdot)$ if the following holds:

$$\sup_{\mathrm{z}'} \mathbb{E}_{\mathcal{S},\mathrm{z},i} \left[\ell(\mathcal{A}(\mathcal{S}), \mathrm{z}') - \ell(\mathcal{A}(\mathcal{S}^{(i)}), \mathrm{z}') \right] \le \mu_1,$$

2. An algorithm $\mathcal{A}$ is μ_2-second-order-on-average stable with respect to loss $\ell(\cdot)$ if the following holds:

$$\sup_{\mathrm{z}'} \mathbb{E}_{\mathcal{S},\mathrm{z},i} \left[\left(\ell(\mathcal{A}(\mathcal{S}), \mathrm{z}') - \ell(\mathcal{A}(\mathcal{S}^{(i)}), \mathrm{z}') \right)^2 \right] \le \mu_2.$$

While the notion of on-average-stability (the first notion above) is milder than uniform stability (Bousquet & Elisseeff, 2002), it is slightly stronger than on-average-replace-one stability (Shalev-Shwartz et al., 2010). The following result bounds the generalization gap of an algorithm in terms of its on-average-stability parameters.

Theorem 2.1 (Theorem 8 of Kuzborskij (2018))**.** Let Algorithm $\mathcal{A}$ be μ_1-on-average-stable and μ_2-second-order-on-average stable. Let $\delta > 0$. Then, given a training set $\mathcal{S} \sim \mathcal{D}^n$ of size n, we have the following for the hypothesis $\mathcal{A}(\mathcal{S})$, with probability at least $1 - \delta$

$$L(\mathcal{A}(\mathcal{S})) - \widehat{L}(\mathcal{A}(\mathcal{S})) \le \mu_1 + \sqrt{4 \log\left(\frac{1}{\delta}\right) \mu_2} + \frac{3M \log\left(\frac{1}{\delta}\right)}{2n}.$$

3. Robust Generalization Guarantees

In this section, we first consider $\mathcal{H}$ to be a reproducing kernel Hilbert space (RKHS) endowed with a symmetric positive semidefinite kernel function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, an inner product $\langle \cdot, \cdot \rangle$ and a norm $\|\cdot\|_k$. For any $\mathrm{x} \in \mathbb{R}^d$, the function $\mathrm{x} \mapsto k(\mathrm{x}, \cdot)$ is contained in $\mathcal{H}$. We study Problem (1) with $\Omega(h) = \|h\|_k^2$. We assume that $\Omega(\cdot)$ is strongly convex w.r.t. $\|\cdot\|_k$ and that the kernel and the auxiliary hypotheses are all bounded.

Assumption 2. We make the following boundedness assumptions on the hypotheses.

(1) Auxiliary hypotheses in $\mathcal{F}^{\text{aux}}$ are bounded point-wise by C, i.e., $\sup_{j \in [k], \mathrm{x} \in \mathcal{X}} \left| f_j^{\text{aux}}(\mathrm{x}) \right| = C < \infty$.

(2) Hypotheses in the RKHS $\mathcal{H}$ are bounded, i.e., the kernel k is bounded by $\kappa \in \mathbb{R}$: $\sup_{\mathrm{x}_1, \mathrm{x}_2 \in \mathcal{X}} k(\mathrm{x}_1, \mathrm{x}_2) = \kappa < \infty$.

Further, we assume that the regularizers Ω, Ψ are strongly convex w.r.t. corresponding norms.

(3) $\Omega(\cdot) = \|\cdot\|_k^2$ is σ-strongly convex w.r.t. RKHS norm.

(4) $\Psi(\cdot)$ is 1-strongly convex w.r.t. the Euclidean norm.

3.1. Bounding Robust Generalization Gap

First, we provide an upper bound on the robust generalization gap for A-RERM.

Theorem 3.1. Assume that the learner is given a weighted linear combination $f_\beta^{\text{aux}}(\cdot) = \sum_{j=1}^k \beta_j f_j^{\text{aux}}(\cdot)$ of auxiliary hypotheses with weights $\beta \in \mathbb{R}^k$ such that $\Psi(\beta) \leq \rho$. Fix any $\delta > 0$, and say Assumptions 1 and 2 hold. Then, given an i.i.d. sample $\mathcal{S} \sim \mathcal{D}^n$ of size n, for any $\lambda > 0$, the A-RERM rule returns $h_{\widehat{\mathrm{w}},\beta}$ such that with probability at least $1-\delta$,

$$\begin{aligned}&\left|L_{\text{adv}}(h_{\widehat{\mathrm{w}},\beta}) - \widehat{L}_{\text{adv}}(h_{\widehat{\mathrm{w}},\beta})\right| \\ &\leq \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}\left(\left(1+\frac{1}{\sqrt{\lambda}}\right)\left(\frac{L_{\text{adv}}^{\text{aux}}}{\sqrt{\lambda}} + \sqrt{L_{\text{adv}}^{\text{aux}}\rho}\right) + \sqrt{L_{\text{adv}}^{\text{aux}}}\right)\right) \\ &\quad + \tilde{\mathcal{O}}\left(\frac{1}{n}\left(\left(1+\frac{1}{\sqrt{\lambda}}\right)\left(\sqrt{\frac{L_{\text{adv}}^{\text{aux}}}{\lambda}} + \sqrt{\rho}\right)\right)\right),\end{aligned}$$

where $L_{\text{adv}}^{\text{aux}} = L_{\text{adv}}^{\alpha}(f_\beta^{\text{aux}})$.

Some remarks are in order.

In the bound above, we use the $\tilde{O}(\cdot)$ notation to hide $\log(1/\delta)$ terms as well as dependence on constants H, κ, C, σ, M; please see Appendix A for a detailed statement and proof.

Not surprisingly, the bound in Theorem 3.1 depends on the robust error of the auxiliary model f_β^{aux}. Indeed, our bound is optimistic in nature. For settings where $L_{\text{adv}}^{\text{aux}} \approx 0$ is small or in small sample regimes where $n = \mathcal{O}(1/L_{\text{adv}}^{\text{aux}})$, the bound above decays at a fast rate of $\mathcal{O}(1/n)$.

If we view the auxiliary model as a warm initialization and A-RERM as performing fine-tuning, then our result is an affirmation of the empirical finding of Hua et al. (2023) that initialization is important for adversarial transfer learning. We can also view $L_{\text{adv}}^{\text{aux}}$ as characterizing transferability to the new domain, playing a role similar to domain divergence in transfer learning (Ben-David et al., 2010) and robust domain adaptation (Deng et al., 2023).

We remark that our proof of Theorem 3.1 allows for general adversarial attacks, wherein the set of perturbations for any input example can be arbitrary, including a discrete set of large-norm perturbations or (image) transformations.

The robust loss of the auxiliary model f_β^{aux} depends on the weights β. While Theorem 3.1 holds for any $\beta \in \mathbb{R}^k$, it is fixed prior to learning. In practice, we can treat β as a hyperparameter and use cross-validation to pick a good model. An alternate approach would be to optimize over h and β simultaneously and consider the following learning rule instead,

$$(h_{\widehat{\mathrm{w}}}, \widehat{\beta}) = \operatorname*{argmin}_{h\in\mathcal{H},\beta\in\mathbb{R}^k} \left(\widehat{L}_{\text{adv}}(h + f_\beta^{\text{aux}}) + \lambda\Omega(h) + \nu\Psi(\beta)\right).$$

Two-layer ReLU Networks Next, we extend our result to two-layer ReLU networks of width m. A hypothesis $h_{\mathrm{w}} \in \mathcal{H}$ is parametrized using top-layer weights $\mathrm{a} = (a_1, \ldots, a_m) \in \mathbb{R}^m$, bottom layer weights $\mathrm{w}_s \in \mathcal{W}$, into each of the hidden neurons, $s \in \{1, \ldots, m\}$, with $\|\mathrm{w}_s\| \leq B$. For any input $\mathrm{x} \in \mathcal{X} := \{\mathrm{x} \in \mathbb{R}^d \mid \|\mathrm{x}\| \leq \kappa\}$, the output of the model $h_w(\cdot)$ is given as $h_{\mathrm{W}}(\mathrm{x}) = \sum_{s=1}^m a_s \psi(\langle \mathrm{w}_s, \mathrm{x}\rangle)$, where $\psi(\mathrm{x}) = \max(0, \mathrm{x})$ is the ReLU activation function. The top-layer weights are set by sampling them uniformly $a_s \sim \text{Unif}\left\{\pm\frac{1}{m}\right\}$ and are kept fixed while training $\{\mathrm{w}_s\}_{s=1}^m$. Slightly abusing the notation, we write $\Omega(h_{\mathrm{W}}) = \sum_{s=1}^m \Omega(\mathrm{w}_s)$. We assume that Ω is σ-strongly convex w.r.t. $\|\cdot\|$ and that $\Omega(0) = 0$. As before, we assume $\Psi(\cdot)$ is 1-strongly convex w.r.t. the Euclidean norm. With the setup above we obtain that Theorem 3.1 holds for two-layer ReLU networks as well.

3.2. Excess Robust Risk

The bound in the previous section is post hoc in nature, stating that if we find a model with a small training loss, then it will generalize well. Here, we give a bound on the excess robust risk that holds a priori, i.e., before the learner even sees any training data.

Theorem 3.2. Assume that the learner is given a weighted linear combination $f_\beta^{\text{aux}}(\cdot) = \sum_{j=1}^k \beta_j f_j^{\text{aux}}(\cdot)$ of auxiliary hypotheses with weights $\beta \in \mathbb{R}^k$ such that $\Psi(\beta) \leq \rho$. Fix any $\delta > 0$, and say Assumptions 1 and 2 hold. Let $\tau > 0$ be such that $\sup_{h\in\mathcal{H}} \Omega(h) \leq \tau$. Then, given an i.i.d. sample $\mathcal{S} \sim \mathcal{D}^n$ of size n, and setting λ as

$$\lambda = \tilde{\mathcal{O}}\left(\sqrt{\frac{1}{\tau}\sqrt{\frac{L_{\text{adv}}^{\text{aux}}}{n}}\left(\sqrt{L_{\text{adv}}^{\text{aux}}} + \sqrt{\rho}\right) + \sqrt{\frac{1}{\tau^2 n}\sqrt{\frac{L_{\text{adv}}^{\text{aux}}}{n}}\left(\sqrt{L_{\text{adv}}^{\text{aux}}} + \sqrt{\rho}\right)}}\right)$$

the A-RERM rule returns $h_{\widehat{\mathrm{w}},\beta}$ such that with probability at least $1-\delta$,

$$\begin{aligned}L_{\text{adv}}(h_{\widehat{\mathrm{w}},\beta}) - \min_{h_{\mathrm{w}}:\Omega(h_{\mathrm{w}})\leq\tau} L_{\text{adv}}(h_{\mathrm{w},\beta}) \leq \tilde{\mathcal{O}}\Bigg(&\frac{1}{n} + \frac{\sqrt{L_{\text{adv}}^{\text{aux}}}}{n^{1/2}} \\ &+ \frac{\sqrt{L_{\text{adv}}^{\text{aux}}} + \sqrt[4]{L_{\text{adv}}^{\text{aux}}\rho}}{n^{1/4}}\sqrt{\tau} + \frac{\sqrt[4]{L_{\text{adv}}^{\text{aux}}} + \sqrt[8]{L_{\text{adv}}^{\text{aux}}\rho}}{n^{3/8}}\sqrt{\tau}\Bigg).\end{aligned}$$

Akin to the bound in the previous section, the bound above is optimistic in nature. If $L_{\text{adv}}^{\text{aux}} \approx 0$, the excess robust risk decays as $\tilde{\mathcal{O}}(1/n)$. However, owing to the non-convexity of the adversarial loss function we obtain a worse rate of $\tilde{\mathcal{O}}(1/n^{1/4})$ in general. We note that both of our results (Theorem 3.1 and 3.2) recover the results in the standard (non-robust) setting (Kuzborskij & Orabona, 2017) for $\alpha = 0$.

4. Robust Generalization Bounds via Proximal Stochastic Adversarial Training

Thus far, we focused on the A-RERM learning rule for adversarial hypothesis transfer. While the A-RERM rule exhibits an optimistic statistical learning rate, it is often computationally hard to implement even in a standard setting without robustness constraints. This motivates a more practical approach to learning based on proximal SGD which we refer to as Proximal Stochastic Adversarial Training (PSAT). Proximal algorithms are standard tool for solving nonsmooth convex optimization problems; see Schmidt et al. (2011); Xiao & Zhang (2014); Ghadimi et al. (2016); Asi & Duchi (2019) for a good introduction.

The setup is the same as in the previous section. We are given training data $\mathcal{S} = \{\mathrm{z}_i = (\mathrm{x}_i, y_i)\}_{i=1}^n \sim \mathcal{D}^n$, an adversarial loss function $\tilde{\ell}(\cdot) : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$, and a hypothesis class parameterized by $\mathrm{w} \in \mathcal{W}$. We consider a possibly nonsmooth regularizer function $\Omega(\cdot)$. Then, at each round t of the PSAT algorithm, we sample an example uniformly randomly from the given dataset $\mathcal{S}$, i.e., sample ξ_t uniformly over $[n]$ without replacement, and perform the following update:

$$\mathrm{w}_{t+1} = \mathrm{prox}_{\gamma_t,\lambda\Omega}\left(\mathrm{w}_t - \gamma_t \nabla\tilde{\ell}(\mathrm{w}_t, \mathrm{z}_{\xi_t})\right),$$

where the proximal map with parameter $\gamma > 0$ is defined as:

$$\mathrm{prox}_{\gamma,\Omega}(\mathrm{w}) := \operatorname*{argmin}_{\mathrm{u}} \Omega(\mathrm{u}) + \frac{1}{2\gamma}\|\mathrm{u} - \mathrm{w}\|^2.$$

We initialize PSAT with the auxiliary model, $\mathrm{w}_0 = f_\beta^{\mathrm{aux}}$. For any $\lambda > 0$, we define the regularized adversarial population and empirical loss, respectively, as follows

$$\Phi_{\mathrm{adv}}(\mathrm{w}) := L_{\mathrm{adv}}(\mathrm{w}) + \lambda\Omega(\mathrm{w}), \quad \text{and}$$
$$\widehat{\Phi}_{\mathrm{adv}}(\mathrm{w}) := \widehat{L}_{\mathrm{adv}}(\mathrm{w}) + \lambda\Omega(\mathrm{w}).$$

For simplicity, we assume $\Omega(\cdot)$ to be 1-strongly convex function, not necessarily differentiable.

Since we work in a stochastic setting, we also assume that the variance of the stochastic gradients is bounded, an assumption that is rather standard in analysis of stochastic gradient-based algorithms for optimization.

Assumption 3. Given any sample $\mathcal{S} = \{\mathrm{z}_1, \ldots, \mathrm{z}_n\} \sim \mathcal{D}^n$, there exists a constant $\nu_{\mathcal{S}} > 0$ such that $\forall \mathrm{w} \in \mathcal{W}$, we have

$$\mathbb{E}_{\xi\sim\mathrm{Uniform}[n]}\left\|\nabla\tilde{\ell}(\mathrm{w};\mathrm{z}_\xi) - \frac{1}{n}\sum_{i=1}^n \nabla\tilde{\ell}(\mathrm{w};\mathrm{z}_i)\right\|^2 \le \nu_{\mathcal{S}}^2.$$

Assumption 4. We assume that the loss function is Lipschitz and satisfies certain smoothness conditions:

(1) $\|\ell(\mathrm{w}_1,\mathrm{z}) - \ell(\mathrm{w}_2,\mathrm{z})\| \le L\|\mathrm{w}_1 - \mathrm{w}_2\|$,
(2) $\|\nabla_{\mathrm{w}}\ell(\mathrm{w}_1,\mathrm{z}) - \nabla_{\mathrm{w}}\ell(\mathrm{w}_2,\mathrm{z})\| \le H\|\mathrm{w}_1 - \mathrm{w}_2\|$,
(3) $\|\nabla_{\mathrm{w}}\ell(\mathrm{w},(\mathrm{x}_1,y)) - \nabla_{\mathrm{w}}\ell(\mathrm{w},(\mathrm{x}_2,y))\| \le H_z\|\mathrm{x}_1 - \mathrm{x}_2\|$.

The assumption above is mild and rather standard in several works studying adversarial training (Sinha et al., 2017; Wang et al., 2021; Farnia & Ozdaglar, 2021; Xing et al., 2021; Xiao et al., 2022a). Note that we do not assume that the loss function is convex.

Next, we establish generalization guarantees for PSAT by showing that it is a stable rule. First, we show that if PSAT is fed two datasets that are similar, then it produces models that are close to each other. Formally, given a training data $\mathcal{S} \sim \mathcal{D}^n$, let $\mathcal{S}^{(i)}$ denote the training data obtained by replacing the i-th example $\mathrm{z}^i \in \mathcal{S}$ by another example $\mathrm{z}' \sim \mathcal{D}$ drawn independently; we refer to $\mathcal{S}, \mathcal{S}^{(i)}$ as neighboring datasets.

Lemma 4.1. Say Assumptions 1, 3 and 4 hold. Let w_T and w'_T denote the outputs on two neighboring datasets $\mathcal{S}, \mathcal{S}'$, respectively, after running PSAT for T iterations on each of the datasets using $\gamma_t = \frac{c}{t+1}$ with $0 < c < \frac{1}{H}$. Then, for $\lambda > H$, we have that:

$$\mathbb{E}_{\xi\sim\mathrm{Uniform}[n],\mathcal{S},\mathcal{S}^{(i)}}\|\mathrm{w}_T - \mathrm{w}'_T\| \le \frac{4(1+\lambda)H_z\alpha}{\lambda - H} + \frac{2(1+\lambda)\sqrt{2H\Phi_{\mathrm{adv}}(\mathrm{w}_0) + (4\mathbb{E}_{\mathcal{S}}[\nu_{\mathcal{S}}^2] + 4H_z^2\alpha^2)\log(T)}}{n(\lambda - H)}.$$

Further, for $\lambda > 2H + 1$, we have

$$\mathbb{E}_{\xi\sim\mathrm{Uniform}[n],\mathcal{S},\mathcal{S}^{(i)}}\|\mathrm{w}_T - \mathrm{w}'_T\|^2 \le \frac{8(H+2)^2(1+\lambda)^2H_z^2\alpha^2}{(2\lambda - 3H - 2)HT} + \frac{(1+\lambda)^2\big(32H\Phi_{\mathrm{adv}}(\mathrm{w}_0) + \big(64\mathbb{E}_{\mathcal{S}}[\nu_{\mathcal{S}}^2] + 16H_z^2\alpha^2\big)\log(T)\big)}{(2\lambda - 3H - 2)HTn}.$$

Using Lemma 4.1 in Theorem 2.1 gives the following bound on the robust generalization error of PSAT.

Theorem 4.2. Say Assumptions 1, 3 and 4 hold. Let w_T denote the output on a sample $\mathcal{S} \sim \mathcal{D}^n$ of size n after running PSAT for T iterations with $\gamma_t = \frac{c}{t+1}$ for $0 < c < \frac{1}{H}$, and let $\lambda > 2H + 1$. Then, $\forall\delta > 0$, w.p. at least $1 - \delta$, we have that

$$L_{\mathrm{adv}}(\mathrm{w}_T) - \widehat{L}_{\mathrm{adv}}(\mathrm{w}'_T) \le \frac{1.5M\log(1/\delta)}{n} + \left(8\sqrt{\frac{2L^2(H\Phi_{\mathrm{adv}}(\mathrm{w}_0) + (4\mathbb{E}_{\mathcal{S}}[\nu_{\mathcal{S}}^2] + H_z^2\alpha^2)\log(T))\log(1/\delta)}{nTH(2\lambda - 3H - 2)}} + \frac{2L}{n(\lambda - H)}\sqrt{2H\Phi_{\mathrm{adv}}(\mathrm{w}_0) + (4\mathbb{E}_{\mathcal{S}}[\nu_{\mathcal{S}}^2] + 4H_z^2\alpha^2)\log(T)} + \frac{4LH_z\alpha}{\lambda - H} + \sqrt{\frac{32(H+2)^2L^2H_z^2\alpha^2\log(1/\delta)}{(2\lambda - 3H - 2)HT}}\right)(1+\lambda)$$

Ignoring the constants and higher order terms, to better understand the result above, we see that the bound scales as $\tilde{\mathcal{O}}(\sqrt{\Phi_{\mathrm{adv}}(\mathrm{w}_0) + \mathbb{E}_{\mathcal{S}}[\nu_{\mathcal{S}}^2]\log(T)}\left(\frac{1}{\sqrt{nT}} + \frac{1}{n}\right) + \frac{1}{n} + \alpha)$.

The results above hold for sufficiently large regularization parameter. Next, we present generalization bound for the setting when λ is relatively small.

Theorem 4.3. Say Assumptions 1, 3 and 4 hold. Let w_T denote the output on a sample $\mathcal{S} \sim \mathcal{D}^n$ of size n after

running PSAT for $T = O(n)$ iterations with $\gamma_t = \frac{c}{t+1}$ for $0 < c < \frac{1}{H}$, and $0 < \lambda < H$, we have that

$$\mathbb{E}_{\xi,\mathcal{S}}\Big[L_{\text{adv}}(\mathrm{w}_T) - \widehat{L}_{\text{adv}}(\mathrm{w}_T)\Big] \leq \mathcal{O}\left(\max\left\{\frac{T^q}{n^{q+1}}\frac{L\,(Q+nH_z\alpha)}{H(1-q)}, \frac{T^{\frac{q}{q+1}}}{n}\left(\mathbb{E}_{\xi,\mathcal{S}}\widehat{L}_{\text{adv}}(\mathrm{w}_T)\right)^{\frac{q}{q+1}}\left[\frac{L\,(Q+nH_z\alpha)}{H(1-q)}\right]^{\frac{1}{q+1}}\right\}\right),$$

where $Q = \sqrt{2H\Phi_{\text{adv}}(\mathrm{w}_0) + (4\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right] + 4H_z^2\alpha^2)\log(T)}$, $q = 1 - \frac{\lambda}{H}$.

The bound above depends on the initialization as well as the expected adversarial empirical loss; i.e. $\mathbb{E}_{\xi,\mathcal{S}}\widehat{L}_{\text{adv}}(\mathrm{w}_T)$. For settings where $\mathbb{E}_{\xi,\mathcal{S}}\widehat{L}_{\text{adv}}(\mathrm{w}_{T,\mathcal{S}}) = \mathcal{O}(\frac{T^q}{n^{1+q}})$, the bound scales as $\mathcal{O}\big(\frac{1}{1-q}\left(\frac{T}{n}\right)^q(\frac{Q}{n}+\alpha)\big)$. Again, setting $T = n$, it simplifies to $\mathcal{O}\big(\frac{1}{1-q}(\frac{Q}{n}+\alpha)\big)$.

5. Conclusion and Discussion

In this paper, we studied the problem of learning adversarially robust models using auxiliary hypotheses. Given a smooth loss and a strongly convex regularizer, we establish robust generalization guarantees for two learning algorithms – adversarial regularized empirical risk minimization (A-RERM) and proximal stochastic adversarial training (PSAT). Our results highlight the importance of a good initialization for achieving fast generalization. There are a several promising directions for future research. Our theoretical analysis highlights the importance of the regularization parameter in achieving fast generalization guarantees. It would be interesting to explore principled approaches such as recursive regularization for controlling the regularizer strength. Further, developing a practical algorithm that mitigates the dependence of the robust generalization gap on the perturbation size would help advance the state-of-the-art.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

This research was supported, in part, by DARPA GARD award HR00112020004 and NSF CAREER award IIS-1943251.

References

Aghbalou, A. and Staerman, G. Hypothesis transfer learning with surrogate classification losses: Generalization bounds through algorithmic stability. In *International Conference on Machine Learning*, pp. 280–303. PMLR, 2023.

Allen-Zhu, Z. and Li, Y. Feature purification: How adversarial training performs robust deep learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 977–988. IEEE, 2022.

Asi, H. and Duchi, J. C. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019.

Awasthi, P., Frank, N., and Mohri, M. Adversarial learning guarantees for linear hypotheses and neural networks. In *International Conference on Machine Learning*, pp. 431–441. PMLR, 2020.

Balcan, M.-F., Khodak, M., and Talwalkar, A. Provable guarantees for gradient-based meta-learning. In *International Conference on Machine Learning*, pp. 424–433. PMLR, 2019.

Balda, E. R., Behboodi, A., Koep, N., and Mathar, R. Adversarial risk bounds for neural networks through sparsity based compression. *arXiv preprint arXiv:1906.00698*, 2019.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Bassily, R., Feldman, V., Guzmán, C., and Talwar, K. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33:4381–4391, 2020.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

Bousquet, O. and Elisseeff, A. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Bousquet, O., Klochkov, Y., and Zhivotovskiy, N. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pp. 610–626. PMLR, 2020.

Cai, Q.-Z., Du, M., Liu, C., and Song, D. Curriculum adversarial training. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, 2018.

Cullina, D., Bhagoji, A. N., and Mittal, P. Pac-learning in the presence of adversaries. *Advances in Neural Information Processing Systems*, 31, 2018.

Denevi, G., Ciliberto, C., Grazzi, R., and Pontil, M. Learning-to-learn stochastic gradient descent with biased regularization. In *International Conference on Machine Learning*, pp. 1566–1575. PMLR, 2019.

Denevi, G., Pontil, M., and Ciliberto, C. The advantage of conditional meta-learning for biased regularization and fine tuning. *Advances in Neural Information Processing Systems*, 33:964–974, 2020.

Deng, Y., Gazagnadou, N., Hong, J., Mahdavi, M., and Lyu, L. On the hardness of robustness transfer: A perspective from rademacher complexity over symmetric difference hypothesis space. *arXiv preprint arXiv:2302.12351*, 2023.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Du, S. S., Koushik, J., Singh, A., and Póczos, B. Hypothesis transfer learning via transformation functions. *Advances in neural information processing systems*, 30, 2017.

Farnia, F. and Ozdaglar, A. Train simultaneously, generalize better: Stability of gradient-based minimax learners. In *International Conference on Machine Learning*, pp. 3174–3185. PMLR, 2021.

Farnia, F., Zhang, J. M., and Tse, D. Generalizable adversarial training via spectral normalization. *arXiv preprint arXiv:1811.07457*, 2018.

Feldman, V. and Vondrak, J. Generalization bounds for uniformly stable algorithms. *Advances in Neural Information Processing Systems*, 31, 2018.

Feldman, V. and Vondrak, J. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, pp. 1270–1279. PMLR, 2019.

Floridi, L. and Chiriatti, M. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.

Ghadimi, S., Lan, G., and Zhang, H. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155 (1-2):267–305, 2016.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 1225–1234. PMLR, 2016.

Hua, A., Gu, J., Xue, Z., Carlini, N., Wong, E., and Qin, Y. Initialization matters for adversarial transfer learning. *arXiv preprint arXiv:2312.05716*, 2023.

Kakade, S. M., Shalev-Shwartz, S., and Tewari, A. Regularization techniques for learning with matrices. *The Journal of Machine Learning Research*, 13(1):1865–1890, 2012.

Khim, J. and Loh, P.-L. Adversarial risk bounds via function transformation. *arXiv preprint arXiv:1810.09519*, 2018.

Klochkov, Y. and Zhivotovskiy, N. Stability and deviation optimal risk bounds with convergence rate $o(1/n)$. *Advances in Neural Information Processing Systems*, 34: 5065–5076, 2021.

Kurakin, A., Goodfellow, I. J., and Bengio, S. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pp. 99–112. Chapman and Hall/CRC, 2018.

Kuzborskij, I. Theory and algorithms for hypothesis transfer learning. Technical report, EPFL, 2018.

Kuzborskij, I. and Lampert, C. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pp. 2815–2824. PMLR, 2018.

Kuzborskij, I. and Orabona, F. Stability and hypothesis transfer learning. In *International Conference on Machine Learning*, pp. 942–950. PMLR, 2013.

Kuzborskij, I. and Orabona, F. Fast rates by transferring from auxiliary hypotheses. *Machine Learning*, 106:171–195, 2017.

Lei, Y. Stability and generalization of stochastic optimization with nonconvex and nonsmooth problems. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 191–227. PMLR, 2023.

Lei, Y. and Ying, Y. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pp. 5809–5819. PMLR, 2020.

Li, J. D. and Telgarsky, M. On achieving optimal adversarial test error. In *International Conference on Learning Representations*, 2023.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Mianjy, P. and Arora, R. Robustness guarantees for adversarially trained neural networks. *Advances in neural information processing systems*, 2023.

Montasser, O., Hanneke, S., and Srebro, N. Reducing adversarially robust learning to non-robust pac learning. *Advances in Neural Information Processing Systems*, 33: 14626–14637, 2020.

Mustafa, W., Lei, Y., and Kloft, M. On the generalization analysis of adversarial learning. In *International Conference on Machine Learning*, pp. 16174–16196. PMLR, 2022.

Perrot, M. and Habrard, A. A theoretical analysis of metric hypothesis transfer learning. In *International Conference on Machine Learning*, pp. 1708–1717. PMLR, 2015.

Schmidt, M., Roux, N., and Bach, F. Convergence rates of inexact proximal-gradient methods for convex optimization. *Advances in neural information processing systems*, 24, 2011.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

Sinha, A., Namkoong, H., Volpi, R., and Duchi, J. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

Srebro, N., Sridharan, K., and Tewari, A. Optimistic rates for learning with a smooth loss. *arXiv preprint arXiv:1009.3896*, 2010.

Takada, M. and Fujisawa, H. Transfer learning via ℓ_1 regularization. *Advances in Neural Information Processing Systems*, 33:14266–14277, 2020.

Tommasi, T. and Caputo, B. The more you know, the less you learn: from knowledge transfer to one-shot learning of object categories. In *Proceedings of the British Machine Vision Conference*, pp. 80–1, 2009.

Tommasi, T., Orabona, F., Castellini, C., and Caputo, B. Improving control of dexterous hand prostheses using adaptive learning. *IEEE Transactions on Robotics*, 29(1): 207–219, 2012.

Viallard, P., VIDOT, E. G., Habrard, A., and Morvant, E. A pac-bayes analysis of adversarial robustness. *Advances in Neural Information Processing Systems*, 34:14421–14433, 2021.

Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., and Gu, Q. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2020.

Wang, Y., Ma, X., Bailey, J., Yi, J., Zhou, B., and Gu, Q. On the convergence and robustness of adversarial training. *arXiv preprint arXiv:2112.08304*, 2021.

Wang, Y., Zhang, K., and Arora, R. Benign overfitting in adversarially trained neural networks. In *International Conference on Machine Learning*. PMLR, 2024.

Xiao, J., Fan, Y., Sun, R., and Luo, Z.-Q. Adversarial rademacher complexity of deep neural networks. *arXiv preprint arXiv:2211.14966*, 2022a.

Xiao, J., Fan, Y., Sun, R., Wang, J., and Luo, Z.-Q. Stability analysis and generalization bounds of adversarial training. *Advances in Neural Information Processing Systems*, 35: 15446–15459, 2022b.

Xiao, J., Sun, R., and Luo, Z.-Q. Pac-bayesian adversarially robust generalization bounds for deep neural networks. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*, 2023.

Xiao, L. and Zhang, T. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Xing, Y., Song, Q., and Cheng, G. On the algorithmic stability of adversarial training. *Advances in neural information processing systems*, 34:26523–26535, 2021.

Yin, D., Kannan, R., and Bartlett, P. Rademacher complexity for adversarially robust generalization. In *International conference on machine learning*, pp. 7085–7094. PMLR, 2019.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019.

Zhang, Y., Zhang, W., Bald, S., Pingali, V., Chen, C., and Goswami, M. Stability of SGD: Tightness analysis and improved bounds. In *Uncertainty in artificial intelligence*, pp. 2364–2373. PMLR, 2022.

Zhou, Y., Liang, Y., and Zhang, H. Understanding generalization error of SGD in nonconvex optimization. *Machine Learning*, pp. 1–31, 2022.

Zou, D., Frei, S., and Gu, Q. Provable robustness of adversarial training for learning halfspaces with noise. In *International Conference on Machine Learning*, pp. 13002–13011. PMLR, 2021.

A. Missing Proofs in Section 3

Before presenting the proof, we first give the definition of a smooth function and Rademacher complexity that will be used later.

Definition (Smoothness)**.** A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is H-smooth if its gradient is H-Lipschitz; i.e., for all $\mathrm{w}_1, \mathrm{w}_2$, $\|\nabla f(\mathrm{w}_1) - \nabla f(\mathrm{w}_2)\| \leq H \|\mathrm{w}_1 - \mathrm{w}_2\|$.

Definition (Rademacher complexity (Bartlett & Mendelson, 2002))**.** Given distribution $\mathcal{D}$, let $\mathcal{S} = \{(\mathrm{x}_i.y_i)\}_{i=1}$ drawn i.i.d. from $\mathcal{D}$. Let $\mathcal{H}$ be a class of functions $h : \mathcal{Z} \to \mathbb{R}$. We define the *empirical Rademacher complexity* of $\mathcal{H}$ measured on $\mathcal{S}$ as

$$\widehat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{H}) = \mathbb{E}_{\sigma_i, i\in[n]}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i h(\mathrm{x}_i)\right)\right].$$

where σ_i is Rademacher random variable such that $\mathrm{P}(\sigma_i = 1) = \mathrm{P}(\sigma_i = -1) = 0.5$. The *Rademacher complexity* of $\mathcal{H}$ is defined as

$$\mathfrak{R}(\mathcal{H}) = \mathbb{E}_{\mathcal{D}}\left[\mathfrak{R}_{\mathcal{S}}(\mathcal{H})\right].$$

Theorem A.1 presents a Rademacher complexity-based generalization bound, which is the basis of the proof of Theorem 3.1. The proof of Theorem A.1 follows from Kuzborskij & Orabona (2017, Proof of Theorem 4) by replacing the standard loss to its adversarial counterpart.

Theorem A.1. Under Assumption 1, let the training set $\mathcal{S}$ of size n be sampled i.i.d. from $\mathcal{D}$. For any $r \geq 0$, define the adversarial loss class w.r.t. the hypothesis class $\mathcal{H}$ as

$$\tilde{\mathcal{L}} := \left\{(\mathrm{x}, y) \mapsto \sup_{\tilde{\mathrm{x}}\in\mathcal{B}_p(\mathrm{x},\alpha)} \ell(h; (\tilde{\mathrm{x}}, y)) : h \in \mathcal{H} \wedge L_{\text{adv}}(h) \leq r\right\}.$$

Fix any $\delta > 0$, for any $h \in \mathcal{H}$ and any training set $\mathcal{S}$ of size n, with probability at least $1 - \delta$,

$$\begin{aligned} L_{\text{adv}}(h) - \widehat{L}_{\text{adv}}(h) &\leq 2\mathfrak{R}(\tilde{\mathcal{L}}) + \frac{3M\log(1/\delta)}{n\log\left(1 + \sqrt{\frac{2M\log(1/\delta)}{(4\mathfrak{R}(\tilde{\mathcal{L}})+r)n}}\right)} \\ &\leq 2\mathfrak{R}(\tilde{\mathcal{L}}) + 3\sqrt{\frac{(4\mathfrak{R}(\tilde{\mathcal{L}}) + r)M\log(1/\delta)}{2n}} + \frac{1.5M\log(1/\delta)}{n}. \end{aligned}$$

We will use the following findings from Kakade et al. (2012) on strongly convex regularizers in a general setting.

Lemma A.2 (Corollary 4 in (Kakade et al., 2012))**.** If Ω is σ-strongly convex w.r.t. $\|\cdot\|$ and $\Omega^*(0) = 0$ (Ω^* is the Fenchel conjugate of Ω), then, denoting the partial sum $\sum_{j\leq i}\mathrm{v}_j$ by $\mathrm{v}_{1:i}$, we have for any sequence $\mathrm{v}_1, \ldots, \mathrm{v}_m$ and for any u,

$$\sum_{i=1}^{m}\langle \mathrm{v}_i, \mathrm{u}\rangle - \Omega(\mathrm{u}) \leq \Omega^*(\mathrm{v}_{1:m}) \leq \sum_{i=1}^{m}\langle \nabla\Omega^*(\mathrm{v}_{1:i-1}), \mathrm{v}_i\rangle + \frac{1}{2\sigma}\sum_{i=1}^{m}\|\mathrm{v}_i\|_*^2$$

We also leverage the following lemma, which demonstrates that the solution to the optimization problem (1) has a bounded radius associated with the given auxiliary hypotheses.

Lemma A.3. Under Assumption 2, the solution of Equation (1) lies in the set $\left\{h \in \mathcal{H}, \|h\|_\infty \leq \sqrt{\frac{\kappa}{\lambda}\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}})}\right\}$.

Proof of Lemma A.3. By the definition of $h_{\widehat{\mathrm{w}}}$, $L_{\text{adv}}(h_{\widehat{\mathrm{w}}} + f_\beta^{\text{aux}}, \mathcal{S}) + \lambda\|h_{\widehat{\mathrm{w}}}\|_k^2 \leq \widehat{L}_{\text{adv}}(f_\beta^{\text{aux}})$, which gives us that $\|h_{\widehat{\mathrm{w}}}\|_k \leq \sqrt{\frac{1}{\lambda}\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}})}$. Therefore,

$$\|h_{\widehat{\mathrm{w}}}\|_\infty = \sup_{\mathrm{x}}|\langle h_{\widehat{\mathrm{w}}}, k(\mathrm{x}, \cdot)\rangle| \leq \|h_{\widehat{\mathrm{w}}}\|_k \sup_{\mathrm{x}}\sqrt{k(\mathrm{x}, \mathrm{x})} \leq \sqrt{\kappa}\|h_{\widehat{\mathrm{w}}}\|_k \leq \sqrt{\frac{\kappa}{\lambda}\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}})}.$$

□

We now study the Rademacher complexity of the adversarial loss function class.

Lemma A.4. Under Assumption 1, define the adversarial loss class w.r.t. the expanded hypothesis class $\tilde{\mathcal{H}}$ as

$$\tilde{\mathcal{L}} := \left\{ (\mathrm{x}, y) \mapsto \tilde{\ell}(h, (\mathrm{x}, y)) : h \in \tilde{\mathcal{H}} \right\}.$$

Then given an i.i.d. sample $\mathcal{S} \sim \mathcal{D}^n$ of size n and the following set

$$\left\{ \tau_i : \tau_i \geq \tilde{\ell}(h, (\mathrm{x}_i, y_i)), \forall (\mathrm{x}_i, y_i) \in \mathcal{S} \wedge \forall h \in \tilde{\mathcal{H}} \right\},$$

we have that

$$\widehat{\mathfrak{R}}_{\mathcal{S}}(\tilde{\mathcal{L}}) \leq \mathbb{E}_{\sigma_i, i \in [n]} \left[\sup_{h \in \tilde{\mathcal{H}}} \left\{ \frac{2\sqrt{3H}}{n} \sum_{i=1}^{n} \sigma_i \sqrt{\tau_i} \min_{\tilde{\mathrm{x}}_i \in \mathcal{B}(\mathrm{x}_i, \alpha)} y_i h(\tilde{\mathrm{x}}_i) \right\} \right]$$

Proof of Lemma A.4. Recall from (Srebro et al., 2010) that for any H-smooth non-negative function $\phi : \mathbb{R} \mapsto \mathbb{R}_+$ and any $z_1, z_2 \in \mathbb{R}$, $|\phi(z_1) - \phi(z_2)| \leq \sqrt{6H(\phi(z_1) + \phi(z_2))}\,|z_1 - z_2|$. Here we define $\tilde{\mathrm{x}}_1 = \mathrm{argmax}_{\tilde{\mathrm{x}} \in \mathcal{B}(\mathrm{x}_1, \alpha)} \ell(h_1, (\tilde{\mathrm{x}}, y))$, $\tilde{\mathrm{x}}_2 = \mathrm{argmax}_{\tilde{\mathrm{x}} \in \mathcal{B}(\mathrm{x}_2, \alpha)} \ell(h_2, (\tilde{\mathrm{x}}, y))$. Then choose $z_1 = y h_1(\tilde{\mathrm{x}}_1)$, $z_2 = y h_2(\tilde{\mathrm{x}}_2)$, we have that

$$|\ell(h_1, (\tilde{\mathrm{x}}_1, y)) - \ell(h_2, (\tilde{\mathrm{x}}_2, y))| \leq \sqrt{6H\left(\ell(h_1, (\tilde{\mathrm{x}}_1, y)) + \ell(h_2, (\tilde{\mathrm{x}}_2, y))\right)}\,|y h_1(\tilde{\mathrm{x}}_1) - y h_2(\tilde{\mathrm{x}}_2)|.$$

Apply adversarial loss gives us that

$$\left|\tilde{\ell}(h_1, (\mathrm{x}, y)) - \tilde{\ell}(h_2, (\mathrm{x}, y))\right| \leq \sqrt{6H\left(\tilde{\ell}(h_1, (\mathrm{x}, y)) + \tilde{\ell}(h_2, (\mathrm{x}, y))\right)} \left| \min_{\tilde{\mathrm{x}}^1 \in \mathcal{B}(\mathrm{x}, \alpha)} y h_1(\tilde{\mathrm{x}}^1) - \min_{\tilde{\mathrm{x}}^2 \in \mathcal{B}(\mathrm{x}, \alpha)} y h_2(\tilde{\mathrm{x}}^2) \right|.$$

Fix the training set $\mathcal{S}$, by the definition of empirical Rademacher complexity, we have that

$$\begin{aligned}
\widehat{\mathfrak{R}}_{\mathcal{S}}(\tilde{\mathcal{L}}) &= \frac{1}{n} \mathbb{E}_{\sigma_i, i \in [n]} \left[\sup_{h \in \tilde{\mathcal{H}}} \left\{ \sum_{i=1}^{n} \sigma_i \tilde{\ell}(h, (\mathrm{x}_i, y_i)) \right\} \right] \\
&= \frac{1}{n} \mathbb{E}_{\sigma_1, \ldots, \sigma_{n-1}} \left[\mathbb{E}_{\sigma_n} \left[\sup_{h \in \tilde{\mathcal{H}}} \left\{ u_{n-1}(h) + \sigma_n \tilde{\ell}(h, (\mathrm{x}_n, y_n)) \right\} \right] \right]
\end{aligned}$$

where $u_{n-1}(h) = \sum_{i=1}^{n-1} \sigma_i \tilde{\ell}(h, (\mathrm{x}_i, y_i))$. By the definition of supremum, for any $\gamma > 0$, there exist $h_1, h_2 \in \tilde{\mathcal{H}}$ such that

$$\begin{aligned}
u_{n-1}(h_1) + \tilde{\ell}(h_1, (\mathrm{x}_n, y_n)) &\geq (1 - \gamma) \left(\sup_{h \in \tilde{\mathcal{H}}} \left\{ u_{n-1}(h) + \tilde{\ell}(h, (\mathrm{x}_n, y_n)) \right\} \right) \\
u_{n-1}(h_2) - \tilde{\ell}(h_2, (\mathrm{x}_n, y_n)) &\geq (1 - \gamma) \left(\sup_{h \in \tilde{\mathcal{H}}} \left\{ u_{n-1}(h) - \tilde{\ell}(h, (\mathrm{x}_n, y_n)) \right\} \right)
\end{aligned}$$

Thus for any $\gamma > 0$, we have

$$\begin{aligned}
&(1 - \gamma) \mathbb{E}_{\sigma_n} \left[\sup_{h \in \tilde{\mathcal{H}}} \left\{ u_{n-1}(h) + \sigma_n \tilde{\ell}(h, (\mathrm{x}_n, y_n)) \right\} \right] \\
&= \frac{1 - \gamma}{2} \left(\sup_{h \in \tilde{\mathcal{H}}} \left\{ u_{n-1}(h) + \tilde{\ell}(h, (\mathrm{x}_n, y_n)) \right\} + \sup_{h \in \tilde{\mathcal{H}}} \left\{ u_{n-1}(h) - \tilde{\ell}(h, (\mathrm{x}_n, y_n)) \right\} \right) \\
&\leq \frac{1}{2} \left(u_{n-1}(h_1) + \tilde{\ell}(h_1, (\mathrm{x}_n, y_n)) + u_{n-1}(h_2) - \tilde{\ell}(h_2, (\mathrm{x}_n, y_n)) \right) \\
&\qquad \text{(Define } h_1 = \arg\sup_{h \in \tilde{\mathcal{H}}} \left\{ u_{n-1}(h) + \tilde{\ell}(h, (\mathrm{x}_n, y_n)) \right\}, h_2 = \arg\sup_{h \in \tilde{\mathcal{H}}} \left\{ u_{n-1}(h) - \tilde{\ell}(h, (\mathrm{x}_n, y_n)) \right\}) \\
&\leq \frac{1}{2} \left(u_{n-1}(h_1) + u_{n-1}(h_2) + \sqrt{6H\left(\tilde{\ell}(h_1; (\mathrm{x}_n, y_n)) + \tilde{\ell}(h_2; (\mathrm{x}_n, y_n)) \right)} \left| \min_{\tilde{\mathrm{x}}_n^1 \in \mathcal{B}(\mathrm{x}_n, \alpha)} y_n h_1(\tilde{\mathrm{x}}_n^1) - \min_{\tilde{\mathrm{x}}_n^2 \in \mathcal{B}(\mathrm{x}_n, \alpha)} y_n h_2(\tilde{\mathrm{x}}_n^2) \right| \right) \\
&\leq \frac{1}{2} \left(u_{n-1}(h_1) + u_{n-1}(h_2) + \sqrt{12H\tau_n} \left| \min_{\tilde{\mathrm{x}}_n^1 \in \mathcal{B}(\mathrm{x}_n, \alpha)} y_n h_1(\tilde{\mathrm{x}}_n^1) - \min_{\tilde{\mathrm{x}}_n^2 \in \mathcal{B}(\mathrm{x}_n, \alpha)} y_n h_2(\tilde{\mathrm{x}}_n^2) \right| \right) \\
&\qquad \text{(Define } s_n = \text{sign}\left(\min_{\tilde{\mathrm{x}}_n^1 \in \mathcal{B}(\mathrm{x}_n, \alpha)} y_n h_1(\tilde{\mathrm{x}}_n^1) - \min_{\tilde{\mathrm{x}}_n^2 \in \mathcal{B}(\mathrm{x}_n, \alpha)} y_n h_2(\tilde{\mathrm{x}}_n^2) \right)) \\
&\leq \frac{1}{2} \sup_{h \in \tilde{\mathcal{H}}} \left\{ u_{n-1}(h) + s_n \sqrt{12H\tau_n} \min_{\tilde{\mathrm{x}}_n \in \mathcal{B}(\mathrm{x}_n, \alpha)} y_n h(\tilde{\mathrm{x}}_n) \right\} + \frac{1}{2} \sup_{h \in \tilde{\mathcal{H}}} \left\{ u_{n-1}(h) - s_n \sqrt{12H\tau_n} \min_{\tilde{\mathrm{x}}_n \in \mathcal{B}(\mathrm{x}_n, \alpha)} y_n h(\tilde{\mathrm{x}}_n) \right\}
\end{aligned}$$

$$= \mathbb{E}_{\sigma_n}\left[\sup_{h\in\tilde{\mathcal{H}}}\left\{u_{n-1}(h) + \sigma_n\sqrt{12H\tau_n}\min_{\tilde{\mathrm{x}}_n\in\mathcal{B}(\mathrm{x}_n,\alpha)} y_n h(\tilde{\mathrm{x}}_n)\right\}\right]$$

Induction in the same way for σ_i with $i \neq n$ proves the result.

□

Theorem A.5. Assume that the learner is given a weighted linear combination $f_\beta^{\text{aux}}(\cdot) = \sum_{j=1}^k \beta_j f_j^{\text{aux}}(\cdot)$ of auxiliary hypotheses with weights $\beta \in \mathbb{R}^k$. Given a scalar $\lambda > 0$, for any i.i.d. sample $\mathcal{S} \sim \mathcal{D}^n$ of size n, define classes

$$\mathcal{H} = \left\{h \in \mathcal{H} : \|h\|_\infty \leq \sqrt{\frac{\kappa}{\lambda}\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}})}\right\}, \quad \mathcal{V} = \{\beta : \Psi(\beta) \leq \rho\},$$

and the adversarial loss class

$$\tilde{\mathcal{L}} = \left\{(\mathrm{x}, y) \mapsto \tilde{\ell}(h(\mathrm{x}) + f_\beta^{\text{aux}}(\mathrm{x}), y) : h \in \mathcal{H} \wedge \beta \in \mathcal{V}\right\}.$$

Under Assumptions 1 and 2, for the adversarial loss class $\tilde{\mathcal{L}}$, we have that

$$\mathfrak{R}(\tilde{\mathcal{L}}) \leq 4\sqrt{3H}(\kappa + \sigma C)\left(1 + \sqrt{\frac{H\kappa}{\lambda}}\right)\frac{L_{\text{adv}}(f_\beta^{\text{aux}})/\sqrt{\lambda} + \sqrt{L_{\text{adv}}(f_\beta^{\text{aux}})\rho}}{\sqrt{n\sigma}}.$$

Proof of Theorem A.5. Define $\tilde{\mathrm{x}}_i(h_\beta) = \min_{\tilde{\mathrm{x}}_i\in\mathcal{B}(\mathrm{x}_i,\alpha)} y_i(h(\tilde{\mathrm{x}}_i) + f_\beta^{\text{aux}}(\tilde{\mathrm{x}}_i)), \forall i \in [n]$. Applying Lemma A.4 gives us that,

$$\begin{aligned}
\widehat{\mathfrak{R}}_{\mathcal{S}}(\tilde{\mathcal{L}}) &\leq \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H},\beta\in\mathcal{V}}\left\{\frac{2\sqrt{3H}}{n}\sum_{i=1}^n \sigma_i\sqrt{\tau_i}\min_{\tilde{\mathrm{x}}_i\in\mathcal{B}(\mathrm{x}_i,\alpha)} y_i(h(\tilde{\mathrm{x}}_i) + f_\beta^{\text{aux}}(\tilde{\mathrm{x}}_i))\right\}\right] \\
&\leq \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H},\beta\in\mathcal{V}}\left\{\frac{2\sqrt{3H}}{n}\sum_{i=1}^n \sigma_i\sqrt{\tau_i}y_i\left\langle h, k(\tilde{\mathrm{x}}_i(h_\beta),\cdot)\right\rangle + \frac{2\sqrt{3H}}{n}\sum_{i=1}^n \sigma_i\sqrt{\tau_i}y_i\left\langle\beta, \mathrm{f}^{\text{aux}}(\tilde{\mathrm{x}}_i(h_\beta))\right\rangle\right\}\right] \\
&\leq \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\left\{\frac{2\sqrt{3H}}{n}\sum_{i=1}^n \sigma_i\sqrt{\tau_i}y_i\left\langle h, k(\tilde{\mathrm{x}}_i(h_\beta),\cdot)\right\rangle\right\}\right] \\
&\qquad + \mathbb{E}_\sigma\left[\sup_{\beta\in\mathcal{V}}\left\{\frac{2\sqrt{3H}}{n}\sum_{i=1}^n \sigma_i\sqrt{\tau_i}y_i\left\langle\beta, \mathrm{f}^{\text{aux}}(\tilde{\mathrm{x}}_i(h_\beta))\right\rangle\right\}\right]
\end{aligned}$$

where f^{aux} is defined as $\mathrm{f}^{\text{aux}} = [f_1^{\text{aux}}, f_2^{\text{aux}}, \ldots, f_k^{\text{aux}}]^\top$. Let $t > 0$. For the first term, consider $\Omega(h) = \|h\|_k^2$, setting $\mathrm{v}_i = t\sigma_i\sqrt{\tau_i}k(\tilde{\mathrm{x}}_i(h_\beta),\cdot)$ and applying Lemma A.2 gives us

$$\begin{aligned}
&\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\left\{\sum_{i=1}^n\left\langle h, t\sigma_i\sqrt{\tau_i}k(\tilde{\mathrm{x}}_i(h_\beta),\cdot)\right\rangle\right\}\right] \\
&\qquad\leq \mathbb{E}_\sigma\left[\frac{t^2}{2\sigma}\sum_{i=1}^n\|\sigma_i\sqrt{\tau_i}k(\tilde{\mathrm{x}}_i(h_\beta),\cdot)\|_*^2 + \sup_{h\in\mathcal{H}}\Omega(h) + \sum_{i=1}^n\left\langle\nabla\Omega^*(\mathrm{v}_{1:i-1}), \sigma_i t\sqrt{\tau_i}k(\tilde{\mathrm{x}}_i(h_\beta),\cdot)\right\rangle\right] \\
&\qquad\leq \frac{t^2\kappa^2}{2\sigma}\sum_{i=1}^n|\tau_i| + \frac{\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}})}{\lambda} \qquad \text{(Assumption 2, definition of } \mathcal{H}, \mathbb{E}_\sigma[\sigma_i] = 0.)
\end{aligned}$$

Similarly, for the second term, we have

$$\mathbb{E}_\sigma\left[\sup_{\beta\in\mathcal{V}}\left\{\sum_{i=1}^n\left\langle\beta, t\sigma_i\sqrt{\tau_i}\mathrm{f}^{\text{aux}}(\tilde{\mathrm{x}}_i(h_\beta))\right\rangle\right\}\right] \leq \frac{t^2C^2}{2}\sum_{i=1}^n|\tau_i| + \rho$$

Combining the two terms and optimizing over t gives us that

$$\widehat{\mathfrak{R}}_{\mathcal{S}}(\tilde{\mathcal{L}}) \leq 4\sqrt{3H}(\kappa + \sigma C)\sqrt{\frac{\frac{1}{n}\sum_{i=1}^n|\tau_i|\left(\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}})/\lambda + \rho\right)}{n\sigma}} \tag{3}$$

Since $\phi(\cdot)$ is a H-smooth monotonic decreasing function, define $\tilde{\mathrm{x}}_1 = \operatorname{argmin}_{\tilde{\mathrm{x}}\in\mathcal{B}(\mathrm{x},\alpha)} y f_\beta^{\text{aux}}(\tilde{\mathrm{x}})$, $\tilde{\mathrm{x}}_2 =$

$\operatorname{argmin}_{\tilde{\mathrm{x}}\in\mathcal{B}(\mathrm{x},\alpha)} y(h(\tilde{\mathrm{x}}) + f_\beta^{\text{aux}}(\tilde{\mathrm{x}}))$, we have

$$\begin{aligned}
\tilde{\ell}(h + f_\beta^{\text{aux}}, (\mathrm{x}, y)) &= \phi(y(h(\tilde{\mathrm{x}}_2) + f_\beta^{\text{aux}}(\tilde{\mathrm{x}}_2))) \\
&\le \phi(y f_\beta^{\text{aux}}(\tilde{\mathrm{x}}_1)) + \phi'(y f_\beta^{\text{aux}}(\tilde{\mathrm{x}}_1)) y(h(\tilde{\mathrm{x}}_2) + f_\beta^{\text{aux}}(\tilde{\mathrm{x}}_2) - f_\beta^{\text{aux}}(\tilde{\mathrm{x}}_1)) + \frac{H}{2}(h(\tilde{\mathrm{x}}_2) + f_\beta^{\text{aux}}(\tilde{\mathrm{x}}_2) - f_\beta^{\text{aux}}(\tilde{\mathrm{x}}_1))^2 \\
&\qquad \text{(By the definition of } \tilde{\mathrm{x}}_1, \tilde{\mathrm{x}}_2\text{, we have } yh(\tilde{\mathrm{x}}_2) \le y(h(\tilde{\mathrm{x}}_2) + f_\beta^{\text{aux}}(\tilde{\mathrm{x}}_2) - f_\beta^{\text{aux}}(\tilde{\mathrm{x}}_1)) \le yf(\tilde{\mathrm{x}}_1)) \\
&\le \phi(y f_\beta^{\text{aux}}(\tilde{\mathrm{x}}_1)) + \phi'(y f_\beta^{\text{aux}}(\tilde{\mathrm{x}}_1)) \max\{h(\tilde{\mathrm{x}}_1), h(\tilde{\mathrm{x}}_2)\} + \frac{H}{2}(\max\{h(\tilde{\mathrm{x}}_1), h(\tilde{\mathrm{x}}_2)\})^2 \\
&\le \tilde{\ell}(y f_\beta^{\text{aux}}(\mathrm{x})) + 2\sqrt{H\tilde{\ell}(y f_\beta^{\text{aux}}(\mathrm{x}))}\,\|h\|_\infty + \frac{H}{2}\|h\|_\infty^2 \\
&\le \tilde{\ell}(y f_\beta^{\text{aux}}(\mathrm{x})) + 2\sqrt{\frac{H\kappa}{\lambda}\tilde{\ell}(y f_\beta^{\text{aux}}(\mathrm{x}))\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}})} + \frac{H\kappa\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}})}{2\lambda}.
\end{aligned}$$

By the definition of τ_i, we have that

$$\tilde{\ell}(h + f_\beta^{\text{aux}}, (\mathrm{x}_i, y_i)) \le \tau_i = \tilde{\ell}(f_\beta^{\text{aux}}, (\mathrm{x}_i, y_i)) + 2\sqrt{\frac{H\kappa}{\lambda}\tilde{\ell}(f_\beta^{\text{aux}}, (\mathrm{x}_i, y_i))\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}})} + \frac{H\kappa\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}})}{2\lambda}.$$

As a result, applying Jensen's inequality gives us that

$$\frac{1}{n}\sum_{i=1}^n |\tau_i| \le \widehat{L}_{\text{adv}}(f_\beta^{\text{aux}}) + 2\sqrt{\frac{H\kappa}{\lambda}}\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}}) + \frac{H\kappa\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}})}{2\lambda} \le \left(1 + \sqrt{\frac{H\kappa}{\lambda}}\right)^2 \widehat{L}_{\text{adv}}(f_\beta^{\text{aux}}).$$

Plugging back into Equation (3) gives us that

$$\begin{aligned}
\mathfrak{R}(\tilde{\mathcal{L}}) = \mathbb{E}_{\mathcal{S}}\left[\widehat{\mathfrak{R}}_{\mathcal{S}}(\tilde{\mathcal{L}})\right] &\le \mathbb{E}_{\mathcal{S}}\left[4\sqrt{3H}(\kappa + \sigma C)\left(1 + \sqrt{\frac{H\kappa}{\lambda}}\right)\sqrt{\frac{\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}})\left(\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}})/\lambda + \rho\right)}{n\sigma}}\right] \\
&\le 4\sqrt{3H}(\kappa + \sigma C)\left(1 + \sqrt{\frac{H\kappa}{\lambda}}\right)\frac{L_{\text{adv}}(f_\beta^{\text{aux}})/\sqrt{\lambda} + \sqrt{L_{\text{adv}}(f_\beta^{\text{aux}})\rho}}{\sqrt{n\sigma}}.
\end{aligned}$$

□

Theorem 3.1. Assume that the learner is given a weighted linear combination $f_\beta^{\text{aux}}(\cdot) = \sum_{j=1}^k \beta_j f_j^{\text{aux}}(\cdot)$ of auxiliary hypotheses with weights $\beta \in \mathbb{R}^k$ such that $\Psi(\beta) \le \rho$. Fix any $\delta > 0$, and say Assumptions 1 and 2 hold. Then, given an i.i.d. sample $\mathcal{S} \sim \mathcal{D}^n$ of size n, for any $\lambda > 0$, the A-RERM rule returns $h_{\widehat{\mathrm{w}},\beta}$ such that with probability at least $1 - \delta$,

$$\begin{aligned}
&\left|L_{\text{adv}}(h_{\widehat{\mathrm{w}},\beta}) - \widehat{L}_{\text{adv}}(h_{\widehat{\mathrm{w}},\beta})\right| \\
&\le \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}\left(\left(1 + \frac{1}{\sqrt{\lambda}}\right)\left(\frac{L_{\text{adv}}^{\text{aux}}}{\sqrt{\lambda}} + \sqrt{L_{\text{adv}}^{\text{aux}}\rho}\right) + \sqrt{L_{\text{adv}}^{\text{aux}}}\right)\right) \\
&\quad + \tilde{\mathcal{O}}\left(\frac{1}{n}\left(\left(1 + \frac{1}{\sqrt{\lambda}}\right)\left(\sqrt{\frac{L_{\text{adv}}^{\text{aux}}}{\lambda}} + \sqrt{\rho}\right)\right)\right),
\end{aligned}$$

where $L_{\text{adv}}^{\text{aux}} = L_{\text{adv}}^{\alpha}(f_\beta^{\text{aux}})$.

Proof of Theorem 3.1. Define the adversarial loss class $\tilde{\mathcal{L}} := \left\{(\mathrm{x}, y) \mapsto \tilde{\ell}(h, (\mathrm{x}, y)) : h \in \mathcal{H}\right\}$, define the expanded hypothesis class:

$$\begin{aligned}
\tilde{\mathcal{H}} := \Bigg\{ &\mathrm{x} \mapsto h_{\mathrm{w},\beta}(\mathrm{x}) : h_{\mathrm{w},\beta} = h_{\mathrm{w}} + f_\beta^{\text{aux}}, h_{\mathrm{w}} \in \mathcal{H}, \\
&\Omega(h_{\mathrm{w}}) \le \frac{\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}})}{\lambda} \wedge \|h\|_\infty \le \sqrt{\frac{\kappa}{\lambda}\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}})} \wedge \Psi(\beta) \le \rho \wedge \widehat{L}_{\text{adv}}(h_{\mathrm{w},\beta}) \le \widehat{L}_{\text{adv}}(f_\beta^{\text{aux}})\Bigg\}.
\end{aligned}$$

Recall the optimization problem:

$$h_{\widehat{\mathrm{w}}} = \operatorname*{argmin}_{h\in\mathcal{H}} \left\{ \widehat{L}_{\mathrm{adv}}(h + f_\beta^{\mathrm{aux}}) + \lambda\Omega(h) \right\}, h_{\widehat{\mathrm{w}},\beta} = h_{\widehat{\mathrm{w}}} + f_\beta^{\mathrm{aux}} \in \tilde{\mathcal{H}}.$$

We have $\widehat{L}_{\mathrm{adv}}(h_{\widehat{\mathrm{w}}} + f_\beta^{\mathrm{aux}}) + \lambda\Omega(h_{\widehat{\mathrm{w}}}) \le \widehat{L}_{\mathrm{adv}}(f_\beta^{\mathrm{aux}})$, which gives us that $\Omega(h_{\widehat{\mathrm{w}}}) \le \frac{1}{\lambda}\widehat{L}_{\mathrm{adv}}(f_\beta^{\mathrm{aux}})$, $\widehat{L}_{\mathrm{adv}}(h_{\widehat{\mathrm{w}},\beta}) \le \widehat{L}_{\mathrm{adv}}(f_\beta^{\mathrm{aux}})$. Leveraging Lemma A.3, we have $h_{\widehat{\mathrm{w}},\beta} \in \tilde{\mathcal{H}}$. From Theorem A.5 we have that

$$\mathfrak{R}(\tilde{\mathcal{L}}) \le \mathcal{O}\left((\kappa + \sigma C)\sqrt{H}\left(1 + \sqrt{\frac{H\kappa}{\lambda}}\right) \frac{L_{\mathrm{adv}}^{\mathrm{aux}}/\sqrt{\lambda} + \sqrt{L_{\mathrm{adv}}^{\mathrm{aux}}\rho}}{\sqrt{n\sigma}} \right).$$

Note that

$$r = \sup_{h\in\tilde{\mathcal{H}}} L_{\mathrm{adv}}(h) = \sup_{h\in\tilde{\mathcal{H}}} \mathbb{E}_{\mathcal{S}}\left[\widehat{L}_{\mathrm{adv}}(h)\right] \le \mathbb{E}_{\mathcal{S}}\left[\sup_{h\in\tilde{\mathcal{H}}} \widehat{L}_{\mathrm{adv}}(h)\right] \le \mathbb{E}_{\mathcal{S}}\left[\widehat{L}_{\mathrm{adv}}(f_\beta^{\mathrm{aux}})\right] = L_{\mathrm{adv}}^{\mathrm{aux}}.$$

Plugging into Theorem A.1 gives us that

$$\begin{aligned}
&L_{\mathrm{adv}}(h_{\widehat{\mathrm{w}},\beta}) - \widehat{L}_{\mathrm{adv}}(h_{\widehat{\mathrm{w}},\beta}) \\
&\le 2\mathfrak{R}(\tilde{\mathcal{L}}) + 3\sqrt{\frac{(4\mathfrak{R}(\tilde{\mathcal{L}}) + L_{\mathrm{adv}}^{\mathrm{aux}})M\log(1/\delta)}{2n}} + \frac{1.5M\log(1/\delta)}{n} \\
&\le 2\mathfrak{R}(\tilde{\mathcal{L}}) + 3\left(\sqrt{L_{\mathrm{adv}}^{\mathrm{aux}}} + \frac{2\mathfrak{R}(\tilde{\mathcal{L}})}{\sqrt{L_{\mathrm{adv}}^{\mathrm{aux}}}}\right)\sqrt{\frac{M\log(1/\delta)}{n}} + \frac{1.5M\log(1/\delta)}{n} \qquad (\sqrt{a+b} \le \sqrt{a} + \tfrac{b}{2\sqrt{a}}) \\
&\le \mathcal{O}\left(\frac{1}{\sqrt{n}}\left(\sqrt{\frac{H}{\sigma}}(\kappa+\sigma C)\left(1+\sqrt{\frac{H\kappa}{\lambda}}\right)\left(\frac{L_{\mathrm{adv}}^{\mathrm{aux}}}{\sqrt{\lambda}} + \sqrt{L_{\mathrm{adv}}^{\mathrm{aux}}\rho}\right) + \sqrt{L_{\mathrm{adv}}^{\mathrm{aux}} M\log(1/\delta)}\right)\right) \\
&\quad + \mathcal{O}\left(\frac{1}{n}\left(\sqrt{\frac{H}{\sigma}}(\kappa+\sigma C)\left(1+\sqrt{\frac{H\kappa}{\lambda}}\right)\left(\sqrt{\frac{L_{\mathrm{adv}}^{\mathrm{aux}}}{\lambda}} + \sqrt{\rho}\right)\sqrt{M\log(1/\delta)} + M\log(1/\delta)\right)\right)
\end{aligned}$$

□

Theorem 3.2. Assume that the learner is given a weighted linear combination $f_\beta^{\mathrm{aux}}(\cdot) = \sum_{j=1}^k \beta_j f_j^{\mathrm{aux}}(\cdot)$ of auxiliary hypotheses with weights $\beta \in \mathbb{R}^k$ such that $\Psi(\beta) \le \rho$. Fix any $\delta > 0$, and say Assumptions 1 and 2 hold. Let $\tau > 0$ be such that $\sup_{h\in\mathcal{H}} \Omega(h) \le \tau$. Then, given an i.i.d. sample $\mathcal{S} \sim \mathcal{D}^n$ of size n, and setting λ as

$$\lambda = \tilde{\mathcal{O}}\left(\sqrt{\frac{1}{\tau}\sqrt{\frac{L_{\mathrm{adv}}^{\mathrm{aux}}}{n}}\left(\sqrt{L_{\mathrm{adv}}^{\mathrm{aux}}} + \sqrt{\rho}\right) + \sqrt{\frac{1}{\tau^2 n}\sqrt{\frac{L_{\mathrm{adv}}^{\mathrm{aux}}}{n}}\left(\sqrt{L_{\mathrm{adv}}^{\mathrm{aux}}} + \sqrt{\rho}\right)}}\right)$$

the A-RERM rule returns $h_{\widehat{\mathrm{w}},\beta}$ such that with probability at least $1-\delta$,

$$\begin{aligned}
L_{\mathrm{adv}}(h_{\widehat{\mathrm{w}},\beta}) - \min_{h_{\mathrm{w}}:\Omega(h_{\mathrm{w}})\le\tau} L_{\mathrm{adv}}(h_{\mathrm{w},\beta}) \le \tilde{\mathcal{O}}\Bigg(&\frac{1}{n} + \frac{\sqrt{L_{\mathrm{adv}}^{\mathrm{aux}}}}{n^{1/2}} \\
&+ \frac{\sqrt{L_{\mathrm{adv}}^{\mathrm{aux}}} + \sqrt[4]{L_{\mathrm{adv}}^{\mathrm{aux}}\rho}}{n^{1/4}}\sqrt{\tau} + \frac{\sqrt[4]{L_{\mathrm{adv}}^{\mathrm{aux}}} + \sqrt[8]{L_{\mathrm{adv}}^{\mathrm{aux}}\rho}}{n^{3/8}}\sqrt{\tau}\Bigg).
\end{aligned}$$

Proof of Theorem 3.2. For any choice of β with $\Psi(\beta) \le \rho$, denote the optimal hypothesis in the class as

$$h_{\mathrm{w}^*} = \operatorname*{argmin}_{h_{\mathrm{w}}:\Omega(h_{\mathrm{w}})\le\tau} L_{\mathrm{adv}}(h_{\mathrm{w},\beta})$$

By the definition of $h_{\widehat{\mathrm{w}}}$, we have

$$\widehat{L}_{\mathrm{adv}}(h_{\widehat{\mathrm{w}},\beta}) + \lambda\Omega(h_{\widehat{\mathrm{w}}}) \le \widehat{L}_{\mathrm{adv}}(h_{\mathrm{w}^*,\beta}) + \lambda\Omega(h_{\mathrm{w}^*})$$

Now denote $Z = (\kappa + \sigma C)\sqrt{\frac{H^2 \kappa L_{\text{adv}}^{\text{aux}}}{n}}(\sqrt{L_{\text{adv}}^{\text{aux}}} + \sqrt{\rho})$. Then follow the proof of Theorem 3.1 gives us that

$$\begin{aligned} L_{\text{adv}}(h_{\widehat{\mathrm{w}},\beta}) &\leq \widehat{L}_{\text{adv}}(h_{\mathrm{w}^*,\beta}) + \lambda\tau + \frac{Z}{\lambda} + \sqrt{\frac{M\log(1/\delta)}{n}}\sqrt{L_{\text{adv}}^{\text{aux}} + \frac{Z}{\lambda}} + \frac{M\log(1/\delta)}{n} \\ &\leq \widehat{L}_{\text{adv}}(h_{\mathrm{w}^*,\beta}) + \lambda\tau + \frac{Z}{\lambda} + \sqrt{\frac{L_{\text{adv}}^{\text{aux}} M\log(1/\delta)}{n}} + \sqrt{\frac{ZM\log(1/\delta)}{n\lambda}} + \frac{M\log(1/\delta)}{n} \end{aligned}$$

Optimize over λ gives us that

$$\begin{aligned} \lambda^* &= \sqrt{\frac{Z}{\tau} + \frac{1}{\tau}\sqrt{\frac{ZM\log(1/\delta)}{n}}} \\ &= \sqrt{\frac{(\kappa+\sigma C)}{\tau}\sqrt{\frac{H^2\kappa L_{\text{adv}}^{\text{aux}}}{n}}(\sqrt{L_{\text{adv}}^{\text{aux}}} + \sqrt{\rho}) + \frac{1}{\tau}\sqrt{(\kappa+\sigma C)\sqrt{\frac{H^2\kappa L_{\text{adv}}^{\text{aux}}}{n}}(\sqrt{L_{\text{adv}}^{\text{aux}}} + \sqrt{\rho})\frac{M\log(1/\delta)}{n}}} \end{aligned}$$

Plug it back gives us that

$$\begin{aligned} L_{\text{adv}}(h_{\widehat{\mathrm{w}},\beta}) &\leq \widehat{L}_{\text{adv}}(h_{\mathrm{w}^*,\beta}) + \sqrt{\tau}\sqrt{Z + \sqrt{\frac{ZM\log(1/\delta)}{n}}} + \sqrt{\frac{L_{\text{adv}}^{\text{aux}} M\log(1/\delta)}{n}} + \frac{M\log(1/\delta)}{n} \\ &\leq \widehat{L}_{\text{adv}}(h_{\mathrm{w}^*,\beta}) + \frac{(\sqrt{L_{\text{adv}}^{\text{aux}}} + (L_{\text{adv}}^{\text{aux}}\rho)^{1/4})}{n^{1/4}}\sqrt{\tau(\kappa+\sigma C)H\sqrt{\kappa}} \\ &\quad + \frac{(L_{\text{adv}}^{\text{aux}})^{1/4} + (L_{\text{adv}}^{\text{aux}}\rho)^{1/8}}{n^{3/8}}\left(M\log(1/\delta)(\kappa+\sigma C)H\sqrt{\kappa}\tau^2\right)^{1/4} \\ &\quad + \sqrt{\frac{L_{\text{adv}}^{\text{aux}} M\log(1/\delta)}{n}} + \frac{M\log(1/\delta)}{n} \end{aligned} \tag{4}$$

We finally use Bernstein's inequality to concentrate $\widehat{L}_{\text{adv}}(h_{\mathrm{w}^*,\beta})$ around $L_{\text{adv}}(h_{\mathrm{w}^*,\beta})$. Formally, with probability at least $1-\delta$,

$$\begin{aligned} \widehat{L}_{\text{adv}}(h_{\mathrm{w}^*,\beta}) &\leq L_{\text{adv}}(h_{\mathrm{w}^*,\beta}) + \sqrt{\frac{2\log(1/\delta)\,\mathbb{E}\left[\sum_{i=1}^n (\ell(h_{\mathrm{w}^*,\beta},(\mathrm{x}_i,y_i)) - L_{\text{adv}}(h_{\mathrm{w}^*,\beta}))^2\right]}{n}} + \frac{2M\log(1/\delta)}{3n} \\ &\leq L_{\text{adv}}(h_{\mathrm{w}^*,\beta}) + 2\sqrt{\frac{L_{\text{adv}}(h_{\mathrm{w}^*,\beta})M\log(1/\delta)}{n}} + \frac{2M\log(1/\delta)}{3n} \\ &\leq L_{\text{adv}}(h_{\mathrm{w}^*,\beta}) + 2\sqrt{\frac{L_{\text{adv}}^{\text{aux}} M\log(1/\delta)}{n}} + \frac{2M\log(1/\delta)}{3n} \end{aligned}$$

Plug it back into Equation (4) gives us that with probability at least $1-\delta$,

$$\begin{aligned} L_{\text{adv}}(h_{\widehat{\mathrm{w}},\beta}) &\leq L_{\text{adv}}(h_{\mathrm{w}^*,\beta}) + \frac{(\sqrt{L_{\text{adv}}^{\text{aux}}} + (L_{\text{adv}}^{\text{aux}}\rho)^{1/4})}{n^{1/4}}\sqrt{\tau(\kappa+\sigma C)H\sqrt{\kappa}} \\ &\quad + \frac{(L_{\text{adv}}^{\text{aux}})^{1/4} + (L_{\text{adv}}^{\text{aux}}\rho)^{1/8}}{n^{3/8}}\left(M\log(1/\delta)(\kappa+\sigma C)H\sqrt{\kappa}\tau^2\right)^{1/4} \\ &\quad + 3\sqrt{\frac{L_{\text{adv}}^{\text{aux}} M\log(1/\delta)}{n}} + \frac{2M\log(1/\delta)}{n} \end{aligned}$$

□

We now provide theoretical results that generalize Section 3 from the RKHS additive hypothesis class to two-layer neural networks with ReLU activation functions.

Theorem A.6. Assume that the learner is given a weighted linear combination $f_\beta^{\text{aux}}(\cdot) = \sum_{j=1}^k \beta_j f_j^{\text{aux}}(\cdot)$ of auxiliary hypotheses with weights $\beta \in \mathbb{R}^k$. Given a scalar $\lambda > 0$, for any i.i.d. sample $\mathcal{S} \sim \mathcal{D}^n$ of size n, define classes

$$\mathcal{H} = \left\{h \in \mathcal{H} : \Omega(h) \leq \frac{\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}})}{\lambda}\right\}, \quad \mathcal{V} = \{\beta : \Psi(\beta) \leq \rho\}.$$

and the adversarial loss class

$$\tilde{\mathcal{L}} = \left\{(\mathrm{x}, y) \mapsto \tilde{\ell}(h(\mathrm{x}) + f_\beta^{\text{aux}}(\mathrm{x}), y) : h \in \mathcal{H} \wedge \beta \in \mathcal{V}\right\}.$$

Under Assumptions 1, for the adversarial loss class $\tilde{\mathcal{L}}$, we have that

$$\mathfrak{R}(\tilde{\mathcal{L}}) \leq 4\sqrt{6H}\,(\kappa + \sigma C)\left(1 + \sqrt{\frac{2H\kappa}{\sigma\lambda}}\right)\sqrt{L_{\text{adv}}(f_\beta^{\text{aux}})\frac{(L_{\text{adv}}(f_\beta^{\text{aux}})/\lambda + \rho/2)}{n\sigma}}$$

Proof of Theorem A.6. We follow the same proof of Theorem A.5.

Define $\tilde{\mathrm{x}}_i(h_\beta) = \min_{\tilde{\mathrm{x}}_i \in \mathcal{B}(\mathrm{x}_i,\alpha)} y_i(h(\tilde{\mathrm{x}}_i) + f_\beta^{\text{aux}}(\tilde{\mathrm{x}}_i)), \forall i \in [n]$. Applying Lemma A.4 gives us that,

$$\begin{aligned}
\widehat{\mathfrak{R}}_{\mathcal{S}}(\tilde{\mathcal{L}}) &\leq \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H},\beta\in\mathcal{V}}\left\{\frac{2\sqrt{3H}}{n}\sum_{i=1}^n \sigma_i\sqrt{\tau_i}\min_{\tilde{\mathrm{x}}_i\in\mathcal{B}(\mathrm{x}_i,\alpha)} y_i(h(\tilde{\mathrm{x}}_i) + f_\beta^{\text{aux}}(\tilde{\mathrm{x}}_i))\right\}\right] \\
&\leq \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\left\{\frac{2\sqrt{3H}}{n}\sum_{i=1}^n \sigma_i\sqrt{\tau_i}y_i h(\tilde{\mathrm{x}}_i(h_\beta))\right\}\right] + \mathbb{E}_\sigma\left[\sup_{\beta\in\mathcal{V}}\left\{\frac{2\sqrt{3H}}{n}\sum_{i=1}^n \sigma_i\sqrt{\tau_i}y_i\left\langle\beta, \mathrm{f}^{\text{aux}}(\tilde{\mathrm{x}}_i(h_\beta))\right\rangle\right\}\right]
\end{aligned}$$

For the first term, recall that $h = \sum_{s=1}^m a_s\psi(\langle \mathrm{w}_s, \tilde{\mathrm{x}}_i\rangle)$, then we have

$$\begin{aligned}
&\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\left\{\frac{2\sqrt{3H}}{n}\sum_{i=1}^n \sigma_i\sqrt{\tau_i}y_i h(\tilde{\mathrm{x}}_i(h_\beta))\right\}\right] \\
&= \mathbb{E}_\sigma\left[\sup_{\mathrm{w}\in\mathcal{W}}\left\{\frac{2\sqrt{3H}}{n}\sum_{i=1}^n \sigma_i\sqrt{\tau_i}y_i\sum_{s=1}^m a_s\psi(\langle \mathrm{w}_s, \tilde{\mathrm{x}}_i(h_\beta)\rangle)\right\}\right] \\
&= \mathbb{E}_\sigma\left[\sup_{\mathrm{w}\in\mathcal{W}}\left\{\frac{2\sqrt{3H}}{n}\sum_{s=1}^m a_s\sum_{i=1}^n \sigma_i\sqrt{\tau_i}y_i\psi(\langle \mathrm{w}_s, \tilde{\mathrm{x}}_i(h_\beta)\rangle)\right\}\right] && (|\textstyle\sum_{s=1}^m a_s| \leq 1) \\
&\leq \mathbb{E}_\sigma\left[\frac{2\sqrt{3H}}{n}\sup_{\mathrm{w}_s\in\mathcal{W}}\left|\sum_{i=1}^n \sigma_i\sqrt{\tau_i}y_i\psi(\langle \mathrm{w}_s, \tilde{\mathrm{x}}_i(h_\beta)\rangle)\right|\right] \\
&\leq \mathbb{E}_\sigma\left[\frac{4\sqrt{3H}}{n}\sup_{\mathrm{w}_s\in\mathcal{W}}\left(\sum_{i=1}^n \sigma_i\sqrt{\tau_i}y_i\psi(\langle \mathrm{w}_s, \tilde{\mathrm{x}}_i(h_\beta)\rangle)\right)\right] \\
&&& (\mathbb{E}_{\sigma,\mathrm{z}}\left[\sup_{h\in\mathcal{H}}\left|\tfrac{1}{n}\textstyle\sum_{i=1}^n \sigma_i h(\mathrm{z}_i)\right|\right] \leq 2\mathbb{E}_{\sigma,\mathrm{z}}\left[\sup_{h\in\mathcal{H}}\tfrac{1}{n}\textstyle\sum_{i=1}^n \sigma_i h(\mathrm{z}_i)\right]) \\
&\leq \mathbb{E}_\sigma\left[\frac{4\sqrt{3H}}{n}\sup_{\mathrm{w}_s\in\mathcal{W}}\left(\sum_{i=1}^n \sigma_i\sqrt{\tau_i}y_i\left\langle \mathrm{w}_s, \tilde{\mathrm{x}}_i(h_\beta)\right\rangle\right)\right] && \text{(Talagrand's contraction Lemma)} \\
&= \mathbb{E}_\sigma\left[\frac{4\sqrt{3H}}{nt}\sup_{\mathrm{w}_s\in\mathcal{W}}\left(\sum_{i=1}^n \left\langle \mathrm{w}_s, t\sigma_i\sqrt{\tau_i}\tilde{\mathrm{x}}_i(h_\beta)\right\rangle\right)\right] \\
&\leq \frac{4\sqrt{3H}}{nt}\mathbb{E}_\sigma\left[\frac{t^2}{2\sigma}\sum_{i=1}^n \|\sigma_i\sqrt{\tau_i}\tilde{\mathrm{x}}_i(h_\beta)\|_*^2 + \sup_{\mathrm{w}_s\in\mathcal{W}}\Omega(\mathrm{w}_s) + \sum_{i=1}^n\left\langle\nabla\Omega^*(\mathrm{v}_{1:i-1}), \sigma_i t\sqrt{\tau_i}\tilde{\mathrm{x}}_i(h_\beta)\right\rangle\right] \\
&&& \text{(Let } t \geq 0 \text{ and set } \mathrm{v}_i = t\sigma_i\sqrt{\tau_i}\tilde{\mathrm{x}}_i(h_\beta), \text{apply Lemma A.2)} \\
&\leq \frac{4\sqrt{3H}}{nt}\left(\frac{t^2\kappa^2}{2\sigma}\sum_{i=1}^n |\tau_i| + \frac{\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}})}{\lambda}\right) \\
&&& \text{(Assumption 2, definition of } \mathcal{H}, \mathbb{E}_\sigma[\sigma_i] = 0, \sup_{\mathrm{w}_s\in\mathcal{W}}\Omega(\mathrm{w}_s) \leq \Omega(h) \leq \tfrac{\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}})}{\lambda}.)
\end{aligned}$$

The second term is derived in the same way as shown in the proof of Theorem A.5.

$$\mathbb{E}_\sigma\left[\sup_{\beta\in\mathcal{V}}\left\{\sum_{i=1}^n\left\langle\beta, t\sigma_i\sqrt{\tau_i}\mathrm{f}^{\text{aux}}(\tilde{\mathrm{x}}_i(h_\beta))\right\rangle\right\}\right] \leq \frac{t^2C^2}{2}\sum_{i=1}^n|\tau_i| + \rho$$

Combining the two terms and optimizing over t gives us that

$$\widehat{\mathfrak{R}}_{\mathcal{S}}(\tilde{\mathcal{L}}) \leq 4\sqrt{6H}(\kappa+\sigma C)\sqrt{\frac{1}{n}\sum_{i=1}^{n}|\tau_i|\,\frac{(\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}})/\lambda+\rho/2)}{n\sigma}} \tag{5}$$

Since $\phi(\cdot)$ is a H-smooth monotonic decreasing function, define $\tilde{\mathrm{x}}_1 = \operatorname{argmin}_{\tilde{\mathrm{x}}\in\mathcal{B}(\mathrm{x},\alpha)} y f_\beta^{\text{aux}}(\tilde{\mathrm{x}})$, $\tilde{\mathrm{x}}_2 = \operatorname{argmin}_{\tilde{\mathrm{x}}\in\mathcal{B}(\mathrm{x},\alpha)} y(h(\tilde{\mathrm{x}}) + f_\beta^{\text{aux}}(\tilde{\mathrm{x}}))$, then we have

$$\begin{aligned}
\tilde{\ell}(h+f_\beta^{\text{aux}},(\mathrm{x},y)) &= \phi(y(h(\tilde{\mathrm{x}}_2)+f_\beta^{\text{aux}}(\tilde{\mathrm{x}}_2))) \\
&\leq \phi(yf_\beta^{\text{aux}}(\tilde{\mathrm{x}}_1)) + \phi'(yf_\beta^{\text{aux}}(\tilde{\mathrm{x}}_1))y(h(\tilde{\mathrm{x}}_2)+f_\beta^{\text{aux}}(\tilde{\mathrm{x}}_2)-f_\beta^{\text{aux}}(\tilde{\mathrm{x}}_1)) + \frac{H}{2}(h(\tilde{\mathrm{x}}_2)+f_\beta^{\text{aux}}(\tilde{\mathrm{x}}_2)-f_\beta^{\text{aux}}(\tilde{\mathrm{x}}_1))^2 \\
&\qquad \text{(By the definition of } \tilde{\mathrm{x}}_1,\tilde{\mathrm{x}}_2\text{, we have } yh(\tilde{\mathrm{x}}_2) \leq y(h(\tilde{\mathrm{x}}_2)+f_\beta^{\text{aux}}(\tilde{\mathrm{x}}_2)-f_\beta^{\text{aux}}(\tilde{\mathrm{x}}_1)) \leq yh(\tilde{\mathrm{x}}_1)) \\
&\leq \phi(yf_\beta^{\text{aux}}(\tilde{\mathrm{x}}_1)) + \phi'(yf_\beta^{\text{aux}}(\tilde{\mathrm{x}}_1))\max\{h(\tilde{\mathrm{x}}_1),h(\tilde{\mathrm{x}}_2)\} + \frac{H}{2}(\max\{h(\tilde{\mathrm{x}}_1),h(\tilde{\mathrm{x}}_2)\})^2
\end{aligned}$$

Recall Ω is σ strongly convex, we have for its minimizer v and any w_s,

$$\|\mathrm{w}_s-\mathrm{v}\|^2 \leq \frac{2}{\sigma}(\Omega(\mathrm{w}_s)-\Omega(\mathrm{v}))$$

Choosing $\mathrm{v}=0$, we have that

$$h(\tilde{\mathrm{x}}_1)^2 = \sum_{s=1}^{m} a_s\psi(\langle \mathrm{w}_s,\mathrm{x}\rangle) \leq \kappa\sum_{s=1}^{m}\|\mathrm{w}_s\|^2 \leq \frac{2\kappa}{\sigma}\sum_{s=1}^{m}\Omega(\mathrm{w}_s) = \frac{2\kappa}{\sigma}\Omega(h) = \frac{2\kappa\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}})}{\sigma\lambda}.$$

By the definition of τ_i, we have that

$$\tilde{\ell}(h+f_\beta^{\text{aux}},(\mathrm{x}_i,y_i)) \leq \tau_i = \tilde{\ell}(f_\beta^{\text{aux}},(\mathrm{x},y)) + 2\sqrt{\frac{2H\kappa}{\sigma\lambda}\tilde{\ell}(f_\beta^{\text{aux}},(\mathrm{x},y))\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}})} + \frac{H\kappa\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}})}{\sigma\lambda}.$$

As a result, applying Jensen's inequality gives us that

$$\frac{1}{n}\sum_{i=1}^{n}|\tau_i| \leq \widehat{L}_{\text{adv}}(f_\beta^{\text{aux}}) + 2\sqrt{\frac{2H\kappa}{\sigma\lambda}}\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}}) + \frac{H\kappa\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}})}{\sigma\lambda} \leq \left(1+\sqrt{\frac{2H\kappa}{\sigma\lambda}}\right)^2\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}}).$$

Plugging back into Equation (5) gives us that

$$\begin{aligned}
\mathfrak{R}(\tilde{\mathcal{L}}) &= \mathbb{E}_{\mathcal{S}}\left[\widehat{\mathfrak{R}}_{\mathcal{S}}(\tilde{\mathcal{L}})\right] \\
&\leq \mathbb{E}_{\mathcal{S}}\left[4\sqrt{6H}\,(\kappa+\sigma C)\left(1+\sqrt{\frac{2H\kappa}{\sigma\lambda}}\right)\sqrt{\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}})\frac{(\widehat{L}_{\text{adv}}(f_\beta^{\text{aux}})/\lambda+\rho/2)}{n\sigma}}\right] \\
&\leq 4\sqrt{6H}\,(\kappa+\sigma C)\left(1+\sqrt{\frac{2H\kappa}{\sigma\lambda}}\right)\sqrt{L_{\text{adv}}(f_\beta^{\text{aux}})\frac{(L_{\text{adv}}(f_\beta^{\text{aux}})/\lambda+\rho/2)}{n\sigma}}
\end{aligned}$$

□

Theorem A.7. Assume that the learner is given a weighted linear combination $f_\beta^{\text{aux}}(\cdot) = \sum_{j=1}^{k}\beta_j f_j^{\text{aux}}(\cdot)$ of auxiliary hypotheses with weights $\beta\in\mathbb{R}^k$ such that $\Psi(\beta)\leq\rho$. Fix any $\delta>0$. Then, given an i.i.d. sample $\mathcal{S}\sim\mathcal{D}^n$ of size n, for any $\lambda>0$, the A-RERM rule returns $h_{\widehat{\mathrm{w}},\beta}$ such that with probability at least $1-\delta$,

$$\begin{aligned}
L_{\text{adv}}(h_{\widehat{\mathrm{w}},\beta}) \leq \widehat{L}_{\text{adv}}(h_{\widehat{\mathrm{w}},\beta}) &+ \mathcal{O}\left(\left(\sqrt{HL_{\text{adv}}^{\text{aux}}}+B\kappa H\right)(\kappa+\sigma C)\sqrt{\frac{(L_{\text{adv}}^{\text{aux}}/\lambda+\rho)}{n}}\right) + \frac{1.5M\log(1/\delta)}{n} \\
&+ \mathcal{O}\left(\left(L_{\text{adv}}^{\text{aux}}+\left(\sqrt{HL_{\text{adv}}^{\text{aux}}}+B\kappa H\right)(\kappa+\sigma C)\sqrt{\frac{(L_{\text{adv}}^{\text{aux}}/\lambda+\rho)}{n}}\right)\sqrt{\frac{M\log(1/\delta)}{n}}\right)
\end{aligned}$$

where $L_{\text{adv}}^{\text{aux}} = L_{\text{adv}}^{\alpha}(f_\beta^{\text{aux}})$.

Proof of Theorem A.7. The procedure is similar as the proof of Theorem 3.1. Define the adversarial loss class $\tilde{\mathcal{L}} :=$

$\left\{(\mathrm{x}, y) \mapsto \tilde{\ell}(h; (\mathrm{x}, y)) : h \in \mathcal{H}\right\}$, define the expanded hypothesis class:

$$\tilde{\mathcal{H}} := \Bigg\{ \mathrm{x} \mapsto h_{\mathrm{w},\beta}(\mathrm{x}) : h_{\mathrm{w},\beta} = h_{\mathrm{w}} + f_\beta^{\mathrm{aux}}, h_{\mathrm{w}} \in \mathcal{H}, \\ \Omega(h_{\mathrm{w}}) \leq \frac{\widehat{L}_{\mathrm{adv}}(f_\beta^{\mathrm{aux}})}{\lambda} \wedge \|h\|_\infty \leq \sqrt{m} B\kappa \wedge \Psi(\beta) \leq \rho \wedge \widehat{L}_{\mathrm{adv}}(h_{\mathrm{w},\beta}) \leq \widehat{L}_{\mathrm{adv}}(f_\beta^{\mathrm{aux}}) \Bigg\}.$$

Recall the optimization problem:

$$h_{\widehat{\mathrm{w}}} = \operatorname*{argmin}_{h \in \mathcal{H}} \left\{ \widehat{L}_{\mathrm{adv}}(h + f_\beta^{\mathrm{aux}}) + \lambda \Omega(h) \right\}, h_{\widehat{\mathrm{w}},\beta} = h_{\widehat{\mathrm{w}}} + f_\beta^{\mathrm{aux}} \in \tilde{\mathcal{H}}.$$

We have $\widehat{L}_{\mathrm{adv}}(h_{\widehat{\mathrm{w}}} + f_\beta^{\mathrm{aux}}) + \lambda\Omega(h_{\widehat{\mathrm{w}}}) \leq \widehat{L}_{\mathrm{adv}}(f_\beta^{\mathrm{aux}})$, which gives us that $\Omega(h_{\widehat{\mathrm{w}}}) \leq \frac{1}{\lambda}\widehat{L}_{\mathrm{adv}}(f_\beta^{\mathrm{aux}})$, $\widehat{L}_{\mathrm{adv}}(h_{\widehat{\mathrm{w}},\beta}) \leq \widehat{L}_{\mathrm{adv}}(f_\beta^{\mathrm{aux}})$. Leveraging Lemma A.3, we have $h_{\widehat{\mathrm{w}},\beta} \in \tilde{\mathcal{H}}$. From Theorem A.5 we have that

$$\mathfrak{R}(\tilde{\mathcal{L}}) \leq \mathcal{O}\left(\sqrt{H}(\kappa + \sigma C) \left(1 + \sqrt{\frac{2H\kappa}{\sigma\lambda}} \right) \sqrt{L_{\mathrm{adv}}(f_\beta^{\mathrm{aux}}) \frac{(L_{\mathrm{adv}}(f_\beta^{\mathrm{aux}})/\lambda + \rho)}{n\sigma}} \right).$$

Note that

$$r = \sup_{h \in \tilde{\mathcal{H}}} L_{\mathrm{adv}}(h) = \sup_{h \in \tilde{\mathcal{H}}} \mathbb{E}_{\mathcal{S}} \left[\widehat{L}_{\mathrm{adv}}(h) \right] \leq \mathbb{E}_{\mathcal{S}} \left[\sup_{h \in \tilde{\mathcal{H}}} \widehat{L}_{\mathrm{adv}}(h) \right] \leq \mathbb{E}_{\mathcal{S}} \left[\widehat{L}_{\mathrm{adv}}(f_\beta^{\mathrm{aux}}) \right] = L_{\mathrm{adv}}^{\mathrm{aux}}.$$

Plugging into Theorem A.1 gives us that

$$\begin{aligned}
&L_{\mathrm{adv}}(h_{\widehat{\mathrm{w}},\beta}) - \widehat{L}_{\mathrm{adv}}(h_{\widehat{\mathrm{w}},\beta}) \\
&\leq 2\mathfrak{R}(\tilde{\mathcal{L}}) + 3\sqrt{\frac{(4\mathfrak{R}(\tilde{\mathcal{L}}) + L_{\mathrm{adv}}^{\mathrm{aux}}) M \log(1/\delta)}{2n}} + \frac{1.5M\log(1/\delta)}{n} \\
&\leq 2\mathfrak{R}(\tilde{\mathcal{L}}) + 3\left(\sqrt{L_{\mathrm{adv}}^{\mathrm{aux}}} + \frac{2\mathfrak{R}(\tilde{\mathcal{L}})}{\sqrt{L_{\mathrm{adv}}^{\mathrm{aux}}}} \right) \sqrt{\frac{M\log(1/\delta)}{n}} + \frac{1.5M\log(1/\delta)}{n} \qquad (\sqrt{a+b} \leq \sqrt{a} + \tfrac{b}{2\sqrt{a}}) \\
&\leq \mathcal{O}\left(\frac{1}{\sqrt{n}} \left(\sqrt{\frac{H}{\sigma}}(\kappa + \sigma C) \left(1 + \sqrt{\frac{H\kappa}{\sigma\lambda}} \right) \left(\frac{L_{\mathrm{adv}}^{\mathrm{aux}}}{\sqrt{\lambda}} + \sqrt{L_{\mathrm{adv}}^{\mathrm{aux}}\rho} \right) + \sqrt{L_{\mathrm{adv}}^{\mathrm{aux}} M \log(1/\delta)} \right) \right) \\
&\quad + \mathcal{O}\left(\frac{1}{n} \left(\sqrt{\frac{H}{\sigma}}(\kappa + \sigma C) \left(1 + \sqrt{\frac{H\kappa}{\sigma\lambda}} \right) \left(\sqrt{\frac{L_{\mathrm{adv}}^{\mathrm{aux}}}{\lambda}} + \sqrt{\rho} \right) \sqrt{M\log(1/\delta)} + M\log(1/\delta) \right) \right)
\end{aligned}$$

□

If we further ignore the dependency on $H, B, \kappa, C, \log(1/\delta)$, then the generalization gap can be rewritten as:

$$L_{\mathrm{adv}}(h_{\widehat{\mathrm{w}},\beta}) - \widehat{L}_{\mathrm{adv}}(h_{\widehat{\mathrm{w}},\beta}) \leq \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}} \left(\left(1 + \frac{1}{\sqrt{\lambda}} \right) \left(\frac{L_{\mathrm{adv}}^{\mathrm{aux}}}{\sqrt{\lambda}} + \sqrt{L_{\mathrm{adv}}^{\mathrm{aux}}\rho} \right) + \sqrt{L_{\mathrm{adv}}^{\mathrm{aux}}} \right) + \frac{1}{n} \left(1 + \frac{1}{\sqrt{\lambda}} \right) \left(\sqrt{\frac{L_{\mathrm{adv}}^{\mathrm{aux}}}{\lambda}} + \sqrt{\rho} \right) \right)$$

Similar as Theorem 3.1, the generalization gap exhibits a fast rate of $\mathcal{O}(\frac{1}{n})$ when $L_{\mathrm{adv}}^{\mathrm{aux}} = \mathcal{O}(1/n)$.

B. Missing Proofs in Section 4

Although adversarial loss is in general non-smooth, it can be characterized via a definition of approximately smoothness, which we introduce below.

Definition (Xiao et al. (2022b)). Let $H > 0$ and $\eta > 0$. We say a differentiable function $g(\mathrm{w})$ is η-approximately H-gradient Lipschitz, if $\forall \mathrm{w}_1$ and w_2, we have

$$\|\nabla g(\mathrm{w}_1) - \nabla g(\mathrm{w}_2)\| \leq H \|\mathrm{w}_1 - \mathrm{w}_2\| + \eta$$

Within the above definition, Lemma B.1 introduces the properties that adversarial loss satisfies.

Lemma B.1 (Xiao et al. (2022b)). Let $\tilde{\ell}$ be the adversarial loss defined as $\tilde{\ell}(\mathrm{w};\mathrm{z}) = \max_{\tilde{\mathrm{z}}\in\mathcal{B}_p(\mathrm{z},\alpha)} \ell(\mathrm{w};\tilde{\mathrm{z}})$[1] with ℓ satisfies Assumption 4. Then $\forall \mathrm{w}_1, \mathrm{w}_2$ and $\forall \mathrm{z} \in \mathcal{Z}$, adversarial loss $\tilde{\ell}$ satisfies:

1. (L-Lipschitz) $\left\|\tilde{\ell}(\mathrm{w}_1;\mathrm{z}) - \tilde{\ell}(\mathrm{w}_2;\mathrm{z})\right\| \leq L\left\|\mathrm{w}_1 - \mathrm{w}_2\right\|$.

2. ($2H_z\alpha$-approximately H-smooth) for all subgradient $d(\mathrm{w},\mathrm{z}) \in \partial_{\mathrm{w}}\tilde{\ell}(\mathrm{w};\mathrm{z})$, we have $\|d(\mathrm{w}_1;\mathrm{z}) - d(\mathrm{w}_2;\mathrm{z})\| \leq H\left\|\mathrm{w}_1 - \mathrm{w}_2\right\| + 2H_z\alpha$.

For any vector $\mathrm{g} \in \mathbb{R}^d$, we define the following quantity:

$$G^{\gamma}(\mathrm{w},\mathrm{g}) := \frac{1}{\gamma}\left(\mathrm{w} - \mathrm{prox}_{\gamma,\lambda\Omega}(\mathrm{w} - \gamma\mathrm{g})\right).$$

Then Proximal Stochastic Adversarial Training can be rewritten as

$$\mathrm{w}_{t+1,\mathcal{S}} = \mathrm{w}_{t,\mathcal{S}} - \gamma_t G^{\gamma_t}(\mathrm{w}_{t,\mathcal{S}}, \nabla\tilde{\ell}(\mathrm{w}_{t,\mathcal{S}}, \mathrm{z}_{\xi_t})).$$

We now present several properties of $G^{\gamma}(\mathrm{w}, g)$ that will be used later in our proof.

Lemma B.2 (Lemma 5 in (Zhou et al., 2022)). Let Ω be a convex and possibly non-smooth function. Then, the following statements hold.

1. For any $\mathrm{w}, \mathrm{g}_1, \mathrm{g}_2 \in \mathcal{W}$, it holds that
$$\|G^{\gamma}(\mathrm{w},\mathrm{g}_1) - G^{\gamma}(\mathrm{w},\mathrm{g}_2)\| \leq \|\mathrm{g}_1 - \mathrm{g}_2\|$$

2. If Ω is λ-strongly convex, then for all $\mathrm{w}, \mathrm{v} \in \mathcal{W}$ and $\gamma > 0$, it holds that
$$\left\|\mathrm{prox}_{\gamma,\Omega}(\mathrm{w}) - \mathrm{prox}_{\gamma,\Omega}(\mathrm{v})\right\| \leq \frac{1}{1+\gamma\lambda}\|\mathrm{w} - \mathrm{v}\|$$

Lemma B.3 (Lemma 1 in (Ghadimi et al., 2016)). For any $\mathrm{w} \in \mathcal{W}, \mathrm{g} \in \mathbb{R}^d$, and $\gamma > 0$, it holds that

$$\langle \mathrm{g}, G^{\gamma}(\mathrm{w},\mathrm{g})\rangle \geq \|G^{\gamma}(\mathrm{w},\mathrm{g})\|^2 + \frac{\lambda}{\gamma}\left(\Omega(\mathrm{prox}_{\gamma,\lambda\Omega}(\mathrm{w} - \gamma\mathrm{g})) - \Omega(\mathrm{w})\right)$$

We first provide the result that connects the initialization with the norm of the gradient of the adversarial loss.

Lemma B.4. Under Assumption 1, 4 and 3, consider applying Proximal Stochastic Adversarial Training with training data $\mathcal{S}$, choose $\gamma_t \leq \frac{c}{t+1}$ with $0 < c < \frac{1}{H}$. Then $\forall i \in [n]$, it holds that

$$\mathbb{E}_{\mathcal{S},\xi}\left\|\nabla\tilde{\ell}(\mathrm{w}_{t,\mathcal{S}};\mathrm{z}_i)\right\| \leq \sqrt{2H\Phi_{\mathrm{adv}}(\mathrm{w}_0) + (4\mathbb{E}_{\mathcal{S}}[\nu_{\mathcal{S}}^2] + 4H_z^2\alpha^2)\log(T)} + 4H_z\alpha.$$

Proof. Proof follows the similar idea as Zhou et al. (2022, Lemma 6). Denoting $\mathrm{g}_{t,\mathcal{S}} = \nabla\tilde{\ell}(\mathrm{w}_{t,\mathcal{S}};\mathrm{z}_{\xi_t})$ as the stochastic gradient of adversarial loss sampled at iteration t. Setting $\mathrm{w} = \mathrm{w}_{t,\mathcal{S}}, \mathrm{g} = \mathrm{g}_{t,\mathcal{S}}$, apply Lemma B.3, we have

$$\left\langle \mathrm{g}_{t,\mathcal{S}}, G^{\gamma}(\mathrm{w}_{t,\mathcal{S}},\mathrm{g}_{t,\mathcal{S}})\right\rangle \geq \left\|G^{\gamma}(\mathrm{w}_{t,\mathcal{S}},\mathrm{g}_{t,\mathcal{S}})\right\|^2 + \frac{\lambda}{\gamma}\left(\Omega(\mathrm{w}_{t+1,\mathcal{S}}) - \Omega(\mathrm{w}_{t,\mathcal{S}})\right) \tag{6}$$

Since $\tilde{\ell}$ is non-negative, η-approximately$H$-smooth (with $\eta = 2H_z\alpha$), apply Xiao et al. (2022a, Lemma 4.2) gives us that

$$\tilde{\ell}(\mathrm{w}_1,\mathrm{z}) - \tilde{\ell}(\mathrm{w}_2,\mathrm{z}) \leq \left\langle \nabla\tilde{\ell}(\mathrm{w}_2,\mathrm{z}), \mathrm{w}_1 - \mathrm{w}_2\right\rangle + \frac{H}{2}\left\|\mathrm{w}_1 - \mathrm{w}_2\right\|^2 + \eta\left\|\mathrm{w}_1 - \mathrm{w}_2\right\|. \tag{7}$$

Choosing $\mathrm{w}_1 = \mathrm{w}_2 - \frac{1}{H}\nabla\tilde{\ell}(\mathrm{w}_2,\mathrm{z})$ gives us that

$$0 \leq \tilde{\ell}(\mathrm{w}_1,\mathrm{z}) \leq \tilde{\ell}(\mathrm{w}_2,\mathrm{z}) - \frac{1}{2H}\left\|\nabla\tilde{\ell}(\mathrm{w}_2,\mathrm{z})\right\|^2 + \frac{\eta}{H}\left\|\nabla\tilde{\ell}(\mathrm{w}_2,\mathrm{z})\right\|.$$

Rearranging gives us that

$$\left\|\nabla\tilde{\ell}(\mathrm{w}_2,\mathrm{z})\right\| \leq 2\eta + \sqrt{2H\tilde{\ell}(\mathrm{w}_2,\mathrm{z})}. \tag{8}$$

[1]Here we slightly abuse the notation, $\tilde{\mathrm{z}} \in \mathcal{B}_p(\mathrm{z},\alpha)$ is equivalent as $\tilde{\mathrm{x}} \in \mathcal{B}_p(\mathrm{x},\alpha)$.

Choose $\mathrm{w}_2 = \mathrm{w}_{t,\mathcal{S}}$, take expectations w.r.t the training data and the randomness gives us that

$$\begin{aligned}
\mathbb{E}_{\xi,\mathcal{S}}\left\|\nabla\tilde{\ell}(\mathrm{w}_{t,\mathcal{S}};\mathrm{z}_i)\right\| &\le \mathbb{E}_{\xi,\mathcal{S}}\sqrt{2H\tilde{\ell}(\mathrm{w}_{t,\mathcal{S}};\mathrm{z}_i)} + 2\eta \\
&\le \sqrt{2H\mathbb{E}_{\xi,\mathcal{S}}\tilde{\ell}(\mathrm{w}_{t,\mathcal{S}};\mathrm{z}_i)} + 2\eta && \text{(Jensen's inequality)} \\
&\le \sqrt{2H\mathbb{E}_{\xi,\mathcal{S}}\frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}(\mathrm{w}_{t,\mathcal{S}};\mathrm{z}_i)} + 2\eta && \text{(All samples in } \mathcal{S} \text{ are generated i.i.d. from } \mathcal{D}) \\
&\le \sqrt{2H\mathbb{E}_{\xi,\mathcal{S}}\widehat{\Phi}_{\text{adv}}(\mathrm{w}_{t,\mathcal{S}})} + 2\eta. && (9)
\end{aligned}$$

Moreover, consider a fixed $\mathcal{S}$, we have

$$\begin{aligned}
&\widehat{L}_{\text{adv}}(\mathrm{w}_{t+1,\mathcal{S}}) - \widehat{L}_{\text{adv}}(\mathrm{w}_{t,\mathcal{S}}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\tilde{\ell}(\mathrm{w}_{t+1,\mathcal{S}},\mathrm{z}_i) - \tilde{\ell}(\mathrm{w}_{t,\mathcal{S}},\mathrm{z}_i)\right] \\
&\le \frac{1}{n}\sum_{i=1}^{n}\left[\left\langle \mathrm{w}_{t+1,\mathcal{S}} - \mathrm{w}_{t,\mathcal{S}}, \nabla\tilde{\ell}(\mathrm{w}_{t,\mathcal{S}},\mathrm{z}_i)\right\rangle + \frac{H}{2}\left\|\mathrm{w}_{t+1,\mathcal{S}} - \mathrm{w}_{t,\mathcal{S}}\right\|^2 + \eta\left\|\mathrm{w}_{t+1,\mathcal{S}} - \mathrm{w}_{t,\mathcal{S}}\right\|\right] && \text{(Equation (7))} \\
&= -\left\langle \gamma_t G^{\gamma_t}(\mathrm{w}_{t,\mathcal{S}},\mathrm{g}_{t,\mathcal{S}}), \nabla\widehat{L}_{\text{adv}}(\mathrm{w}_{t,\mathcal{S}})\right\rangle + \frac{H\gamma_t^2}{2}\left\|G^{\gamma_t}(\mathrm{w}_{t,\mathcal{S}},\mathrm{g}_{t,\mathcal{S}})\right\|^2 + \eta\gamma_t\left\|G^{\gamma_t}(\mathrm{w}_{t,\mathcal{S}},\mathrm{g}_{t,\mathcal{S}})\right\| \\
&= -\gamma_t\left\langle G^{\gamma_t}(\mathrm{w}_{t,\mathcal{S}},\mathrm{g}_{t,\mathcal{S}}),\mathrm{g}_{t,\mathcal{S}}\right\rangle - \left\langle \gamma_t G^{\gamma_t}(\mathrm{w}_{t,\mathcal{S}},\mathrm{g}_{t,\mathcal{S}}), \nabla\widehat{L}_{\text{adv}}(\mathrm{w}_{t,\mathcal{S}}) - \mathrm{g}_{t,\mathcal{S}}\right\rangle \\
&\qquad + \frac{H\gamma_t^2}{2}\left\|G^{\gamma_t}(\mathrm{w}_{t,\mathcal{S}},\mathrm{g}_{t,\mathcal{S}})\right\|^2 + \eta\gamma_t\left\|G^{\gamma_t}(\mathrm{w}_{t,\mathcal{S}},\mathrm{g}_{t,\mathcal{S}})\right\| \\
&= -\gamma_t\left\langle G^{\gamma_t}(\mathrm{w}_{t,\mathcal{S}},\mathrm{g}_{t,\mathcal{S}}),\mathrm{g}_{t,\mathcal{S}}\right\rangle - \gamma_t\left\langle G^{\gamma_t}(\mathrm{w}_{t,\mathcal{S}},\nabla\widehat{L}_{\text{adv}}(\mathrm{w}_{t,\mathcal{S}})), \nabla\widehat{L}_{\text{adv}}(\mathrm{w}_{t,\mathcal{S}}) - \mathrm{g}_{t,\mathcal{S}}\right\rangle + \frac{H\gamma_t^2}{2}\left\|G^{\gamma_t}(\mathrm{w}_{t,\mathcal{S}},\mathrm{g}_{t,\mathcal{S}})\right\|^2 \\
&\qquad + \gamma_t\left\langle G^{\gamma_t}(\mathrm{w}_{t,\mathcal{S}},\nabla\widehat{L}_{\text{adv}}(\mathrm{w}_{t,\mathcal{S}})) - G^{\gamma_t}(\mathrm{w}_{t,\mathcal{S}},\mathrm{g}_{t,\mathcal{S}}), \nabla\widehat{L}_{\text{adv}}(\mathrm{w}_{t,\mathcal{S}}) - \mathrm{g}_{t,\mathcal{S}}\right\rangle + \eta\gamma_t\left\|G^{\gamma_t}(\mathrm{w}_{t,\mathcal{S}},\mathrm{g}_{t,\mathcal{S}})\right\|
\end{aligned}$$

Combined with Equation (6) gives us that

$$\begin{aligned}
&\widehat{\Phi}_{\text{adv}}(\mathrm{w}_{t+1,\mathcal{S}}) - \widehat{\Phi}_{\text{adv}}(\mathrm{w}_{t,\mathcal{S}}) \\
&= \widehat{L}_{\text{adv}}(\mathrm{w}_{t+1};\mathcal{S}) - \widehat{L}_{\text{adv}}(\mathrm{w}_{t,\mathcal{S}}) + \lambda\left(\Omega(\mathrm{w}_{t+1,\mathcal{S}}) - \Omega(\mathrm{w}_{t,\mathcal{S}})\right) \\
&\le \left(\frac{H\gamma_t^2}{2} - \gamma_t\right)\left\|G^{\gamma_t}(\mathrm{w}_{t,\mathcal{S}},\mathrm{g}_{t,\mathcal{S}})\right\|^2 - \gamma_t\left\langle G^{\gamma_t}(\mathrm{w}_{t,\mathcal{S}},\nabla\widehat{L}_{\text{adv}}(\mathrm{w}_{t,\mathcal{S}})), \nabla\widehat{L}_{\text{adv}}(\mathrm{w}_{t,\mathcal{S}}) - \mathrm{g}_{t,\mathcal{S}}\right\rangle \\
&\qquad + \gamma_t\left\langle G^{\gamma_t}(\mathrm{w}_{t,\mathcal{S}},\nabla\widehat{L}_{\text{adv}}(\mathrm{w}_{t,\mathcal{S}})) - G^{\gamma_t}(\mathrm{w}_{t,\mathcal{S}},\mathrm{g}_{t,\mathcal{S}}), \nabla\widehat{L}_{\text{adv}}(\mathrm{w}_{t,\mathcal{S}}) - \mathrm{g}_{t,\mathcal{S}}\right\rangle + \eta\gamma_t\left\|G^{\gamma_t}(\mathrm{w}_{t,\mathcal{S}},\mathrm{g}_{t,\mathcal{S}})\right\| \\
&\le \left(\frac{H\gamma_t^2}{2} - \gamma_t\right)\left\|G^{\gamma_t}(\mathrm{w}_{t,\mathcal{S}},\mathrm{g}_{t,\mathcal{S}})\right\|^2 - \gamma_t\left\langle G^{\gamma_t}(\mathrm{w}_{t,\mathcal{S}},\nabla\widehat{L}_{\text{adv}}(\mathrm{w}_{t,\mathcal{S}})), \nabla\widehat{L}_{\text{adv}}(\mathrm{w}_{t,\mathcal{S}}) - \mathrm{g}_{t,\mathcal{S}}\right\rangle \\
&\qquad + \gamma_t\left\|G^{\gamma_t}(\mathrm{w}_{t,\mathcal{S}},\nabla\widehat{L}_{\text{adv}}(\mathrm{w}_{t,\mathcal{S}})) - G^{\gamma_t}(\mathrm{w}_{t,\mathcal{S}},\mathrm{g}_{t,\mathcal{S}})\right\|\left\|\nabla\widehat{L}_{\text{adv}}(\mathrm{w}_{t,\mathcal{S}}) - \mathrm{g}_{t,\mathcal{S}}\right\| + \eta\gamma_t\left\|G^{\gamma_t}(\mathrm{w}_{t,\mathcal{S}},\mathrm{g}_{t,\mathcal{S}})\right\| \\
&\le \left(\frac{H\gamma_t^2}{2} - \gamma_t\right)\left(\left\|G^{\gamma_t}(\mathrm{w}_{t,\mathcal{S}},\mathrm{g}_{t,\mathcal{S}})\right\| - \frac{\eta}{H\gamma_t - 2}\right)^2 - \gamma_t\left\langle G^{\gamma_t}(\mathrm{w}_{t,\mathcal{S}},\nabla\widehat{L}_{\text{adv}}(\mathrm{w}_{t,\mathcal{S}})), \nabla\widehat{L}_{\text{adv}}(\mathrm{w}_{t,\mathcal{S}}) - \mathrm{g}_{t,\mathcal{S}}\right\rangle \\
&\qquad + \gamma_t\left\|\nabla\widehat{L}_{\text{adv}}(\mathrm{w}_{t,\mathcal{S}}) - \mathrm{g}_{t,\mathcal{S}}\right\|^2 + \frac{\eta^2\gamma_t}{4 - 2H\gamma_t}
\end{aligned}$$

where the last line uses Lemma B.2. Conditioning on $\mathrm{w}_{t,\mathcal{S}}$ and taking the expectation w.r.t. ξ, we further have

$$\begin{aligned}
\mathbb{E}_\xi\left[\widehat{\Phi}_{\text{adv}}(\mathrm{w}_{t+1,\mathcal{S}}) - \widehat{\Phi}_{\text{adv}}(\mathrm{w}_{t,\mathcal{S}})|\mathrm{w}_{t,\mathcal{S}}\right] &\le \left(\frac{H\gamma_t^2}{2} - \gamma_t\right)\mathbb{E}_\xi\left[\left(\left\|G^{\gamma_t}(\mathrm{w}_{t,\mathcal{S}},\mathrm{g}_{t,\mathcal{S}})\right\| - \frac{\eta}{H\gamma_t - 2}\right)^2|\mathrm{w}_{t,\mathcal{S}}\right] \\
&\quad + \gamma_t\mathbb{E}_\xi\left[\left\|\nabla\widehat{L}_{\text{adv}}(\mathrm{w}_{t,\mathcal{S}}) - \mathrm{g}_{t,\mathcal{S}}\right\|^2|\mathrm{w}_{t,\mathcal{S}}\right] + \frac{\eta^2\gamma_t}{4 - 2H\gamma_t} && (\gamma_t \le \tfrac{2}{H}, \forall t \in [T])
\end{aligned}$$

$$\leq \gamma_t \mathbb{E}_\xi \left[\left\| \nabla \widehat{L}_{\text{adv}}(\mathrm{w}_{t,\mathcal{S}}) - \mathrm{g}_{t,\mathcal{S}} \right\|^2 | \mathrm{w}_{t,\mathcal{S}} \right] + \frac{\eta^2 \gamma_t}{4 - 2H\gamma_t}$$

Further taking expectation w.r.t. the randomness of $\mathrm{w}_{t,\mathcal{S}}$ and $\mathcal{S}$, telescoping the above inequality gives us that

$$\mathbb{E}_{\xi,\mathcal{S}} \left[\widehat{\Phi}_{\text{adv}}(\mathrm{w}_{T,\mathcal{S}}) \right] \leq \mathbb{E}_{\xi,\mathcal{S}} \left[\widehat{\Phi}_{\text{adv}}(\mathrm{w}_0) \right] + \mathbb{E}_{\mathcal{S}} \left[\nu_{\mathcal{S}}^2 \right] \sum_{t=0}^{T-1} \gamma_t + \sum_{t=0}^{T-1} \frac{\eta^2 \gamma_t}{4 - 2H\gamma_t} \qquad (\gamma_t \leq \tfrac{c}{t+1}, cH \leq 1)$$

$$\leq \mathbb{E}_{\xi,\mathcal{S}} \left[\widehat{\Phi}_{\text{adv}}(\mathrm{w}_0) \right] + 2c\mathbb{E}_{\mathcal{S}} \left[\nu_{\mathcal{S}}^2 \right] \log(T) + \frac{c\eta^2}{2} \log(T) \qquad (10)$$

As a result, for $cH \leq 1$, $\forall i \in [n]$, we have

$$\mathbb{E}_{\xi,\mathcal{S}} \left\| \nabla \tilde{\ell}(\mathrm{w}_{T,\mathcal{S}}; \mathrm{z}_i) \right\| \leq \sqrt{2H\mathbb{E}_{\xi,\mathcal{S}} \widehat{\Phi}_{\text{adv}}(\mathrm{w}_{t,\mathcal{S}})} + 2\eta \qquad \text{(Equation (9))}$$

$$\leq \sqrt{2H\mathbb{E}_{\xi,\mathcal{S}} \left[\widehat{\Phi}_{\text{adv}}(\mathrm{w}_0) \right] + (4\mathbb{E}_{\mathcal{S}} \left[\nu_{\mathcal{S}}^2 \right] + \eta^2) \log(T)} + 2\eta$$

$$\leq \sqrt{2H\Phi_{\text{adv}}(\mathrm{w}_0) + (4\mathbb{E}_{\mathcal{S}} \left[\nu_{\mathcal{S}}^2 \right] + 4H_z^2\alpha^2) \log(T)} + 4H_z\alpha \qquad (\eta = 2H_z\alpha)$$

□

We first consider the case when λ is relatively large.

Lemma 4.1. Say Assumptions 1, 3 and 4 hold. Let w_T and w'_T denote the outputs on two neighboring datasets $\mathcal{S}, \mathcal{S}'$, respectively, after running PSAT for T iterations on each of the datasets using $\gamma_t = \frac{c}{t+1}$ with $0 < c < \frac{1}{H}$. Then, for $\lambda > H$, we have that:

$$\mathbb{E}_{\xi \sim \text{Uniform}[n], \mathcal{S}, \mathcal{S}^{(i)}} \|\mathrm{w}_T - \mathrm{w}'_T\| \leq \frac{4(1+\lambda)H_z\alpha}{\lambda - H} + \frac{2(1+\lambda)\sqrt{2H\Phi_{\text{adv}}(\mathrm{w}_0) + (4\mathbb{E}_{\mathcal{S}}[\nu_{\mathcal{S}}^2] + 4H_z^2\alpha^2)\log(T)}}{n(\lambda - H)}.$$

Further, for $\lambda > 2H + 1$, we have

$$\mathbb{E}_{\xi \sim \text{Uniform}[n], \mathcal{S}, \mathcal{S}^{(i)}} \|\mathrm{w}_T - \mathrm{w}'_T\|^2 \leq \frac{8(H+2)^2(1+\lambda)^2 H_z^2 \alpha^2}{(2\lambda - 3H - 2)HT} + \frac{(1+\lambda)^2 \big(32H\Phi_{\text{adv}}(\mathrm{w}_0) + \big(64\mathbb{E}_{\mathcal{S}}[\nu_{\mathcal{S}}^2] + 16H_z^2\alpha^2\big)\log(T)\big)}{(2\lambda - 3H - 2)HTn}.$$

Proof of Lemma 4.1. Given the training set $\mathcal{S} \sim \mathcal{D}^n$ and an additional example $\mathrm{z} \sim \mathcal{D}$, let $\mathcal{S}^{(i)}$ be the training set obtained by replacing the i-th example of $\mathcal{S}$ with z; namely, $\mathcal{S}^{(i)} = (\mathrm{z}_1, \ldots, \mathrm{z}_{i-1}, \mathrm{z}, \mathrm{z}_{i+1}, \ldots, \mathrm{z}_n)$. We define $\delta_{t,\mathcal{S},\mathcal{S}^{(i)}} = \left\| \mathrm{w}_{t,\mathcal{S}} - \mathrm{w}_{t,\mathcal{S}^{(i)}} \right\|$. Recall that $\tilde{\ell}$ is η-approximately H-smooth ($\eta = 2H_z\alpha$). At the t-th iteration, if $i \notin \xi_t$, which happens w.p. $\frac{n-1}{n}$, we have

$$\delta_{t+1,\mathcal{S},\mathcal{S}^{(i)}} = \left\| \text{prox}_{\gamma_t,\lambda\Omega}(\mathrm{w}_{t,\mathcal{S}} - \gamma_t \nabla \tilde{\ell}(\mathrm{w}_{t,\mathcal{S}}; \mathrm{z}_{\xi_t})) - \text{prox}_{\gamma_t,\lambda\Omega}(\mathrm{w}_{t,\mathcal{S}^{(i)}} - \gamma_t \nabla \tilde{\ell}(\mathrm{w}_{t,\mathcal{S}^{(i)}}; \mathrm{z}_{\xi_t})) \right\|$$

$$\leq \frac{1}{1+\gamma_t\lambda} \left\| \mathrm{w}_{t,\mathcal{S}} - \gamma_t \nabla \tilde{\ell}(\mathrm{w}_{t,\mathcal{S}}; \mathrm{z}_{\xi_t}) - \mathrm{w}_{t,\mathcal{S}^{(i)}} + \gamma_t \nabla \tilde{\ell}(\mathrm{w}_{t,\mathcal{S}^{(i)}}; \mathrm{z}_{\xi_t}) \right\| \qquad \text{(Lemma B.2)}$$

$$\leq \frac{1+\gamma_t H}{1+\gamma_t\lambda} \delta_{t,\mathcal{S},\mathcal{S}^{(i)}} + \frac{\gamma_t \eta}{1+\gamma_t\lambda} \qquad (\tilde{\ell} \text{ is } \eta\text{-approximately } H\text{-smooth})$$

On the other hand, if $i \in \xi_t$, which happens w.p. $\frac{1}{n}$, we have

$$\delta_{t+1,\mathcal{S},\mathcal{S}^{(i)}} = \left\| \text{prox}_{\gamma_t,\lambda\Omega}(\mathrm{w}_{t,\mathcal{S}} - \gamma_t \nabla \tilde{\ell}(\mathrm{w}_{t,\mathcal{S}}; \mathrm{z}_i)) - \text{prox}_{\gamma_t,\lambda\Omega}(\mathrm{w}_{t,\mathcal{S}^{(i)}} - \gamma_t \nabla \tilde{\ell}(\mathrm{w}_{t,\mathcal{S}^{(i)}}; \mathrm{z})) \right\|$$

$$\leq \frac{1}{1+\gamma_t\lambda} \left\| \mathrm{w}_{t,\mathcal{S}} - \gamma_t \nabla \tilde{\ell}(\mathrm{w}_{t,\mathcal{S}}; \mathrm{z}_i) - \mathrm{w}_{t,\mathcal{S}^{(i)}} + \gamma_t \nabla \tilde{\ell}(\mathrm{w}_{t,\mathcal{S}^{(i)}}; \mathrm{z}) \right\| \qquad \text{(Lemma B.2)}$$

$$\leq \frac{1}{1+\gamma_t\lambda} \delta_{t,\mathcal{S},\mathcal{S}^{(i)}} + \frac{\gamma_t}{1+\gamma_t\lambda} \left(\left\| \nabla \tilde{\ell}(\mathrm{w}_{t,\mathcal{S}}; \mathrm{z}_i) \right\| + \left\| \nabla \tilde{\ell}(\mathrm{w}_{t,\mathcal{S}^{(i)}}; \mathrm{z}) \right\| \right)$$

Combining the above two cases and taking expectation w.r.t. the randomness of $\xi, \mathcal{S}, \mathcal{S}^{(i)}$, we have

$$\begin{aligned}
&\mathbb{E}_{\xi,\mathcal{S},\mathcal{S}^{(i)}}\left[\delta_{t+1,\mathcal{S},\mathcal{S}^{(i)}}\right] \\
&\le \left[\frac{n-1}{n}\frac{1+\gamma_t H}{1+\gamma_t\lambda} + \frac{1}{n}\frac{1}{1+\gamma_t\lambda}\right]\mathbb{E}_{\xi,\mathcal{S},\mathcal{S}^{(i)}}\left[\delta_{t,\mathcal{S},\mathcal{S}^{(i)}}\right] + \frac{n-1}{n}\frac{\gamma_t\eta}{1+\gamma_t\lambda} + \frac{2}{n}\frac{\gamma_t}{1+\gamma_t\lambda}\mathbb{E}_{\xi,\mathcal{S}}\left\|\nabla\tilde{\ell}(\mathrm{w}_{t,\mathcal{S}};\mathrm{z}_i)\right\| \\
&\le \frac{1+\gamma_t H}{1+\gamma_t\lambda}\mathbb{E}_{\xi,\mathcal{S},\mathcal{S}^{(i)}}\left[\delta_{t,\mathcal{S},\mathcal{S}^{(i)}}\right] + \frac{2}{n}\frac{\gamma_t}{1+\gamma_t\lambda}\sqrt{2H\Phi_{\text{adv}}(\mathrm{w}_0) + (4\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right]+\eta^2)\log(T)} + \frac{(n+3)\gamma_t\eta}{n(1+\gamma_t\lambda)} \quad \text{(Lemma B.4)} \\
&\le \exp\left(\frac{\gamma_t}{1+\lambda}(H-\lambda)\right)\mathbb{E}_{\xi,\mathcal{S},\mathcal{S}^{(i)}}\left[\delta_{t,\mathcal{S},\mathcal{S}^{(i)}}\right] + \frac{2\gamma_t}{n}\sqrt{2H\Phi_{\text{adv}}(\mathrm{w}_0) + (4\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right]+\eta^2)\log(t)} + 4\gamma_t\eta \\
&\qquad\qquad (1+x\le\exp(x))
\end{aligned}$$

Note that the relation $x_{t+1} \le a_t x_t + b_t$ with $x_0 = 0$ unwinds from T to 0 as $x_T \le \sum_{t=1}^{T} b_t \prod_{k=t+1}^{T} a_k$. Recursively applying the above inequality over $t = 0, 1, \ldots, T-1$, with $\delta_{0,\mathcal{S},\mathcal{S}^{(i)}} = 0$, $\gamma_t = \frac{c}{t+1}$ gives us that

$$\begin{aligned}
&\mathbb{E}_{\xi,\mathcal{S},\mathcal{S}^{(i)}}\left[\delta_{T,\mathcal{S},\mathcal{S}^{(i)}}\right] \\
&\le \sum_{t=0}^{T-1}\left[\prod_{k=t+1}^{T-1}\exp\left(\frac{\gamma_k}{1+\lambda}(H-\lambda)\right)\right]\left(\frac{2\gamma_t}{n}\sqrt{2H\Phi_{\text{adv}}(\mathrm{w}_0) + (4\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right]+\eta^2)\log(t)} + 4\gamma_t\eta\right) \\
&= \sum_{t=0}^{T-1}\left[\exp\left((H-\lambda)\sum_{k=t+1}^{T-1}\frac{c}{(k+1)(1+\lambda)}\right)\right]\left(\frac{2\gamma_t}{n}\sqrt{2H\Phi_{\text{adv}}(\mathrm{w}_0) + (4\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right]+\eta^2)\log(t)} + 4\gamma_t\eta\right) \\
&\le \sum_{t=0}^{T-1}\left(\frac{t+1}{T}\right)^{\frac{c}{1+\lambda}(\lambda-H)}\left(\frac{2c\sqrt{2H\Phi_{\text{adv}}(\mathrm{w}_0) + (4\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right]+\eta^2)\log(t)}}{n(t+1)} + \frac{4c\eta}{t+1}\right) \qquad (11) \\
&\le \frac{2(1+\lambda)}{n(\lambda-H)}\sqrt{2H\Phi_{\text{adv}}(\mathrm{w}_0) + (4\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right]+\eta^2)\log(T)} + \frac{2(1+\lambda)\eta}{\lambda-H} \\
&\qquad\qquad (\lambda > H, \textstyle\sum_{t=0}^{T-1}(t+1)^{c(\lambda-H)/(1+\lambda)-1} \lesssim \int_{t=1}^{T} t^{c(\lambda-H)/(1+\lambda)-1}.) \\
&= \frac{2(1+\lambda)}{n(\lambda-H)}\sqrt{2H\Phi_{\text{adv}}(\mathrm{w}_0) + (4\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right]+4H_z^2\alpha^2)\log(T)} + \frac{4(1+\lambda)H_z\alpha}{\lambda-H} \qquad (\eta = 2H_z\alpha)
\end{aligned}$$

We can bound $\mathbb{E}_{\xi,\mathcal{S},\mathcal{S}^{(i)}}\left[\delta^2_{T,\mathcal{S},\mathcal{S}^{(i)}}\right]$ in a similar fashion. $\tilde{\ell}$ is η-approximately H-smooth gives us

$$\begin{aligned}
\mathbb{E}_{\xi,\mathcal{S}}\left\|\nabla\tilde{\ell}(\mathrm{w}_{t,\mathcal{S}};\mathrm{z}_i)\right\|^2 &\le 4H\mathbb{E}_{\xi,\mathcal{S}}\tilde{\ell}(\mathrm{w}_{t,\mathcal{S}},\mathrm{z}_i) + 8\eta^2 && \text{(Equation (8))} \\
&\le 4H\mathbb{E}_{\xi,\mathcal{S}}\left[\widehat{\Phi}_{\text{adv}}(\mathrm{w}_{t,\mathcal{S}})\right] + 8\eta^2 && \text{(Equation (10))} \\
&\le 4H\Phi_{\text{adv}}(\mathrm{w}_0) + \left(8\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right]+2\eta^2\right)\log(t) + 8\eta^2 && (12)
\end{aligned}$$

At the t-th iteration, if $i \notin \xi_t$, which happens w.p. $\frac{n-1}{n}$, we have

$$\begin{aligned}
\delta^2_{t+1,\mathcal{S},\mathcal{S}^{(i)}} &= \left\|\text{prox}_{\gamma_t,\lambda\Omega}(\mathrm{w}_{t,\mathcal{S}} - \gamma_t\nabla\tilde{\ell}(\mathrm{w}_{t,\mathcal{S}};\mathrm{z}_{\xi_t})) - \text{prox}_{\gamma_t,\lambda\Omega}(\mathrm{w}_{t,\mathcal{S}^{(i)}} - \gamma_t\nabla\tilde{\ell}(\mathrm{w}_{t,\mathcal{S}^{(i)}};\mathrm{z}_{\xi_t}))\right\|^2 \\
&\le \frac{1}{(1+\gamma_t\lambda)^2}\left\|\mathrm{w}_{t,\mathcal{S}} - \gamma_t\nabla\tilde{\ell}(\mathrm{w}_{t,\mathcal{S}};\mathrm{z}_{\xi_t}) - \mathrm{w}_{t,\mathcal{S}^{(i)}} + \gamma_t\nabla\tilde{\ell}(\mathrm{w}_{t,\mathcal{S}^{(i)}};\mathrm{z}_{\xi_t})\right\|^2 && \text{(Lemma B.2)} \\
&= \frac{1}{(1+\gamma_t\lambda)^2}\Bigg(\left\|\mathrm{w}_{t,\mathcal{S}} - \mathrm{w}_{t,\mathcal{S}^{(i)}}\right\|^2 - 2\gamma_t\left\langle \mathrm{w}_{t,\mathcal{S}} - \mathrm{w}_{t,\mathcal{S}^{(i)}}, \nabla\tilde{\ell}(\mathrm{w}_{t,\mathcal{S}};\mathrm{z}_{\xi_t}) - \nabla\tilde{\ell}(\mathrm{w}_{t,\mathcal{S}^{(i)}};\mathrm{z}_{\xi_t})\right\rangle \\
&\qquad + \gamma_t^2\left\|\nabla\tilde{\ell}(\mathrm{w}_{t,\mathcal{S}};\mathrm{z}_{\xi_t}) - \nabla\tilde{\ell}(\mathrm{w}_{t,\mathcal{S}^{(i)}};\mathrm{z}_{\xi_t})\right\|^2\Bigg) \\
&\le \frac{1}{(1+\gamma_t\lambda)^2}\left(\delta^2_{t,\mathcal{S},\mathcal{S}^{(i)}} + 2\gamma_t\delta_{t,\mathcal{S},\mathcal{S}^{(i)}}\left(H\delta_{t,\mathcal{S},\mathcal{S}^{(i)}} + \eta\right) + \gamma_t^2\left(H\delta_{t,\mathcal{S},\mathcal{S}^{(i)}} + \eta\right)^2\right) \\
&= \frac{1}{(1+\gamma_t\lambda)^2}\left((1+H\gamma_t)^2\delta_{t,\mathcal{S},\mathcal{S}^{(i)}} + 2\gamma_t\eta(1+H\gamma_t)\delta_{t,\mathcal{S},\mathcal{S}^{(i)}} + \gamma_t^2\eta^2\right) \\
&\le \frac{(1+\gamma_t H)(1+\gamma_t H+\gamma_t)}{(1+\gamma_t\lambda)^2}\delta^2_{t,\mathcal{S},\mathcal{S}^{(i)}} + \frac{\gamma_t(1+\gamma_t H+\gamma_t)}{(1+\gamma_t\lambda)^2}\eta^2
\end{aligned}$$

On the other hand, if $i \in \xi_t$, which happens w.p. $\frac{1}{n}$, we have

$$\begin{aligned}
\delta^2_{t+1,\mathcal{S},\mathcal{S}^{(i)}} &= \left\|\mathrm{prox}_{\gamma_t,\lambda\Omega}(\mathrm{w}_{t,\mathcal{S}} - \gamma_t\nabla\tilde{\ell}(\mathrm{w}_{t,\mathcal{S}};\mathrm{z}_i)) - \mathrm{prox}_{\gamma_t,\lambda\Omega}(\mathrm{w}_{t,\mathcal{S}^{(i)}} - \gamma_t\nabla\tilde{\ell}(\mathrm{w}_{t,\mathcal{S}^{(i)}};\mathrm{z}_i))\right\|^2 \\
&\le \frac{1}{(1+\gamma_t\lambda)^2}\left\|\mathrm{w}_{t,\mathcal{S}} - \gamma_t\nabla\tilde{\ell}(\mathrm{w}_{t,\mathcal{S}};\mathrm{z}_i) - \mathrm{w}_{t,\mathcal{S}^{(i)}} + \gamma_t\nabla\tilde{\ell}(\mathrm{w}_{t,\mathcal{S}^{(i)}};\mathrm{z}_i)\right\|^2 && \text{(Lemma B.2)} \\
&\le \frac{2}{(1+\gamma_t\lambda)^2}\delta^2_{t,\mathcal{S},\mathcal{S}^{(i)}} + \frac{4\gamma_t^2}{(1+\gamma_t\lambda)^2}\left(\left\|\nabla\tilde{\ell}(\mathrm{w}_{t,\mathcal{S}};\mathrm{z}_i)\right\|^2 + \left\|\nabla\tilde{\ell}(\mathrm{w}_{t,\mathcal{S}^{(i)}};\mathrm{z}_i)\right\|^2\right)
\end{aligned}$$

Combining the above two cases and taking expectation w.r.t. the randomness of $\xi, \mathcal{S}, \mathcal{S}^{(i)}$, we have

$$\begin{aligned}
&\mathbb{E}_{\xi,\mathcal{S},\mathcal{S}^{(i)}}\left[\delta^2_{t+1,\mathcal{S},\mathcal{S}^{(i)}}\right] \\
&\le \left[\frac{n-1}{n}\frac{(1+\gamma_tH)(1+\gamma_tH+\gamma_t)}{(1+\gamma_t\lambda)^2} + \frac{1}{n}\frac{1}{(1+\gamma_t\lambda)^2}\right]\mathbb{E}_{\xi,\mathcal{S},\mathcal{S}^{(i)}}\left[\delta^2_{t,\mathcal{S},\mathcal{S}^{(i)}}\right] + \frac{n-1}{n}\frac{\gamma_t(1+\gamma_tH+\gamma_t)\eta^2}{(1+\gamma_t\lambda)^2} \\
&\qquad + \frac{2}{n}\frac{4\gamma_t^2}{(1+\gamma_t\lambda)^2}\mathbb{E}_{\xi,\mathcal{S}}\left\|\nabla\tilde{\ell}(\mathrm{w}_{t,\mathcal{S}};\mathrm{z}_i)\right\|^2 \\
&\le \frac{(1+\gamma_tH)(1+\gamma_tH+\gamma_t)}{(1+\gamma_t\lambda)^2}\mathbb{E}_{\xi,\mathcal{S},\mathcal{S}^{(i)}}\left[\delta^2_{t,\mathcal{S},\mathcal{S}^{(i)}}\right] + \frac{n-1}{n}\frac{\gamma_t(1+\gamma_tH+\gamma_t)\eta^2}{(1+\gamma_t\lambda)^2} + \frac{64\gamma_t^2\eta^2}{n(1+\gamma_t^2\lambda^2)^2} && \text{(Equation(12))} \\
&\qquad + \frac{2}{n}\frac{4\gamma_t^2}{(1+\gamma_t\lambda)^2}\left(4H\mathbb{E}_{\xi,\mathcal{S}}\left[\Phi_{\text{adv}}(\mathrm{w}_0)\right] + \left(8\mathbb{E}_{\mathcal{S}}\left[\nu^2_{\mathcal{S}}\right] + 2\eta^2\right)\log(t)\right) \\
&\le \exp\left(\frac{\gamma_t(2H-2\lambda+2)+\gamma_t^2(H^2-\lambda^2)}{(1+\lambda)^2}\right)\mathbb{E}_{\xi,\mathcal{S},\mathcal{S}^{(i)}}\left[\delta^2_{t,\mathcal{S},\mathcal{S}^{(i)}}\right] + \frac{8\gamma_t^2}{n}\left(4H\Phi_{\text{adv}}(\mathrm{w}_0) + \left(8\mathbb{E}_{\mathcal{S}}\left[\nu^2_{\mathcal{S}}\right] + 2\eta^2\right)\log(t)\right) \\
&\qquad + 2(H+2)\gamma_t^2\eta^2 && (n \ge 33)
\end{aligned}$$

Note that the relation $x_{t+1} \le a_tx_t + b_t$ with $x_0 = 0$ unwinds from T to 0 as $x_T \le \sum_{t=1}^T b_t\prod_{k=t+1}^T a_k$. Recursively applying the above inequality over $t = 0, 1, \ldots, T-1$, with $\delta_{0,\mathcal{S},\mathcal{S}^{(i)}} = 0$, $\gamma_t = \frac{c}{t+1}$ gives us that

$$\begin{aligned}
&\mathbb{E}_{\xi,\mathcal{S},\mathcal{S}^{(i)}}\left[\delta^2_{T,\mathcal{S},\mathcal{S}^{(i)}}\right] \\
&\le \sum_{t=0}^{T-1}\left[\prod_{k=t+1}^{T-1}\exp\left(\frac{\gamma_k(2H-2\lambda+2))}{(1+\lambda)^2}\right)\cdot\prod_{k=t+1}^{T-1}\exp\left(\frac{\gamma_k^2}{(1+\lambda)^2}(H^2-\lambda^2)\right)\right] \\
&\qquad\cdot\left(\frac{8\gamma_t^2}{n}\left(4H\Phi_{\text{adv}}(\mathrm{w}_0) + \left(8\mathbb{E}_{\mathcal{S}}\left[\nu^2_{\mathcal{S}}\right] + 2\eta^2\right)\log(t)\right) + 2(H+2)^2\gamma_t^2\eta^2\right) \\
&= \sum_{t=0}^{T-1}\left[\exp\left((2H-2\lambda+2)\sum_{k=t+1}^{T-1}\frac{c}{(k+1)(1+\lambda)^2}\right)\cdot\exp\left((H^2-\lambda^2)\sum_{k=t+1}^{T-1}\frac{c^2}{(k+1)^2(1+\lambda)^2}\right)\right] \\
&\qquad\cdot\left(\frac{8\gamma_t^2}{n}\left(4H\Phi_{\text{adv}}(\mathrm{w}_0) + \left(8\mathbb{E}_{\mathcal{S}}\left[\nu^2_{\mathcal{S}}\right] + 2\eta^2\right)\log(t)\right) + 2(H+2)^2\gamma_t^2\eta^2\right) \\
&\le \sum_{t=0}^{T-1}\exp\left(\frac{c^2(H^2-\lambda^2)}{(t+2)(1+\lambda)^2}\right)\cdot\left(\frac{t+1}{T}\right)^{2c(\lambda-H-1)/(1+\lambda)^2} \\
&\qquad\cdot\left(\frac{8\gamma_t^2}{n}\left(4H\Phi_{\text{adv}}(\mathrm{w}_0) + \left(8\mathbb{E}_{\mathcal{S}}\left[\nu^2_{\mathcal{S}}\right] + 2\eta^2\right)\log(t)\right) + 2(H+2)^2\gamma_t^2\eta^2\right) \\
&\le \sum_{t=0}^{T-1}\left(\frac{t+1}{T}\right)^{2c(\lambda-H-1)/(1+\lambda)^2}\left(\frac{8c^2\left(4H\Phi_{\text{adv}}(\mathrm{w}_0) + \left(8\mathbb{E}_{\mathcal{S}}\left[\nu^2_{\mathcal{S}}\right] + 2\eta^2\right)\log(T)\right)}{n(t+1)^2} + \frac{2c^2(H+2)^2\eta^2}{(t+1)^2}\right) \\
&\qquad\qquad (\lambda \ge H+1) \\
&\le \frac{2c^2(1+\lambda)^2}{(2c(\lambda-H-1)-1)T}\left(\frac{4\left(4H\Phi_{\text{adv}}(\mathrm{w}_0) + \left(8\mathbb{E}_{\mathcal{S}}\left[\nu^2_{\mathcal{S}}\right] + 2\eta^2\right)\log(T)\right)}{n} + (H+2)^2\eta^2\right) \\
&\qquad\qquad \text{(choose } c = \tfrac{1}{H} \text{ to maximize } \tfrac{c^2}{(2c(\lambda-H)-1)T}\text{.)}
\end{aligned}$$

$$\leq \frac{2(1+\lambda)^2}{(2\lambda-3H-2)HT}\left(\frac{\left(16H\Phi_{\text{adv}}(\mathrm{w}_0)+\left(32\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right]+8\eta^2\right)\log\left(T\right)\right)}{n}+(H+2)^2\eta^2\right)$$
$$= \frac{2(1+\lambda)^2}{(2\lambda-3H-2)HT}\left(\frac{\left(16H\Phi_{\text{adv}}(\mathrm{w}_0)+\left(32\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right]+8H_z^2\alpha^2\right)\log\left(T\right)\right)}{n}+4(H+2)^2H_z^2\alpha^2\right) \qquad (\eta = 2H_z\alpha)$$

□

Leveraging Lemma 4.1 gives us the robust generalization guarantees.

Theorem 4.2. Say Assumptions 1, 3 and 4 hold. Let w_T denote the output on a sample $\mathcal{S}\sim\mathcal{D}^n$ of size n after running PSAT for T iterations with $\gamma_t=\frac{c}{t+1}$ for $0<c<\frac{1}{H}$, and let $\lambda>2H+1$. Then, $\forall\delta>0$, w.p. at least $1-\delta$, we have that

$$\begin{aligned}
L_{\text{adv}}(\mathrm{w}_T)-\widehat{L}_{\text{adv}}(\mathrm{w}'_T) &\leq \frac{1.5M\log\left(1/\delta\right)}{n}\\
&+\Bigg(8\sqrt{\frac{2L^2(H\Phi_{\text{adv}}(\mathrm{w}_0)+(4\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right]+H_z^2\alpha^2)\log\left(T\right))\log\left(1/\delta\right)}{nTH(2\lambda-3H-2)}}\\
&+\frac{2L}{n(\lambda-H)}\sqrt{2H\Phi_{\text{adv}}(\mathrm{w}_0)+(4\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right]+4H_z^2\alpha^2)\log\left(T\right)}\\
&+\frac{4LH_z\alpha}{\lambda-H}+\sqrt{\frac{32(H+2)^2L^2H_z^2\alpha^2\log\left(1/\delta\right)}{(2\lambda-3H-2)HT}}\Bigg)(1+\lambda)
\end{aligned}$$

Proof of Theorem 4.2. Apply Lemma 4.1 gives us μ_1 and μ_2 as follows.

$$\begin{aligned}
&\sup_{\mathrm{z}'}\mathbb{E}_{\xi,\mathcal{S},\mathcal{S}^{(i)}}\left[\tilde{\ell}(\mathrm{w}_{T,\mathcal{S}},\mathrm{z}')-\tilde{\ell}(\mathrm{w}_{T,\mathcal{S}^{(i)}},\mathrm{z}')\right]\\
&\leq L\mathbb{E}_{\xi,\mathcal{S},\mathcal{S}^{(i)}}\left\|\mathrm{w}_{T,\mathcal{S}}-\mathrm{w}_{T,\mathcal{S}^{(i)}}\right\|\\
&\leq \frac{2L(1+\lambda)}{n(\lambda-H)}\sqrt{2H\Phi_{\text{adv}}(\mathrm{w}_0)+(4\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right]+4H_z^2\alpha^2)\log\left(T\right)}+\frac{4L(1+\lambda)H_z\alpha}{\lambda-H}:=\mu_1
\end{aligned}$$

$$\begin{aligned}
&\sup_{\mathrm{z}'}\mathbb{E}_{\xi,\mathcal{S},\mathcal{S}^{(i)}}\left[\left(\tilde{\ell}(\mathrm{w}_{T,\mathcal{S}},\mathrm{z}')-\tilde{\ell}(\mathrm{w}_{T,\mathcal{S}^{(i)}},\mathrm{z}')\right)^2\right]\\
&\leq L^2\mathbb{E}_{\xi,\mathcal{S},\mathcal{S}^{(i)}}\left\|\mathrm{w}_{T,\mathcal{S}}-\mathrm{w}_{T,\mathcal{S}^{(i)}}\right\|^2\\
&\leq \frac{32L^2(1+\lambda)^2H\Phi_{\text{adv}}(\mathrm{w}_0)+\left(64\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right]+16H_z^2\alpha^2\right)L^2(1+\lambda)^2\log\left(T\right)}{(2\lambda-3H-2)HTn}+\frac{8(H+2)^2L^2(1+\lambda)^2H_z^2\alpha^2}{(2\lambda-3H-2)HT}:=\mu_2
\end{aligned}$$

Apply Theorem 2.1 on the adversarial loss gives us that

$$\begin{aligned}
&L_{\text{adv}}(\mathrm{w}_{T,\mathcal{S}})-\widehat{L}_{\text{adv}}(\mathrm{w}_{T,\mathcal{S}})\\
&\qquad\leq \frac{2L(1+\lambda)}{n(\lambda-H)}\sqrt{2H\Phi_{\text{adv}}(\mathrm{w}_0)+(4\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right]+4H_z^2\alpha^2)\log\left(T\right)}+\frac{4L(1+\lambda)H_z\alpha}{\lambda-H}+\frac{1.5M\log\left(1/\delta\right)}{n}\\
&\qquad+\sqrt{4(1+\lambda)^2\left(\frac{32L^2H\Phi_{\text{adv}}(\mathrm{w}_0)+(64\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right]+16H_z^2\alpha^2)L^2\log\left(T\right)}{(2\lambda-3H-2)HTn}+\frac{8L^2(H+2)^2H_z^2\alpha^2}{(2\lambda-3H-2)HT}\right)\log\left(1/\delta\right)}
\end{aligned}$$

□

We next discuss the case when λ is relatively small. We will be using the following Lemma.

Lemma B.5 (Lemma 5 in Kuzborskij & Lampert (2018))**.** Assume that the loss $\ell(\cdot,\mathrm{z})\in[0,M]$ is L-Lipschitz for all z. Then for every $t_0\in\{1,\ldots,n\}$, we have

$$\mathbb{E}_{\mathcal{S},\mathrm{z},\xi}\left[\ell(\mathrm{w}_{\mathcal{S},T},\mathrm{z})-\ell(\mathrm{w}_{\mathcal{S}^{(i)},T},\mathrm{z})\right]\leq L\mathbb{E}_{\mathcal{S},\mathrm{z}}\left[\mathbb{E}_{\xi}\left[\delta_T(\mathcal{S},\mathrm{z})=0\right]|\delta_{t_0}(\mathcal{S},\mathrm{z})=0\right]+\mathbb{E}_{\mathcal{S},\xi}\left[L_{\text{adv}}(\mathrm{w}_{T,\mathcal{S}})\right]\frac{t_0}{n}$$

Lemma B.6. Let $a,y>0$ and $0<b<1$. Then $x-ax^b-y\leq 0$ implies

$$x\leq\max\left\{2^{\frac{b}{1-b}}a^{\frac{1}{1-b}},a(2y)^b\right\}+y$$

Theorem 4.3. Say Assumptions 1, 3 and 4 hold. Let w_T denote the output on a sample $\mathcal{S} \sim \mathcal{D}^n$ of size n after running PSAT for $T = O(n)$ iterations with $\gamma_t = \frac{c}{t+1}$ for $0 < c < \frac{1}{H}$, and $0 < \lambda < H$, we have that

$$\mathbb{E}_{\xi,\mathcal{S}}\Big[L_{\text{adv}}(\mathrm{w}_T) - \widehat{L}_{\text{adv}}(\mathrm{w}_T)\Big] \leq \mathcal{O}\left(\max\left\{\frac{T^q}{n^{q+1}}\frac{L\,(Q+nH_z\alpha)}{H(1-q)}, \right.\right.$$
$$\left.\left.\frac{T^{\frac{q}{q+1}}}{n}\left(\mathbb{E}_{\xi,\mathcal{S}}\widehat{L}_{\text{adv}}(\mathrm{w}_T)\right)^{\frac{q}{q+1}}\left[\frac{L\,(Q+nH_z\alpha)}{H(1-q)}\right]^{\frac{1}{q+1}}\right\}\right),$$

where $Q = \sqrt{2H\Phi_{\text{adv}}(\mathrm{w}_0) + (4\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right] + 4H_z^2\alpha^2)\log(T)}$, $q = 1 - \frac{\lambda}{H}$.

Proof of Theorem 4.3. Given the training set $\mathcal{S} \sim \mathcal{D}^n$ and an additional example $\mathrm{z} \sim \mathcal{D}$, let $\mathcal{S}^{(i)}$ be the training set obtained by replacing the i-th example of $\mathcal{S}$ with z; namely, $\mathcal{S}^{(i)} = (\mathrm{z}_1, \ldots, \mathrm{z}_{i-1}, \mathrm{z}, \mathrm{z}_{i+1}, \ldots, \mathrm{z}_n)$. We define $\delta_{t,\mathcal{S},\mathcal{S}^{(i)}} = \left\|\mathrm{w}_{t,\mathcal{S}} - \mathrm{w}_{t,\mathcal{S}^{(i)}}\right\|$. Let $t_0 \in \{1, \ldots, n\}$ be the iteration that $\delta_{t_0,\mathcal{S},\mathcal{S}^{(i)}} = 0$, and PSAT picks two different samples from $\mathcal{S}$ and $\mathcal{S}^{(i)}$ in iteration $t_0 + 1$.

We follow the proof of Lemma 4.1 until Equation (11), which gives us that

$$\mathbb{E}_{\xi,\mathcal{S},\mathcal{S}^{(i)}}\left[\delta_{T,\mathcal{S},\mathcal{S}^{(i)}}\right] \leq \sum_{t=0}^{T-1}\left(\frac{t+1}{T}\right)^{c(\lambda-H)}\left(\frac{2c\sqrt{2H\Phi_{\text{adv}}(\mathrm{w}_0) + (4\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right] + \eta^2)\log(t)}}{n(t+1)} + \frac{4c\eta}{t+1}\right)$$

As we consider $\lambda < H$, we have

$$\mathbb{E}_{\xi,\mathcal{S},\mathcal{S}^{(i)}}\left[\delta_{T,\mathcal{S},\mathcal{S}^{(i)}} | \delta_{t_0,\mathcal{S},\mathcal{S}^{(i)}} = 0\right] \leq \frac{4}{(H-\lambda)}\left(\frac{T}{t_0}\right)^{1-\frac{\lambda}{H}}\left(\frac{\sqrt{2H\Phi_{\text{adv}}(\mathrm{w}_0) + (4\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right] + \eta^2)\log(T)}}{n} + 2\eta\right)$$

We know from Lemma B.1 that adversarial loss $\tilde{\ell}$ is L-Lipschitz. Recall that the PSAT assumes sampling from the uniform distribution over $[n]$ without replacement, therefore apply Lemma B.5 gives us that

$$\mathbb{E}_{\mathcal{S},\mathrm{z},\xi}\left[\tilde{\ell}(\mathrm{w}_{\mathcal{S},T}, \mathrm{z}) - \tilde{\ell}(\mathrm{w}_{\mathcal{S}^{(i)},T}, \mathrm{z})\right]$$
$$\leq \frac{4L}{(H-\lambda)}\left(\frac{T}{t_0}\right)^{1-\frac{\lambda}{H}}\left(\frac{\sqrt{2H\Phi_{\text{adv}}(\mathrm{w}_0) + (4\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right] + \eta^2)\log(T)}}{n} + 2\eta\right) + \mathbb{E}_{\mathcal{S},\xi}\left[L_{\text{adv}}(\mathrm{w}_{T,\mathcal{S}})\right]\frac{t_0}{n}$$

Define

$$t_0 = \left\lceil \frac{4L}{H\mathbb{E}_{\mathcal{S},\xi}\left[L_{\text{adv}}(\mathrm{w}_{T,\mathcal{S}})\right]}\left(\sqrt{2H\Phi_{\text{adv}}(\mathrm{w}_0) + (4\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right] + \eta^2)\log(T)} + 2\eta n\right)\right\rceil^{\frac{1}{q+1}} T^{\frac{q}{1+q}},$$

where we define $q = 1 - \frac{\lambda}{H} \in (0,1)$. As we are consider small $T \lesssim n$, we have $t_0 \leq n$. Plugging t_0 back gives us that

$$\mathbb{E}_{\xi,\mathcal{S}}\left[L_{\text{adv}}(\mathrm{w}_{T,\mathcal{S}}) - \widehat{L}_{\text{adv}}(\mathrm{w}_{T,\mathcal{S}})\right]$$
$$\leq \left(1 + \frac{1}{q}\right)\left[\frac{\mathbb{E}_{\mathcal{S},\xi}\left[L_{\text{adv}}(\mathrm{w}_{T,\mathcal{S}})\right]T}{n}\right]^{\frac{q}{q+1}}\left[\frac{4Lq}{H-\lambda}\left(\frac{\sqrt{2H\Phi_{\text{adv}}(\mathrm{w}_0) + (4\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right] + \eta^2)\log(T)}}{n} + 2\eta\right)\right]^{\frac{1}{q+1}}$$
$$\leq \left(1 + \frac{1}{q}\right)\left[\frac{\mathbb{E}_{\mathcal{S},\xi}\left[L_{\text{adv}}(\mathrm{w}_{T,\mathcal{S}})\right]T}{n}\right]^{\frac{q}{q+1}}\left[\frac{4Lq}{H-\lambda}\left(\frac{\sqrt{2H\Phi_{\text{adv}}(\mathrm{w}_0) + (4\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right] + 4H_z^2\alpha^2)\log(T)}}{n} + 4H_z\alpha\right)\right]^{\frac{1}{q+1}}$$

Applying Lemma B.6 with

$$x = \mathbb{E}_{\xi,\mathcal{S}}L_{\text{adv}}(\mathrm{w}_{T,\mathcal{S}}), y = \mathbb{E}_{\xi,\mathcal{S}}\widehat{L}_{\text{adv}}(\mathrm{w}_{T,\mathcal{S}}), b = \frac{q}{q+1}$$
$$a = \left(1 + \frac{1}{q}\right)\left(\frac{T}{n}\right)^{\frac{q}{q+1}}\left[\frac{4Lq}{H-\lambda}\left(\frac{\sqrt{2H\Phi_{\text{adv}}(\mathrm{w}_0) + (4\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right] + 4H_z^2\alpha^2)\log(T)}}{n} + 4H_z\alpha\right)\right]^{\frac{1}{q+1}}$$

gives us that

$$\mathbb{E}_{\xi,\mathcal{S}}\left[L_{\text{adv}}(\mathrm{w}_{T,\mathcal{S}}) - \widehat{L}_{\text{adv}}(\mathrm{w}_{T,\mathcal{S}})\right]$$
$$\leq \max\left\{2^q a^{q+1}, a(2y)^{\frac{q}{q+1}}\right\}$$

$$= \max\left\{ \left(1+\frac{1}{q}\right)^{q+1} \left(\frac{2T}{n}\right)^{q} \frac{4Lq}{H-\lambda} \left(\frac{\sqrt{2H\Phi_{\text{adv}}(\mathrm{w}_0) + (4\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right] + 4H_z^2\alpha^2)\log\left(T\right)}}{n} + 4H_z\alpha \right), \right.$$
$$\left. \left(1+\frac{1}{q}\right) \left(\frac{2T\mathbb{E}_{\xi,\mathcal{S}}\widehat{L}_{\text{adv}}(\mathrm{w}_{T,\mathcal{S}})}{n}\right)^{\frac{q}{q+1}} \left[\frac{4Lq}{H-\lambda} \left(\frac{\sqrt{2H\Phi_{\text{adv}}(\mathrm{w}_0) + (4\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right] + 4H_z^2\alpha^2)\log\left(T\right)}}{n} + 4H_z\alpha \right)\right]^{\frac{1}{q+1}} \right\}$$
$$= 4\max\left\{ \left(\frac{2T}{n}\right)^{q} \frac{4L}{H(1-q)} \left(\frac{\sqrt{2H\Phi_{\text{adv}}(\mathrm{w}_0) + (4\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right] + 4H_z^2\alpha^2)\log\left(T\right)}}{n} + 4H_z\alpha \right), \right.$$
$$\left. \left(\frac{2T\mathbb{E}_{\xi,\mathcal{S}}\widehat{L}_{\text{adv}}(\mathrm{w}_{T,\mathcal{S}})}{n}\right)^{\frac{q}{q+1}} \left[\frac{4L}{H(1-q)} \left(\frac{\sqrt{2H\Phi_{\text{adv}}(\mathrm{w}_0) + (4\mathbb{E}_{\mathcal{S}}\left[\nu_{\mathcal{S}}^2\right] + 4H_z^2\alpha^2)\log\left(T\right)}}{n} + 4H_z\alpha \right)\right]^{\frac{1}{q+1}} \right\}$$

where the last line holds because $\left(1+\frac{1}{q}\right)^{q+1} q$ and $\left(1+\frac{1}{q}\right) q^{\frac{1}{q+1}}$ are bounded when $0 < q < 1$.

□