

# Blessings and Curses of Covariate Shifts: Adversarial Learning Dynamics, Directional Convergence, and Equilibria

Tengyuan Liang

*Booth School of Business*

*University of Chicago*

*Chicago, IL 60637, USA*

TENGYUAN.LIANG@CHICAGOBOOTH.EDU

## Abstract

Covariate distribution shifts and adversarial perturbations present robustness challenges to the conventional statistical learning framework: mild shifts in the test covariate distribution can significantly affect the performance of the statistical model learned based on the training distribution. The model performance typically deteriorates when extrapolation happens: namely, covariates shift to a region where the training distribution is scarce, and naturally, the learned model has little information. For robustness and regularization considerations, adversarial perturbation techniques are proposed as a remedy; however, careful study needs to be carried out about what extrapolation region adversarial covariate shift will focus on, given a learned model. This paper precisely characterizes the extrapolation region, examining both regression and classification in an infinite-dimensional setting. We study the implications of adversarial covariate shifts to subsequent learning of the equilibrium—the Bayes optimal model—in a sequential game framework. We exploit the dynamics of the adversarial learning game and reveal the curious effects of the covariate shift to equilibrium learning and experimental design. In particular, we establish two directional convergence results that exhibit distinctive phenomena: (1) a blessing in regression, the adversarial covariate shifts in an exponential rate to an optimal experimental design for rapid subsequent learning; (2) a curse in classification, the adversarial covariate shifts in a subquadratic rate to the hardest experimental design trapping subsequent learning.

**Keywords:** covariate distribution shift, adversarial learning, experimental design, directional convergence, dynamics, equilibria.

## 1 Introduction

In supervised learning, a folklore rule is that the test data set should follow the same, or at least resemble the probability distribution from which the training data set is drawn, for strong guarantees of learnability and generalization (Vapnik, 1999). The reason is grounded since, if not, either (a) concept shift, namely the conditional distribution of  $P(Y|X)$  changes, hence the underlying Bayes optimal prediction model  $f^* : X \rightarrow Y$  could shift, or (b) covariate shift, namely the covariate distribution  $\mu \in \mathcal{P}(X)$  shifts so that the underlying evaluation metric for the learned model  $f$  changes.<sup>1</sup> Nevertheless, supervised learning is often deployed “in the wild,” meaning the test data distribution typically extrapolates the training data distribution.

Concept shift is inherently a complex problem, as if shooting for a moving target. However, covariate shift may be less severe a problem if the concept stays the same. Historically,

---

1. One typical evaluation metric is  $\|f - f^*\|_{L^2(\mu)}$ , where the underlying covariate distribution  $\mu \in \mathcal{P}(X)$  varies.

specific statistical methods allow for mild extrapolation,<sup>2</sup> say the (fixed-design) linear regression and the (local) nonparametric regression (Stone, 1980). Recently, a few notable lines of work have arisen to study the covariate shift. Shimodaira (2000), Sugiyama et al. (2007), and Sugiyama and Mueller (2005) studied covariate shift adaption assuming knowledge of the density ratio of the covariate shift. Ben-David et al. (2006) and Blitzer et al. (2007) initiated the learning-theoretic study of domain adaptation. There, the generalization error on the target/test domain is bounded by the standard generalization error on the source/training domain, plus a discrepancy term—for instance, total variation or its tighter analog induced by the hypothesis class—quantifying the covariate shift between training and test distributions. Later, on the one hand, a collection of research extended the theory to allow for concept shifts with general loss functions by proposing other notions of discrepancy measures or quantities contrasting two distributions (Mansour et al., 2009; Ben-David et al., 2010a; Mohri and Muñoz Medina, 2012; Kpotufe and Martinet, 2018). On the other hand, by establishing lower bounds, Ben-David et al. (2010b) studied assumptions on the relationship between training and test distributions necessary for successful domain adaptation. The learning-theoretic framework brought forth new domain adaptation algorithms, for instance, reweighing the empirical distribution to minimize the discrepancy between source and target (Mansour et al., 2009), and finding common representation space with small discrepancy while maintaining good performance on the training data (Ben-David et al., 2006; Ganin et al., 2016). A considerable body of domain adaptation literature primarily focuses on when the conditional relationship  $P(Y|X)$  is invariant, and thus the Bayes optimal model stays fixed, yet the covariate distribution  $P(X)$  shifts. We follow this tradition of an invariant Bayes optimal model and investigate adversarial perturbations to the covariate distribution.

The quest for robust domain adaptation also prompted the recent development in adversarial learning (Ganin et al., 2016; Goodfellow et al., 2014; Ilyas et al., 2019; Bubeck et al., 2021). Akin to covariate shift, adversarial perturbations have recently revived interest in machine learning and robust optimization communities (Goodfellow et al., 2014; Delage and Ye, 2010). Adversarial perturbation is motivated by the following observation: small local perturbations to the covariate distribution can significantly compromise the supervised learning performance (Goodfellow et al., 2014; Madry et al., 2017; Javanmard and Soltanolkotabi, 2022). For example, given a supervised learning model  $f$ , adversarial perturbations shift the covariate distribution  $\mu \in \mathcal{P}(X)$  locally under a specific metric on the measure space  $\mathcal{P}(X)$ , such that it makes the model suffer the most in predictive performance. The familiar reader will immediately identify the minimax game between the supervised learning model  $f$  and the covariate distribution  $\mu$ : the model  $f$  minimizes the risk from a model class, yet the covariate distribution  $\mu$  maximizes the risk from a probability distribution class. Such a game perspective between the learning model  $f$  and data distribution  $\mu$  has been influential since the seminal work of boosting (Freund and Schapire, 1997, 1999). Inspired by the above, we study covariate shift from a game-theoretic perspective. The connection between boosting and adversarial perturbation will be elaborated later; in a nutshell, instead of taking Kullback-Leibler divergence as a metric, we use the

---

2. Here we mean the region of the extrapolation is still contained in the region of the seen data, but the test distribution can differ from the training distribution.

Wasserstein metric for adversarial perturbation, thus allowing for extrapolation outside the current covariate support.

This paper studies a particular form of covariate shifts following adversarial perturbations, with the underlying concept (i.e., Bayes optimal model) held fixed. We take a game-theoretic view to examine covariate shifts and discover curious insights. We study both regression and classification in an infinite-dimensional setting. As hinted, we will exploit the dynamics of the adversarial learning game and reveal the curious effects of the covariate shift to subsequent learning and experimental design. The models we study are discriminative in nature rather than generative<sup>3</sup> for invariance considerations: for the latter, the underlying concept  $P(Y|X)$  could shift as a result of adversarial perturbations on covariates.

Now we are ready to state the main goal of this paper:

*Adversarial covariate shifts move the current covariate domain to an extrapolation region. We precisely characterize the extrapolation region and, subsequently, the implications of adversarial covariate shifts to subsequent learning of the equilibrium, the Bayes optimal model.*

Curiously, we show two directional convergence results that exhibit distinctive phenomena: (1) a blessing in regression, the adversarial covariate shifts in an exponential rate to an optimal experimental design for rapid subsequent learning, (2) a curse in classification, the adversarial covariate shifts in a subquadratic rate to the hardest experimental design trapping subsequent learning. The theoretical results will be later coupled with numerical validations. The theoretical study is admittedly based on idealized models to demonstrate clean new insights and curious dichotomy on covariate shift and adversarial learning; potential future directions will be discussed in the last section. Before diving into the problem setup, we elaborate on the background and some related literature.

### 1.1 Background and Literature Review

We first fix some notations to make the discussions on covariate shift and adversarial perturbation concrete. Let  $X$  be space for the covariates, and  $Y \subset \mathbb{R}$  be the space for a real-valued response variable. When a pair of covariate and response data  $(x, y) \in X \times Y$  is generated based on the probability measure  $\pi \in \mathcal{P}(X \times Y)$ , we denote  $(x, y) \sim \pi$ . Let  $f : X \rightarrow Y$  be a statistical model and  $\ell(\cdot, \cdot) : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  be a risk or loss function ( $f(x), y \mapsto \ell(f(x), y)$ ) that quantifies how the model  $f$  performs on the data pair  $(x, y)$ .

Given a statistical model  $f$  and a probability measure for data set  $\pi$ , one can define the utility function that accesses the risk of model  $f$  on data  $\pi$

$$\mathcal{R}(f, \pi) := \mathbb{E}_{(x,y) \sim \pi} [\ell(f(x), y)] .$$

**Models, covariate distributions, and Bayes optimality** Following the literature (Shimodaira, 2000; Quinero-Candela et al., 2008), we consider the covariate shift but not the concept shift. For a valid marginal probability measure for the covariate  $\mu \in \mathcal{P}(X)$ ,

---

3. Here discriminative refers to modeling  $Y|X$ , and generative refers to modeling  $X|Y$ .

we define the induced joint measure for the covariate and response pair

$$\pi_\mu := \int \delta_x \otimes \pi_x^* d\mu(x) \in \mathcal{P}(X \times Y), \quad (1)$$

where  $\pi_x^* \in \mathcal{P}(Y)$  denotes a fixed conditional data generating process for  $\mathbf{y}|\mathbf{x} = x$  that does not vary with  $\mu$ , and  $\delta_x$  denotes the delta measure at point  $x$ . Equation (1) should be read as disintegration of measure (Villani, 2021), meaning for all bounded continuous function  $h \in C_b(X \times Y)$

$$\int_{X \times Y} h(x, y) d\pi_\mu(x, y) = \int_X \left[ \int_Y h(x, y) d\pi_x^*(y) \right] d\mu(x).$$

Given a fixed conditional distribution  $\pi_x^*$  and a loss function  $\ell(\cdot, \cdot)$ , one can thus define the Bayes optimal model (suppose for now that this map is well-defined)

$$f_{\text{Bayes}}^* : x \mapsto \arg \min_{y' \in Y} \int \ell(y', y) d\pi_x^*(y). \quad (2)$$

Observe that the Bayes optimal model does not change with  $\mu$ , the distribution of covariates  $x$ .

Now, we can define the objective of the game between the model  $f : X \rightarrow Y$  and the covariate distribution  $\mu \in \mathcal{P}(X)$ ,

$$\mathcal{U}(f, \mu) := \mathcal{R}(f, \pi_\mu) = \int_X \left[ \int_Y \ell(f(x), y) d\pi_x^*(y) \right] d\mu(x). \quad (3)$$

It turns out Bayes optimal model  $f_{\text{Bayes}}^*$  is an equilibrium of the game, as we shall show shortly.

Note that classical statistical learning theory studies  $\mathcal{U}(\hat{f}_\mu, \mu)$  where  $\hat{f}_\mu$  is learned based on an empirical data set drawn from the same distribution  $\pi_\mu$ . However, when the covariate distribution shift to another measure  $\nu$  that is different from the training data distribution  $\mu$ , the performance  $\mathcal{U}(\hat{f}_\mu, \nu)$  deteriorates, see Shimodaira (2000) and Quinero-Candela et al. (2008) for a review on covariate shifts. Assuming the knowledge of the density ratio  $d\nu/d\mu$ , importance weighting methods have been proposed as an adaptation to covariate shifts.

**Adversarial perturbation** The theoretical insights toward understanding adversarial perturbations have so far centered around robustness and regularization in various formulations, see Xu et al. (2009) (Theorem 3) for support vector machines, Ross and Doshi-Velez (2018); Madry et al. (2017); Sinha et al. (2017) for neural network models, and Delage and Ye (2010) for distributionally robust optimization. Given a metric measure space  $(\mathcal{P}(X), d)$ , a covariate distribution  $\mu \in \mathcal{P}(X)$ , and a current model  $f$ , consider the following population version of the adversarial perturbation,

$$\mathcal{U}_\gamma(f, \mu) := \max_{\nu : d^2(\nu, \mu) \leq \gamma} \mathcal{U}(f, \nu), \quad (4)$$

where  $\mathcal{U}(\cdot, \cdot)$  is defined in (3).

Adversarial perturbation can be viewed as smoothly regularizing the original loss function, thus enforcing stability. To see this, consider the Wasserstein metric  $W_2$ ; as done in Sinha et al. (2017) (Proposition 1), one can write the Lagrangian of (4) and characterize the coupling analytically

$$\max_{\lambda \geq 0} \min_{\nu \in \mathcal{P}(X)} -\mathcal{U}(f, \nu) + \lambda [W_2^2(\nu, \mu) - \gamma] = \max_{\lambda \geq 0} \mathbb{E}_{x \sim \mu} \left[ \underbrace{\min_{x' \in X} \left( -\mathcal{U}(f, \delta_{x'}) + \frac{\lambda}{2} \|x' - x\|^2 \right)}_{\text{Moreau–Yosida regularization}} \right] - \gamma \lambda,$$

where  $\delta_x$  denotes the delta measure at point  $x$ . Both the robustness and regularization perspectives readily unveil, as the above is the Moreau–Yosida envelope of the function  $-\mathcal{U}(f, \delta_x) : X \rightarrow \mathbb{R}$  with parameter  $\lambda^{-1}$ , thus serving as a smoothed regularization to the original loss. We refer the readers to Sinha et al. (2017) for detailed derivations.

The adversarial perturbation provides a robust notion of covariate shifts without requiring the explicit knowledge of density ratio. Perhaps more importantly, it extends to the extrapolation case when the support of  $\nu$  differs from  $\mu$ . The literature on adversarial learning is growing too fast to give a complete review. To name a few: Bubeck et al. (2021); Bartlett et al. (2021) studied adversarial examples using gradient steps for two/multi-layer ReLU networks with Gaussian weights; for regression, Javanmard et al. (2020) studied precise tradeoffs between adversarial risk  $\mathcal{U}_\gamma(\hat{f}, \mu)$  and standard risk  $\mathcal{U}(\hat{f}, \mu)$  for a range of models  $\hat{f}$  interpolating between empirical risk minimization and adversarial risk minimization, Xing et al. (2021) studied properties of the adversarially robust estimate; for classification, Bao et al. (2020) introduced surrogate losses that are calibrated with the adversarial 0-1 loss, Hu et al. (2018) identified certain failure mode of distributionally robust classification under  $f$ -divergences, Javanmard and Soltanolkotabi (2022) precisely characterized the adversarial 0-1 loss with Gaussian covariate distributions.

**Wasserstein gradient flow** Adversarial distribution shift is inherently connected to Wasserstein gradient flow. Given a current model  $f$ , the covariate distribution is perturbed incrementally within a Wasserstein ball in an adversarial way, with a stepsize  $\gamma \in \mathbb{R}_+$ ,

$$\nu := \arg \min_{\nu \in \mathcal{P}(X)} -\mathcal{U}(f, \nu) + \frac{1}{\gamma} W_2^2(\nu, \mu). \tag{5}$$

Denote the distribution shift map  $\text{Ds}_\gamma : x \mapsto \arg \min_{x' \in X} \left( -\mathcal{U}(f, \delta_{x'}) + \frac{1}{2\gamma} \|x' - x\|^2 \right)$  defined by the Moreau–Yosida envelope. Informally, such a map defines the worst-case covariate shift for the model  $f$  evaluated at measure  $\mu$ , as the maximizer of the adversarial perturbation is attained at

$$\nu = (\text{Ds}_{\lambda^{-1}})_\# \mu$$

where  $\lambda > 0$  is the solution to the dual. In the infinitesimal limit  $\gamma \asymp \lambda^{-1} \rightarrow 0$ , one can show that (Ambrosio et al., 2005; Guo et al., 2022)

$$\frac{(\text{Ds}_\gamma - \text{Id})[x]}{\gamma} \rightarrow \frac{\partial}{\partial x} \mathcal{U}(f, \delta_x). \tag{6}$$

The distribution shift map  $\text{Ds}_\gamma$  presents a way of constructing adversarial examples (Goodfellow et al., 2014; Ilyas et al., 2019; Bubeck et al., 2021), namely a couple  $x \approx x'$  such that  $\mathcal{U}(f, \delta_{x'}) - \mathcal{U}(f, \delta_x)$  is large.

The continuous-time analog of the adversarial perturbation (5) is called the Wasserstein gradient flow as  $\gamma \rightarrow 0$ , where the density  $\rho_t$  (associated with  $\nu_t$ ), w.r.t. the Lebesgue measure, evolves according to the following PDE (Ambrosio et al., 2005)

$$\partial_t \rho_t + \nabla \cdot (\rho_t V) = 0, \text{ where } V : x \mapsto \frac{\partial}{\partial x} \mathcal{U}(f, \delta_x). \quad (7)$$

## 1.2 Problem Setup

This paper considers regression and classification problems in an infinite-dimensional setting. Let  $X \subset \mathbb{R}^{\mathbb{N}}$  be a subset of a possibly infinite-dimensional space for the covariates, and  $Y \subset \mathbb{R}$  be the space for a real-valued response variable. We concern an infinite-dimensional linear model class  $\mathcal{F} := \{f_\theta \mid f_\theta(x) := \langle x, \theta \rangle, \theta \in \ell_{\mathbb{N}}^2\}$  where the inner-product corresponds to the Hilbert space  $\ell_{\mathbb{N}}^2$ . Slightly abusing the notation, we write for convenience the utility function

$$\mathcal{U}(\theta, \mu) = \mathbb{E}_{(x,y) \sim \pi_\mu} [\ell(f_\theta(x), y)] = \int_X \left[ \int_Y \ell(f_\theta(x), y) d\pi_x^*(y) \right] d\mu(x). \quad (8)$$

We investigate two types of conditional relationships for  $\pi_x^*$  in (1), namely for some  $\theta^* \in \ell_{\mathbb{N}}^2$ :

Regression:  $\mathbf{y} | \mathbf{x} = x \sim \text{Gaussian}(\langle x, \theta^* \rangle, 1)$ ,  $\ell(f, y) = (f - y)^2$ ;

Classification:  $\mathbf{y} | \mathbf{x} = x \sim \text{Bernoulli}(\sigma(\langle x, \theta^* \rangle))$ ,  $\ell(f, y) = -fy + \log(1 + e^f)$ .

Here  $\sigma(z) = 1/(1 + e^{-z})$  is the sigmoid function. Note that in the classification setup,  $y \in \{0, 1\}$ . In both the regression and classification settings we study, the Bayes optimal model (2) is uniquely defined

$$f_{\text{Bayes}}^*(x) = \langle x, \theta^* \rangle.$$

**Game and equilibria** In the game between the model  $\theta \in \ell_{\mathbb{N}}^2$  and the covariate distribution  $\mu \in \mathcal{P}(X)$ ,

$$\begin{aligned} \min_{\theta} \max_{\mu} \mathcal{U}(\theta, \mu) &\geq \max_{\mu} \min_{\theta} \mathcal{U}(\theta, \mu) \geq \max_{\mu} \int_X \left[ \min_{\theta \in \ell_{\mathbb{N}}^2} \int_Y \ell(f_\theta(x), y) d\pi_x^*(y) \right] d\mu(x) \quad (9) \\ &= \max_{\mu} \mathcal{U}(\theta^*, \mu) \geq \min_{\theta} \max_{\mu} \mathcal{U}(\theta, \mu). \end{aligned}$$

Therefore, the min-max theorem holds in this infinite-dimensional context when the Bayes optimal model  $f_{\text{Bayes}}^*(x) = \langle x, \theta^* \rangle$  is well-defined. A Nash equilibrium of  $\mathcal{U}(\cdot, \cdot)$  is precisely the Bayes optimal model  $f_{\text{Bayes}}^*$ . In plain language, the covariate distribution shift does not affect the notion of equilibrium of the game.

The game perspective is not new. For example, the celebrated boosting literature (Freund and Schapire, 1997, 1999; Telgarsky, 2013; Liang and Sur, 2022) is precisely harnessing the duality between a linear predictive model (aggregating  $p$  weak learners) indexed by  $\theta \in \mathbb{R}^p$  and a finitely supported data distribution (with cardinality  $n$ ) parametrized by a weight on the probability simplex  $\mu \in \Delta_n$ . There, rather than adversarially perturbing data

using the Wasserstein metric, the probability weight vector  $\mu$ —and its induced joint distribution  $\pi_\mu = \sum_{i=1}^n \mu_i \delta_{(x_i, y_i)}$ —is perturbed as in (5) under the Kullback-Leibler divergence. A crucial difference between Wasserstein and Kullback-Leibler is that in the latter case, only the weights are allowed to vary but not the domain. Another analogy is regarding the equilibrium concept: when the data set is linearly separable, the equilibrium concept for boosting is the max-margin solution; for our problem, the equilibrium concept is the Bayes optimal solution. The game perspective is also instrumental to the generative modeling and adversarial learning literature (Goodfellow et al., 2020; Dziugaite et al., 2015; Daskalakis et al., 2017; Liang, 2021; Liang and Stokes, 2019; Mokhtari et al., 2020), where the duality between the probability distribution given by the generative model and discriminative function is leveraged.

**Best response and information sets** Given a covariate distribution  $\mu^{(0)}$  whose support is on a linear subspace and does not span the full space  $\text{supp}(\mu^{(0)}) \subset X = \ell_N^2$  (so that extrapolation is meaningful), the best response model  $f_{\theta^{(0)}} \in \mathcal{F}$  solves the following risk minimization associated with measure  $\mu^{(0)}$

$$\theta^{(0)} \in \text{BR}(\mu^{(0)}) := \arg \min_{\theta \in \ell_N^2} \mathcal{U}(\theta, \mu^{(0)}) .$$

In both the Gaussian and Bernoulli conditional models, the best response model  $\theta^{(0)}$  associated with the measure  $\mu^{(0)}$  takes the form

$$\theta^{(0)} \in \text{BR}(\mu^{(0)}) = \left\{ \Pi_{\text{supp}(\mu^{(0)})} \theta^* + \Pi_{\text{supp}(\mu^{(0)})}^\perp \xi \mid \forall \xi \in \ell_N^2 \right\} ,$$

namely, projected to the linear subspace spanned by  $\text{supp}(\mu^{(0)}) \subset X$ , the perceived best response model  $\theta^{(0)}$  collides the Bayes optimal model  $\Pi_{\text{supp}(\mu^{(0)})} \theta^*$ , while on the orthogonal domain  $\Pi_{\text{supp}(\mu^{(0)})}^\perp$  no information is learned.

The minimum Hilbert space norm solution in the over-identified set  $\text{BR}(\mu^{(0)})$  is  $\Pi_{\text{supp}(\mu^{(0)})} \theta^*$ . Clearly,  $\theta^{(0)} = \Pi_{\text{supp}(\mu^{(0)})} \theta^*$  is inconsistent with the Bayes optimal model  $\theta^*$  (Shimodaira, 2000). It is, therefore, natural to consider, for typical adversarial distribution shifts  $\mu^{(0)} \rightarrow \mu^{(1)}$ , how does the information set  $\text{BR}(\mu^{(1)})$  differ from  $\text{BR}(\mu^{(0)})$ ? The information set question is useful to understand whether  $\theta^{(1)}$  improves upon  $\theta^{(0)}$  in approaching the Bayes optimal model  $\theta^*$ . To answer this, we will probe precisely how the  $\text{supp}(\mu^{(1)})$  varies from  $\text{supp}(\mu^{(0)})$  for natural adversarial covariate shifts: what extrapolation regions  $\text{supp}(\mu^{(1)})$  focus on.

**Adversarial dynamics** For the adversarial distribution shifts, we follow the Wasserstein gradient flow setup, given a current model  $\theta^{(0)}$ , the covariate distribution is perturbed incrementally within a Wasserstein ball in an adversarial way: with a stepsize  $\gamma \in \mathbb{R}_+$ , initialize  $\nu_0 := \mu^{(0)}$ ,

$$\nu_{t+1} := \arg \min_{\nu \in \mathcal{P}(X)} -\mathcal{U}(\theta^{(0)}, \nu) + \frac{1}{\gamma} W_2^2(\nu, \nu_t) , \text{ for } t = 0, 1, \dots, T ,$$

and then set  $\mu^{(1)} := \nu_{T+1}$ . The continuous analog of the adversarial perturbation is called the Wasserstein gradient flow as  $\gamma \rightarrow 0$ , where the density  $\rho_t$  (associated with  $\nu_t$ ) evolves

according to the following PDE

$$\partial_t \rho_t + \nabla \cdot (\rho_t \mathbf{V}) = 0, \text{ where } \mathbf{V} : x \mapsto \frac{\partial}{\partial x} \mathcal{U}(\theta^{(0)}, \delta_x). \quad (10)$$

Conceptually, the adversarial distribution shift is a gradient ascent flow on the Wasserstein space  $(\mathcal{P}(X), W_2)$  defined in (10) with effective time  $\gamma T$ . In this paper, we study a discretization of (10) with stepsize  $\gamma$  and iterations  $T$ , as follows

$$x_{t+1} = x_t + \gamma \cdot \frac{\partial}{\partial x} \mathcal{U}(\theta^{(0)}, \delta_x)|_{x=x_t}, \text{ for } t = 0, 1, \dots, T, \text{ where } x_0 \sim \mu^{(0)}. \quad (11)$$

## 2 Main Results

### 2.1 Adversarial Covariate Shifts: Blessings and Curses

Let  $\theta^{(0)} \in \ell_{\mathbb{N}}^2$  be the current learning model and  $\theta^* - \theta^{(0)}$  be the remaining signal to be identified. Let  $\ell_{\mathbb{N}}^2(1)$  denote the unit norm ball. We now define two unit-norm directions: the blessing direction  $\Delta_b \in \mathbb{R}^{\mathbb{N}}$  and the curse direction  $\Delta_c \in \mathbb{R}^{\mathbb{N}}$

$$\Delta_b := \frac{\theta^* - \theta^{(0)}}{\|\theta^* - \theta^{(0)}\|} \in \ell_{\mathbb{N}}^2(1), \quad (12)$$

$$\Delta_c := -\frac{\|\theta^{(0)}\|}{\|\theta^*\|} \cdot \frac{\theta^* - \theta^{(0)}}{\|\theta^* - \theta^{(0)}\|} + \frac{\|\theta^* - \theta^{(0)}\|}{\|\theta^*\|} \cdot \frac{\theta^{(0)}}{\|\theta^{(0)}\|} \in \ell_{\mathbb{N}}^2(1). \quad (13)$$

The name blessing comes from the fact that  $\Delta_b // \theta^* - \theta^{(0)}$  as stated in Theorem 1, namely, the direction is parallel to the remaining signal direction, the most informative signal direction given the current model  $\theta^{(0)}$ .

The name curse arises as  $\Delta_c \perp \theta^*$  under the assumption in Theorem 4, that is, the direction is perpendicular to the signal direction. To see why this assumption makes sense, we recall that the best response model  $\theta^{(0)}$  associated with the measure  $\mu^{(0)}$  takes the form  $\theta^{(0)} \in \text{BR}(\mu^{(0)}) = \left\{ \Pi_{\text{supp}(\mu^{(0)})} \theta^* + \Pi_{\text{supp}(\mu^{(0)})}^{\perp} \xi \mid \forall \xi \in \ell_{\mathbb{N}}^2 \right\}$ . The minimum-norm solution for the best response set satisfies the assumption in Theorem 4, namely  $\theta^{(0)} \perp \theta^* - \theta^{(0)}$  and therefore  $\Delta_c \perp \theta^*$ .

Curiously, we will show that the adversarial learning dynamic collapses to a probability measure along the blessing direction  $\Delta_b$  in the regression problem; in sharp contrast, the probability measure induced by the adversarial learning dynamic converges to the curse direction  $\Delta_c$  in the classification problem. The formal directional convergence results are stated in Theorems 1 and 4. For the flow of the exposition, the primary intuition of the proof is deferred to Section 2.4. Appendix A collects the detailed proof of all theorems.

We first state the result for the infinite-dimensional regression problem. As a reminder, all the relevant notations in Theorems 1 and 4 were introduced in Section 1.2.

**Theorem 1 (Regression: directional convergence)** *Consider the regression setting where  $\ell(y', y) = (y' - y)^2$  and  $\mathbf{y}|\mathbf{x} = x \sim \text{Gaussian}(\langle x, \theta^* \rangle, 1)$ . Let  $x_0 \in \text{supp}(\mu^{(0)})$  that satisfies  $\langle x_0, \theta^* - \theta^{(0)} \rangle \neq 0$ . Then the induced adversarial distribution shift dynamic (11) satisfies*

$$\lim_{T \rightarrow \infty} \left| \left\langle \frac{x_T}{\|x_T\|}, \Delta_b \right\rangle \right| = 1, \text{ where } \Delta_b // \theta^* - \theta^{(0)} \text{ is defined in (12)}. \quad (14)$$

Moreover, the directional convergence is exponential in  $T$ ,

$$\left| \left\langle \frac{x_T}{\|x_T\|}, \Delta_b \right\rangle \right| \in [1 - O\left(\frac{1}{e^{cT}}\right), 1] ,$$

where  $c = 2 \log(1 + 2\gamma \|\theta^* - \theta^{(0)}\|^2)$ .

**Remark 2** This theorem concerns the case when the current model  $\theta^{(0)}$  is imperfect—namely  $\|\theta^* - \theta^{(0)}\| \neq 0$ —the only case when distribution shifts that vary  $\mu \in \mathcal{P}(X)$  could impact learning the conditional relationship  $\pi_x^*$  and hence identifying the equilibrium  $f_{\text{Bayes}}^*$  defined in (2). The current model could be imperfect due to (a)  $\text{supp}(\mu^{(0)}) \subsetneq X$  the covariate does not span the full infinite-dimensional space, or (b) the learner only has finite sample access to the measure  $\pi_{\mu^{(0)}} \in \mathcal{P}(X \times Y)$ . The theorem states that the adversarial distribution shift dynamics  $\mu^{(0)} \rightarrow \mu^{(1)}$  align all the mass of the covariates along the most informative direction for the next stage of learning: the shifted distribution  $\mu^{(1)}$  is asymptotically a measure along a one-dimensional “blessing” direction  $\Delta_b$ , reducing the subsequent learning to a one-dimensional problem. Namely, the adversarial distribution shift asymptotically constructs the **optimal covariate design** for the next stage of learning: making the current model  $\theta^{(0)}$  suffer is revealing the information towards the equilibrium of learning, the Bayes optimal model  $\theta^*$ . The impact of the distribution shifts on the next stage learner, in this sequential game perspective, is formally stated in Theorem 6. The proof is based on power iterations as in principle component analysis.

Now, we state the result for the infinite-dimensional classification problem, which contrasts sharply with the regression problem. We start by stating the conditions and discuss the assumption before stating the theorem.

**Assumption 1 (Initial condition)** *Given a fixed  $r > 0$ , we call that two real numbers  $(a_0, b_0)$  satisfy the initial condition, if  $a_0 + b_0 < 0$  and*

$$\frac{e^{a_0+b_0} a_0}{1 - e^{2(a_0+b_0)}} < 1 , \tag{15}$$

$$\frac{e^{a_0+b_0} a_0}{1 + e^{a_0+b_0}} \geq \frac{1 + \frac{1}{a_0}}{1 + r + \frac{1}{a_0}} , \tag{16}$$

with  $a_0 > c$  for some large enough constant  $c > 0$ .

**Remark 3** When the initialization  $a_0 > c$  with  $c$  not too small, the assumption holds for a range of  $b_0$ : (15) and (16) are equivalent to

$$a_0 \frac{1 + \sqrt{1 + 4a_0^{-2}}}{2} < e^{-(a_0+b_0)} < a_0 \frac{1 + r + a_0^{-1}}{1 + a_0^{-1}} - 1 .$$

This is nonempty whenever  $(1 + \frac{r}{1+a_0^{-1}} - a_0^{-1})^2 - 1 - 4a_0^{-2} > 0$ , which is true for  $a_0$  not too small.

**Theorem 4 (Classification: directional convergence)** *Consider the classification setting where  $\ell(y', y) = -y'y + \log(1 + e^{y'})$  and  $\mathbf{y}|\mathbf{x} = x \sim \text{Bernoulli}(\sigma(\langle x, \theta^* \rangle))$ . Assume  $\theta^{(0)} \perp \theta^* - \theta^{(0)}$  and define  $r := \frac{\|\theta^* - \theta^{(0)}\|^2}{\|\theta^{(0)}\|^2} > 0$ . Consider fixed, small step-size in the adversarial distribution shift dynamic (11) that satisfies  $\eta := \gamma\|\theta^{(0)}\|^2 < \frac{1}{2(2+r)}$ . Let  $x_0 \in \text{supp}(\mu^{(0)})$  and assume there exists a finite  $t_0 \in \mathbb{N}$  with  $(a_0, b_0) := (\langle x_{t_0}, \theta^{(0)} \rangle, \langle x_{t_0}, \theta^* - \theta^{(0)} \rangle)$  satisfying Assumption 1 for constants  $r$  defined above and  $c > 1$ . Then*

$$\lim_{T \rightarrow \infty} \left| \left\langle \frac{x_T}{\|x_T\|}, \Delta_c \right\rangle \right| = 1, \text{ where } \Delta_c \perp \theta^* \text{ is defined in (13)}. \quad (17)$$

Moreover, the directional convergence is quadratic in  $T/\log(T)$ ,

$$\left| \left\langle \frac{x_T}{\|x_T\|}, \Delta_c \right\rangle \right| \in \left[ 1 - O\left(\frac{\log^2(T)}{T^2}\right), 1 \right].$$

**Remark 5** This theorem concerns also the case when the current model  $\theta^{(0)}$  is useful yet imperfect—namely  $\|\theta^* - \theta^{(0)}\| \neq 0$ , and  $\|\theta^*\| \neq 0$ . In particular, this theorem studies when  $\text{supp}(\mu^{(0)}) \subsetneq X$ . As stated before, the best response model  $\theta^{(0)}$  associated with the measure  $\mu^{(0)}$  takes the form  $\theta^{(0)} \in \text{BR}(\mu^{(0)}) = \left\{ \Pi_{\text{supp}(\mu^{(0)})} \theta^* + \Pi_{\text{supp}(\mu^{(0)})}^\perp \xi \mid \forall \xi \in \ell_{\mathbb{N}}^2 \right\}$ . The minimum-norm solution for the best response set satisfies the assumption in the theorem, namely  $\theta^{(0)} \perp \theta^* - \theta^{(0)}$ . The theorem is also about directional convergence: the adversarial distribution shift dynamics  $\mu^{(0)} \rightarrow \mu^{(1)}$  asymptotically align all the mass of the covariates along a one-dimensional, “curse” direction  $\Delta_c \perp \theta^*$ , orthogonal to the Bayes optimal model. This directional alignment will reduce the subsequent learning to a one-dimensional problem that is the hardest, namely, the  $(x, y) \sim \pi_{\mu^{(1)}}$  where  $y$  is a Bernoulli coin-flip that is independent of  $x$ ! Note the adversarial distribution shift (under the logistic loss) asymptotically constructs the **hardest covariate design** under the 0-1 loss, since the Bernoulli coin-flip is impossible to predict for the next stage of learning. Qualitatively, this contrasts sharply with the phenomenon in the regression setting. The adversarial distribution shift constructs a difficult covariate design, trapping the next stage of learning. Namely, making the current model  $\theta^{(0)}$  suffer is constructing the hardest experimental design for identifying the Bayes optimal model  $\theta^*$ . The impact of the distribution shifts on the next stage learner, in this sequential game perspective, is formally stated in Theorem 7. Quantitatively, the directional convergence to  $\Delta_c$  in the classification setting is quadratic in  $T$ , much slower than the directional convergence to  $\Delta_b$  in the regression setting, exponential in  $T$ . We invite the readers to Section 2.3 for a preliminary numerical experiment illustrating the sharp contrast of Theorems 1 and 4, visualized in Figure 1 and Figure 2. We find the contrast in the speed of directional convergence interesting in its own right.

Tracking down the exact behavior of the distribution shift dynamic (non-convex and non-linear) and establishing a sharp directional convergence rate are this paper’s main technical innovations and difficulties. To overcome these challenges, we carefully construct two bounding envelopes refined recursively to characterize the dynamics analytically. Section 2.4 elaborates on the main steps of the technical proof and key ideas.

Before studying the subsequent impact on the learner, we briefly discuss why a sharp dichotomy exists between regression and classification. A crucial factor is the loss function.

One intuition that helps shed light on the distinction between classification and regression can be seen in the following special case. Consider when the learner  $\theta^{(0)}$  has already learned the Bayes optimal function  $\theta^*$ : (a) for the regression problem, the covariate distribution will not shift, since the gradient flow for covariate distribution has velocity zero, as a result of the quadratic loss function; (b) in contrast, for the classification problem, the covariate distribution will further shift to decrease the margin so to decrease the signal-to-noise ratio, as a result of the logistic loss function. For the general case when  $\theta^* - \theta^{(0)} \neq 0$ , the explicit directional convergence phenomenon is curious and technically challenging to derive, which is the main focus of the current paper.

## 2.2 Impact on the Learner: Sequential Game Perspective

In this section, we investigate the impact of the covariate distribution shift on the next stage learner’s gradient descent dynamic. The goal here is to demonstrate that the directional convergence results established in Theorems 1 and 4 can translate to a direct impact on the learner’s subsequent learning towards the Bayes optimal model  $f_{\text{Bayes}}^*(x) = \langle x, \theta^* \rangle$ , the equilibrium. The impact is a dichotomy that either comes as a blessing or a curse.

The sequential game between model  $f_\theta$  and covariate distribution  $\mu$  evolves according to the following protocol. Here we only focus on one round, namely from stage  $t = 0$  to  $t = 1$ .

1. Adversarial covariate shift: the covariate distribution shifts from  $\widehat{\mu}^{(0)} \rightarrow \widehat{\mu}^{(1)}$  given the previous model  $\theta^{(0)}$ . Here  $\widehat{\mu}^{(0)} = \frac{1}{n} \sum_{i=1}^n \delta_{x_0^i / \|x_0^i\|}$  where  $x_0^i \sim \mu^{(0)}$ , and  $\widehat{\mu}^{(1)} := \frac{1}{n} \sum_{i=1}^n \delta_{x_T^i / \|x_T^i\|}$ , where  $x_T^i$  evolves according to (11) after  $T$  steps.
2. Learner’s subsequent action: the learner performs a one-step improvement using gradient descent given the shifted distribution  $\widehat{\mu}^{(1)}$ . Draw  $n$ -i.i.d. samples  $(x^i, y^i) \sim \pi_{\widehat{\mu}^{(1)}}$  defined in (1), and update

$$\theta^{(1)} = \theta^{(0)} - \eta \cdot \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} \ell(\langle x^i, \theta \rangle, y^i) |_{\theta = \theta^{(0)}} . \tag{18}$$

Note that  $\theta^{(1)}$  implicitly depends on  $(n, T, \eta)$  so we subsequently denote as  $\theta_{n,T,\eta}^{(1)}$ .

The curious reader may wonder about the renormalization such that  $\widehat{\mu}^{(1)}$  is a probability measure on the unit sphere. Note that this is for convenience of analysis and does not change the qualitative phenomenon. The assumption is currently used to invoke the Hanson–Wright inequality in the proofs of Theorems 6 and 7.

**Theorem 6 (Regression: blessing to the learner)** *Consider the same setting as in Theorem 1. For all  $\theta^{(0)}$  such that  $\|\theta^* - \theta^{(0)}\| \neq 0$ , the learner’s one-step reaction to the distribution shift as in (18) with  $\eta = 1/2$  satisfies*

$$\lim_{n \rightarrow \infty} \lim_{T \rightarrow \infty} \|\theta^* - \theta_{n,T,\eta}^{(1)}\| = 0 \quad a.s. . \tag{19}$$

**Theorem 7 (Classification: curse to the learner)** *Consider the same setting as in Theorem 4. For all  $\theta^{(0)}$  such that  $\langle \theta^* - \theta^{(0)}, \theta^* \rangle \neq 0$ , the learner’s one-step reaction to the distribution shift as in (18) with any fixed  $\eta > 0$  satisfies*

$$\lim_{n \rightarrow \infty} \lim_{T \rightarrow \infty} \frac{\langle \theta^* - \theta_{n,T,\eta}^{(1)}, \theta^* \rangle}{\langle \theta^* - \theta^{(0)}, \theta^* \rangle} = 1 .$$

Moreover,

$$\liminf_{n \rightarrow \infty} \lim_{T \rightarrow \infty} \|\theta^* - \theta_{n,T,\eta}^{(1)}\| > 0 . \quad (20)$$

**Remark 8** The result in Theorem 7 holds valid for any fixed number of gradient descent steps for the learner

$$\theta^{(t+1)} = \theta^{(t)} - \eta \cdot \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} \ell(\langle x^i, \theta \rangle, y^i) |_{\theta=\theta^{(t)}} .$$

Therefore, the learner gets stuck with the adversarial distribution  $\hat{\mu}^{(1)}$  in making progress towards  $\theta^*$ . Comparing Theorems 6 and 7, we see that the adversarial distribution shifts in the regression setting make the learner’s one-step subsequent move optimal! (19) shows that one-step improvement using gradient descent dynamic as in (18) will reach the Bayes optimal model, also the equilibrium to the minimax game as in (9). On the contrary, in the classification setting, (20) shows that subsequent learner’s move using gradient descent dynamic (regardless of the number of steps) will be trapped with no improvement, preventing the learner from reaching the Bayes optimal model.

### 2.3 Numerical Illustration

In this section, we provide two simple numerical simulations, one for regression and one for classification, to contrast the sharp differences of the directional convergence established in Theorems 1 and 4.

**Experiment setup** Both simulations are based on a high-dimensional setting where  $X = \mathbb{R}^{200}$  and  $\text{supp}(\mu^{(0)}) = \mathbb{R}^{100}$  a randomly drawn subspace from the Haar measure (a uniformly drawn subspace of dimension 100). Specify a Bayes optimal model  $\theta^* = [1, 1/2, \dots, 1/i, \dots, 1/200]^\top$ , denoted by the **Yellow  $\star$**  in the figures. The best response model restricted to the subspace  $\text{supp}(\mu^{(0)})$ ,  $\{\frac{\theta^{(0)}}{\|\theta^{(0)}\|}$  direction is noted by **Green  $\blacktriangle$** , and the remaining signal direction  $\frac{\theta^* - \theta^{(0)}}{\|\theta^* - \theta^{(0)}\|}$  is denoted by **Red  $\times$** . The probability distribution of the covariates is shown by the **Blue  $\bullet$** . For the experiment, we consider the case when  $\theta^{(0)} \perp \theta^* - \theta^{(0)}$ . To visualize the adversarial distribution shift on a two-dimensional plane, we project all points to the two-dimensional subspace with horizontal and vertical axes  $\{\frac{\theta^{(0)}}{\|\theta^{(0)}\|}, \frac{\theta^* - \theta^{(0)}}{\|\theta^* - \theta^{(0)}\|}\}$ . In each simulation, the adversarial distribution shift dynamic is visualized by the motion of the **Blue  $\bullet$**  data clouds. Since we focus on directional convergence, every data point is visualized by its direction, normalized to the unit ball in  $\mathbb{R}^{200}$ . Concretely, for each draw  $x_0 \sim \mu^{(0)}$ , we plot at each timestamp  $\frac{x_t}{\|x_t\|}$ , projected to the plane.



Figure 1: Regression setting, directional convergence. From left to right, top to bottom, we plot the directional information at timestamp  $t = 0, 5, 10, \dots, 40$ , once every 5 iterations.

**Directional convergence: regression** Consider (8) with  $\mathbf{y}|\mathbf{x} = x \sim \text{Gaussian}(\langle x, \theta^* \rangle, 1)$  and  $\ell(y', y) = (y' - y)^2$ . The adversarial distribution shift evolves according to (11). Here the blessing direction  $\Delta_b$  defined in (12) is precisely marked by **Red x**. As seen in Figure 1, the **Blue •** data clouds collapsed to a align perfectly to  $\Delta_b$ , rapidly. We emphasize that since the simulation is done in high dimensions, directional convergence  $|\langle \frac{x_t}{\|x_t\|}, \Delta_b \rangle| = 1$  is only true when the **Blue •** perfectly lands on  $\{\pm \Delta_b\}$ , shown in  $t = 40$  (the bottom right subfigure); just aligning to a direction in the two-dimensional domain does not imply directional convergence in  $\mathbb{R}^{200}$ .

**Directional convergence: classification** Consider (8) with  $\mathbf{y}|\mathbf{x} = x \sim \text{Bernoulli}(\sigma(\langle x, \theta^* \rangle))$  and  $\ell(y', y) = -y'y + \log(1 + e^{y'})$ . The adversarial distribution shift evolves according to (11). Here the curse direction  $\Delta_c$  defined in (13) is perpendicular to **Yellow \***. Seen in Figure 2, the **Blue •** data clouds eventually land on the directions  $\{\pm \Delta_c\}$ . Compared to Figure 1, the direction  $\{\pm \Delta_c\}$  is different from the regression case  $\{\pm \Delta_b\}$ , and the convergence is much slower ( $T^{-2} \log^2(T)$  vs.  $\exp(-T)$ ), as proved by Theorems 6 and 7. Again we emphasize that alignment to a direction in the two-dimensional domain does not imply directional convergence in  $\mathbb{R}^{200}$ :  $t = 50$  (the top right subfigure)  $|\langle \frac{x_t}{\|x_t\|}, \Delta_c \rangle| \neq 1$ ; only when  $t = 175$  (the bottom middle subfigure), we have roughly directional convergence  $|\langle \frac{x_t}{\|x_t\|}, \Delta_c \rangle| = 1$ .

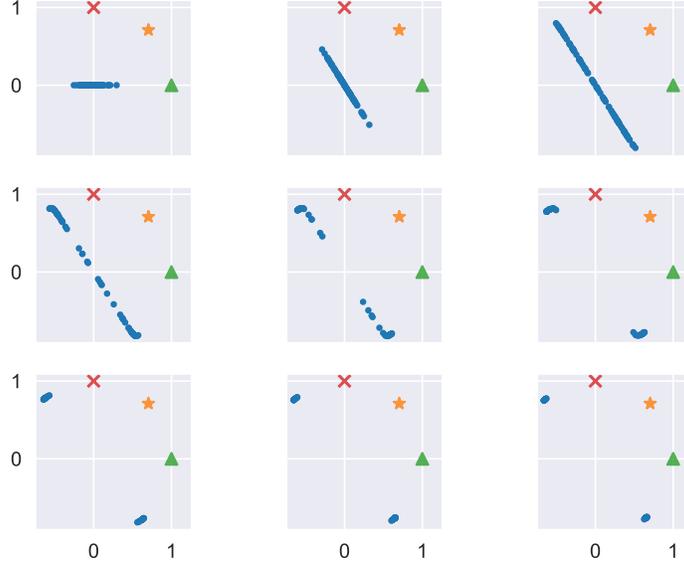


Figure 2: Classification setting, directional convergence. From left to right, top to bottom, we plot the directional information at timestamp  $t = 0, 25, 50, \dots, 200$ , once every 25 iterations.

## 2.4 Intuition of the Proof

In this section, we elaborate on the intuition and technical innovation behind the proof of Theorem 4. Let  $\sigma(z) = 1/(1 + e^{-z})$  be the sigmoid function and  $\sigma'(z) = \sigma(z)(1 - \sigma(z))$ . The dynamic of the distribution shift is non-convex and non-linear, following the equation

$$x_{t+1} = x_t + \gamma \cdot \left[ -\sigma(\langle x_t, \theta^* \rangle) (1 - \sigma(\langle x_t, \theta^* \rangle)) \langle x_t, \theta^{(0)} \rangle \cdot \theta^* + (\sigma(\langle x_t, \theta^{(0)} \rangle) - \sigma(\langle x_t, \theta^* \rangle)) \cdot \theta^{(0)} \right].$$

Define  $\eta := \gamma \|\theta^{(0)}\|^2 > 0$  and  $r := \frac{\|\theta^* - \theta^{(0)}\|^2}{\|\theta^{(0)}\|^2}$ . We focus on two summary statistics to keep track of the directional convergence. For all  $t \geq 0$ , define a sequence of real values  $a_t, b_t \in \mathbb{R}$

$$a_t := \langle x_t, \theta^{(0)} \rangle, \quad b_t := \langle x_t, \theta^* - \theta^{(0)} \rangle.$$

The non-linear evolution of the summary statistics is thus defined,

$$\begin{aligned} a_{t+1} &= a_t - \eta \cdot \sigma'(a_t + b_t) a_t + \eta \cdot (\sigma(a_t) - \sigma(a_t + b_t)), \\ b_{t+1} &= b_t - \eta \cdot r \sigma'(a_t + b_t) a_t. \end{aligned}$$

Recall that  $b_0 = \langle x_0, \theta^* - \theta^{(0)} \rangle = \langle x_0, \theta^* - \Pi_{\text{supp}(\mu^{(0)})} \theta^* \rangle = 0$  since  $\theta^{(0)} \in \text{BR}(\mu^{(0)})$  and  $x_0 \in \text{supp}(\mu^{(0)})$ . Without loss of generality, we consider  $a_0 = \langle x_0, \theta^{(0)} \rangle > 0$ . The directional convergence now hinges on studying  $\{a_t, b_t\}_{t \geq 0}$ , done in Lemma 11.

The proof builds upon the following “rough” observation of  $\{a_t, b_t\}_{t \geq 0}$ : after a finite time  $t_0$ , a key quantity ( $L$  for Lyapunov)

$$L_t := \frac{\sigma'(a_t + b_t)a_t}{\sigma(a_t) - \sigma(a_t + b_t)} < 1 \quad (21)$$

will cross below threshold 1 and deviate away from the threshold 1 for  $t \geq t_0$ . However, perhaps surprisingly, one can show even when  $t \rightarrow \infty$ , the quantity never cross below a threshold

$$L_t \geq \frac{1}{1+r}, \quad \forall t \geq t_0.$$

Namely, the threshold  $\frac{1}{1+r}$  is a stable fixed point for the quantity  $L_t$ . The quantity  $L_t$  regulates the monotonicity of the updates  $a_t, b_t, a_t + b_t$  and determines the order of magnitude for each term, see Lemma 9.

The above intuition is educative but hard to directly operate upon over iterations, due to the non-linear form of (21) and the nonlinear recursions of  $\{a_t, b_t\}$ . Instead of directly working with  $L_t$ , we build two envelopes inspired by  $L_t$  that are easier to control during recursions, done in Lemma 12. We define

$$L_t^{\text{env-U}} := \frac{e^{a_t+b_t}a_t}{1 - e^{2(a_t+b_t)}}, \quad \text{and} \quad L_t^{\text{env-L}} := \frac{e^{a_t+b_t}a_t}{1 + e^{a_t+b_t}}.$$

We show that these two envelopes are related to  $L_t$  in the following sense (Lemma 10), but not sandwiching it (we only have  $L_t^{\text{env-L}} \leq \min\{L_t, L_t^{\text{env-U}}\}$ )

$$L_t^{\text{env-U}} < 1 \implies L_t < 1, \quad \text{and} \quad L_t^{\text{env-L}} > \frac{1}{1+r} \implies L_t > \frac{1}{1+r}.$$

It turns out that the envelopes intervene cleanly with the non-linear recursions for  $t \rightarrow t+1$ . The crux of the argument lies in a strengthened version of the “rough observation” outlined in the previous paragraph, specifically done in Lemma 12. On the one hand, if the lower envelope function  $L_t^{\text{env-L}} > \frac{1+a_t^{-1}}{1+r+a_t^{-1}} \in [\frac{1}{1+r}, 1]$ , then the upper envelope function decreases in the recursion,  $L_{t+1}^{\text{env-U}} < L_t^{\text{env-U}} < 1$ ; On the other hand, the lower envelope function cannot decrease too much, in the following sense

$$L_t^{\text{env-L}} > \frac{1+a_t^{-1}}{1+r+a_t^{-1}} \implies L_{t+1}^{\text{env-L}} > \frac{1+a_{t+1}^{-1}}{1+r+a_{t+1}^{-1}} > \frac{1}{1+r}.$$

The two envelope functions also ensure the monotonicity of  $a_t + b_t \downarrow -\infty$  and  $a_t \uparrow +\infty$ . To sum up, we can show that the  $L_t$  has  $\frac{1}{1+r}$  as the stable fixed point. The analytical characterization of  $L_t$  ensures an explicit rate on the directional convergence. Finally, the Assumption 1 is a mild condition requiring the dynamic of  $L_t$  to reach below 1. All the detailed proof can be found in Appendix A.

### 3 Discussion and Future Work

This paper studied covariate shifts from game-theoretic and dynamic viewpoints, with the underlying Bayes optimal model being invariant. In particular, we show that under the Wasserstein gradient flow, the distribution of covariates will converge in directions in both regression and classification. However, the result presents as a dichotomy: a blessing in regression, the adversarial covariate shifts in an exponential rate to an optimal experimental design for rapid subsequent learning; a curse in classification, the adversarial covariate shifts in a subquadratic rate fast to the hardest experimental design trapping subsequent learning. We view the work as a starting point for unveiling new insights for adversarial learning and covariate shift: it cautions the researchers to be aware that robust optimization is brittle depending on the learning losses, model complexity, and early stopping. In particular, following potential directions are left as future work.

**Discriminative vs. Generative** This paper considers discriminative models for the joint distribution, where the conditional distribution  $P(Y|X)$  stays invariant, yet the covariate distribution  $P(X)$  can shift. The main reason is to define the Bayes optimal model as the equilibrium, invariant regardless of the covariate distribution  $P(X)$ . We study at the population level to simplify the main results and analysis. Another line of literature on adversarial examples, in the classification setting only, considers a generative model, where  $P(Y)$  stays untouched, yet  $P(X|Y)$  are allowed to shift. By simple Bayes rule, the Bayes optimal model  $P(Y|X)$  will consequently vary, making the concept a moving target. It is still to be determined if the notion of equilibrium or invariance exists in the generative setting. We leave it as a future direction for investigation.

**Iterative Game Updates** This paper only considers running the subsequent learning after the covariate shift reaches stationarity in direction. Generally, one may envision the game between the learner and nature by iteratively running gradient descent and ascent dynamics. The non-asymptotic analysis for iterative game updates, and the trade-offs on step sizes between the learner and the covariate shift will require future work. The analysis of iterative game updates could also benefit the understanding of generative models like generative adversarial networks.

**Complex Models** This paper shows curious insights into studying infinite-dimensional linear models with square and logistic loss. What will happen for other nonlinear models, such as neural networks? Extensions to nonlinear models require highly technical work and could lead to new insights on covariate shifts.

### Acknowledgments and Disclosure of Funding

Liang acknowledges the generous support from the NSF CAREER Grant (DMS-2042473) and the William Ladany Faculty Fellowship from the University of Chicago Booth School of Business.

## References

- L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.
- H. Bao, C. Scott, and M. Sugiyama. Calibrated surrogate losses for adversarially robust classification. In *Conference on Learning Theory*, pages 408–451. PMLR, 2020.
- P. Bartlett, S. Bubeck, and Y. Cherapanamjeri. Adversarial examples in multi-layer random relu networks. *Advances in Neural Information Processing Systems*, 34:9241–9252, 2021.
- S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. *Advances in Neural Information Processing Systems*, 19, 2006.
- S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010a.
- S. Ben-David, T. Lu, T. Luu, and D. Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 129–136. JMLR Workshop and Conference Proceedings, 2010b.
- J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman. Learning bounds for domain adaptation. *Advances in Neural Information Processing Systems*, 20, 2007.
- S. Bubeck, Y. Cherapanamjeri, G. Gidel, and R. Tachet des Combes. A single gradient step finds adversarial examples on random two-layers neural networks. *Advances in Neural Information Processing Systems*, 34:10081–10091, 2021.
- C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.
- E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.
- G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*, 2015.
- Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.
- Y. Freund and R. E. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103, 1999.
- Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(1):2096–2030, 2016.
- I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

- I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- W. Guo, Y. Hur, T. Liang, and C. Ryan. Online learning to transport via the minimal selection principle. In *Conference on Learning Theory*, volume 178 of *Proceedings of machine learning research*, pages 4085–4109. PMLR, July 2022.
- W. Hu, G. Niu, I. Sato, and M. Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037. PMLR, 2018.
- A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. Adversarial examples are not bugs, they are features. *Advances in Neural Information Processing Systems*, 32, 2019.
- A. Javanmard and M. Soltanolkotabi. Precise statistical analysis of classification accuracies for adversarial training. *The Annals of Statistics*, 50(4):2127–2156, Aug. 2022. ISSN 0090-5364, 2168-8966. doi: 10.1214/22-AOS2180.
- A. Javanmard, M. Soltanolkotabi, and H. Hassani. Precise tradeoffs in adversarial training for linear regression. In *Conference on Learning Theory*, pages 2034–2078. PMLR, 2020.
- S. Kpotufe and G. Martinet. Marginal singularity, and the benefits of labels in covariate-shift. In *Conference On Learning Theory*, pages 1882–1886. PMLR, 2018.
- T. Liang. How Well Generative Adversarial Networks Learn Distributions. *Journal of Machine Learning Research*, 22(228):1–41, 2021. ISSN 1533-7928.
- T. Liang and J. Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *International Conference on Artificial Intelligence and Statistics*, pages 907–915. PMLR, 2019.
- T. Liang and P. Sur. A precise high-dimensional asymptotic theory for boosting and minimum-l1-norm interpolated classifiers. *The Annals of Statistics*, 50(3), June 2022. ISSN 0090-5364. doi: 10.1214/22-AOS2170.
- A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
- Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.
- M. Mohri and A. Muñoz Medina. New analysis and algorithm for learning with drifting distributions. In *Algorithmic Learning Theory*, pages 124–138. Springer, 2012.
- A. Mokhtari, A. Ozdaglar, and S. Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pages 1497–1507. PMLR, 2020.

- J. Quinero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.
- A. Ross and F. Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.
- A. Sinha, H. Namkoong, R. Volpi, and J. Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.
- C. J. Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, pages 1348–1360, 1980.
- M. Sugiyama and K. Mueller. Generalization error estimation under covariate shift. In *Workshop on Information-Based Induction Sciences*, pages 21–26. Citeseer, 2005.
- M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.
- M. Telgarsky. Margins, shrinkage, and boosting. In *International Conference on Machine Learning*, pages 307–315. PMLR, 2013.
- V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.
- C. Villani. *Topics in optimal transportation*, volume 58. American Mathematical Society, 2021.
- Y. Xing, R. Zhang, and G. Cheng. Adversarially robust estimate and risk analysis in linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 514–522. PMLR, 2021.
- H. Xu, C. Caramanis, and S. Mannor. Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, 10(7), 2009.

## Appendix A. Proofs

### A.1 Proofs in Section 2.1

**Proof** [Proof of Theorem 1] In the regression setting, the utility evaluated at delta measure  $\delta_x$  takes the form

$$\mathcal{U}(\theta, \delta_x) = \int_Y \ell(f_\theta(x), y) d\pi_x^*(y) = \langle x, \theta^* - \theta \rangle^2 + 1 .$$

For each particle  $x_0 \sim \mu^{(0)}$ , the adversarial distribution shift updates following the iteration

$$\begin{aligned} x_{t+1} &= x_t + \gamma \cdot \frac{\partial}{\partial x} \mathcal{U}(\theta^{(0)}, \delta_{x_t}) , \\ &= [I + 2\gamma(\theta^* - \theta^{(0)})(\theta^* - \theta^{(0)})^\top] x_t , \\ &= (I + \tilde{\gamma} \Delta_b \Delta_b^\top) x_t , \end{aligned}$$

where  $\tilde{\gamma} := 2\gamma \|\theta^* - \theta^{(0)}\|^2$ . Therefore

$$x_T = (1 + \tilde{\gamma})^T \langle x_0, \Delta_b \rangle \Delta_b + (I - \Delta_b \Delta_b^\top) x_0 .$$

It is then clear to see that

$$\begin{aligned} \langle x_T, \Delta_b \rangle &= (1 + \tilde{\gamma})^T \langle x_0, \Delta_b \rangle , \\ \|x_T\| &= \left[ (1 + \tilde{\gamma})^{2T} \langle x_0, \Delta_b \rangle^2 + \langle x_0, (I - \Delta_b \Delta_b^\top) x_0 \rangle \right]^{1/2} , \end{aligned}$$

and thus, we can conclude the proof by noting

$$\left| \frac{\langle x_T, \Delta_b \rangle}{\|x_T\|} \right| = \frac{1}{\left[ 1 + \frac{\langle x_0, (I - \Delta_b \Delta_b^\top) x_0 \rangle}{(1 + \tilde{\gamma})^{2T} \langle x_0, \Delta_b \rangle^2} \right]^{1/2}} .$$

■

**Proof** [Proof of Theorem 4] Let  $\sigma(z) = 1/(1 + e^{-z})$  be the sigmoid function and  $\sigma'(z) = \sigma(z)(1 - \sigma(z))$ . In the classification setting,

$$\mathcal{U}(\theta, \delta_x) = \int_Y \ell(f_\theta(x), y) d\pi_x^*(y) = \frac{1}{1 + e^{-\langle x, \theta^* \rangle}} \log \left( 1 + e^{-\langle x, \theta \rangle} \right) + \frac{1}{1 + e^{\langle x, \theta^* \rangle}} \log \left( 1 + e^{\langle x, \theta \rangle} \right) .$$

For each particle  $x_0 \sim \mu^{(0)}$ , the adversarial distribution shift reads

$$\begin{aligned} x_{t+1} &= x_t + \gamma \cdot \frac{\partial}{\partial x} \mathcal{U}(\theta^{(0)}, \delta_{x_t}) , \\ &= x_t + \gamma \cdot \left[ -\sigma(\langle x_t, \theta^* \rangle) (1 - \sigma(\langle x_t, \theta^* \rangle)) \langle x_t, \theta^{(0)} \rangle \cdot \theta^* + (\sigma(\langle x_t, \theta^{(0)} \rangle) - \sigma(\langle x_t, \theta^* \rangle)) \cdot \theta^{(0)} \right] . \end{aligned}$$

For all  $t \geq 0$ , define a sequence of real values  $(a_t, b_t) \in \mathbb{R}^2$

$$a_t := \langle x_t, \theta^{(0)} \rangle , \quad b_t := \langle x_t, \theta^* - \theta^{(0)} \rangle .$$

Let  $\eta := \gamma \|\theta^{(0)}\|^2 > 0$ . Then the recursive relationship on  $\{a_t, b_t\}$  induced by the dynamics reads

$$\begin{aligned} a_{t+1} &= a_t - \eta \cdot (1 + \sqrt{r}q)\sigma'(a_t + b_t)a_t + \eta \cdot (\sigma(a_t) - \sigma(a_t + b_t)) , \\ b_{t+1} &= b_t - \eta \cdot (\sqrt{r}q + r)\sigma'(a_t + b_t)a_t + \eta \cdot \sqrt{r}q(\sigma(a_t) - \sigma(a_t + b_t)) , \end{aligned}$$

where  $q := \left\langle \frac{\theta^* - \theta^{(0)}}{\|\theta^* - \theta^{(0)}\|}, \frac{\theta^{(0)}}{\|\theta^{(0)}\|} \right\rangle$  and  $r := \frac{\|\theta^* - \theta^{(0)}\|^2}{\|\theta^{(0)}\|^2}$ . To study the directional convergence, we need sharp characterizations of the sequence  $\{a_t, b_t\}$  following the above non-linear, non-convex updates. The main technical hurdle is to derive a precise estimate on the following quantity, done in Lemma 9,

$$\left| \frac{a_T + b_T}{a_T} \right| = O\left(\frac{\log(T)}{T}\right) .$$

To apply Lemma 9, we note the assumption  $\theta^{(0)} \perp \theta^* - \theta^{(0)}$ , which in turn simplifies the dynamics with  $q = 0$ .

Recall that the direction  $\Delta_c$  can be written as

$$\Delta_c = -\frac{1}{\sqrt{1+r}} \cdot \frac{\theta^* - \theta^{(0)}}{\|\theta^* - \theta^{(0)}\|} + \frac{\sqrt{r}}{\sqrt{1+r}} \cdot \frac{\theta^{(0)}}{\|\theta^{(0)}\|} .$$

Therefore

$$\begin{aligned} \langle x_T, \Delta_c \rangle &= -\frac{1}{\sqrt{1+r}} \cdot \frac{b_T}{\|\theta^* - \theta^{(0)}\|} + \frac{\sqrt{r}}{\sqrt{1+r}} \cdot \frac{a_T}{\|\theta^{(0)}\|} \\ \|x_T\| &= \left[ \frac{b_T^2}{\|\theta^* - \theta^{(0)}\|^2} + \frac{a_T^2}{\|\theta^{(0)}\|^2} + \langle x_0, \Pi_{\{\theta^{(0)}, \theta^* - \theta^{(0)}\}}^\perp x_0 \rangle \right]^{1/2} , \end{aligned}$$

and thus we conclude the proof recalling Lemma 9

$$\begin{aligned} \frac{\langle x_T, \Delta_c \rangle}{\|x_T\|} &= \frac{(1+r)a_T - (a_T + b_T)}{\sqrt{1+r} \sqrt{b_T^2 + ra_T^2 + \|\theta^* - \theta^{(0)}\|^2 \langle x_0, \Pi_{\{\theta^{(0)}, \theta^* - \theta^{(0)}\}}^\perp x_0 \rangle}} \\ &= \frac{1 - \frac{1}{1+r} \frac{a_T + b_T}{a_T}}{\sqrt{1 - \frac{2}{1+r} \frac{a_T + b_T}{a_T} + \frac{1}{1+r} \left[ \left( \frac{a_T + b_T}{a_T} \right)^2 + \frac{\langle x_0, \Pi_{\{\theta^{(0)}, \theta^* - \theta^{(0)}\}}^\perp x_0 \rangle}{a_T^2} \right]}} . \end{aligned}$$

■

**Lemma 9 (Nonlinear recursions)** *Let  $\sigma(z) = 1/(1 + e^{-z})$  be the sigmoid function and  $\sigma'(z) = \sigma(z)(1 - \sigma(z))$ . Define the nonlinear recursion for fixed  $r, \eta > 0$ , with  $a_0 > 0$  and  $b_0 = 0$*

$$a_{t+1} = a_t - \eta \cdot \sigma'(a_t + b_t)a_t + \eta \cdot (\sigma(a_t) - \sigma(a_t + b_t)) , \quad (22)$$

$$b_{t+1} = b_t - \eta \cdot r \sigma'(a_t + b_t)a_t . \quad (23)$$

Assume there exists some  $t_0 \in \mathbb{N}$ , such that  $(a_{t_0}, b_{t_0})$  satisfy Assumption 1 with  $r > 0, c > 1$ . Then as  $T \rightarrow \infty$ ,  $a_T + b_T \rightarrow -\infty$  and  $a_T \rightarrow +\infty$  and

$$\begin{aligned} |a_T + b_T| &= O(\log(T)), \\ a_T &= \Theta(T). \end{aligned}$$

**Proof** [Proof of Lemma 9]  $a_{t_0}$  and  $b_{t_0}$  satisfy the conditions in Lemma 11, therefore we know  $a_{t+1} > a_t > 0$ , and  $a_{t+1} + b_{t+1} < a_t + b_t < 0$  for  $t \geq t_0$ . Given the monotonicity, the proof proceeds in the following steps. First, note that

$$ra_{t+1} - b_{t+1} = ra_t - b_t + \eta \cdot r(\sigma(a_t) - \sigma(a_t + b_t)), \quad (24)$$

$$a_{t+1} + b_{t+1} = (1 - \eta\sigma'(a_t + b_t))(a_t + b_t) - \eta\sigma'(a_t + b_t)(ra_t - b_t) + \eta(\sigma(a_t) - \sigma(a_t + b_t)). \quad (25)$$

Below we use the Bachmann–Landau notation: we say two sequences of positive real numbers  $x_t = O(z_t)$  iff  $\limsup_{t \rightarrow \infty} x_t/z_t < \infty$ ,  $x_t = \Omega(z_t)$  iff  $\liminf_{t \rightarrow \infty} x_t/z_t > 0$ , and  $x_t = \Theta(z_t)$  iff  $x_t = O(z_t)$  and  $x_t = \Omega(z_t)$ .

1. Observe that  $a_t > 0$ ,  $b_t \leq 0$  and  $b_{t+1} < b_t$  monotonic decreasing. The proof is straightforward from induction observing the form of (22)-(23).
2. Claim  $ra_t - b_t \rightarrow +\infty$ . Proof by contradiction. If the monotonic sequence  $ra_t - b_t$  is uniformly upper bounded by  $M$ . Then  $\sigma(a_t) - \sigma(a_t + b_t) \leq \sigma(a_t) - \sigma(a_t + b_t) = o(1)$ , which implies  $a_t \rightarrow \infty$ , and therefore contradicts  $ra_t \leq ra_t - b_t \leq M$ .
3. Claim  $b_t \rightarrow -\infty$ . Proof by contradiction. If  $b_t$  is bounded below by  $-M$ , then  $\eta \cdot r\sigma'(a_t + b_t)a_t \rightarrow 0$ . Note  $a_t \geq a_{t_0} > 0$ , then  $\sigma'(a_t + b_t) \rightarrow 0$ . Recall that  $a_t + b_t \leq a_{t_0} + b_{t_0} < 0$  and is monotonic decreasing, then  $a_t + b_t \rightarrow -\infty$  and thus  $b_t$  must go to negative infinity. We therefore reach a contradiction.
4. Claim  $a_t \rightarrow +\infty$ . Proof by contradiction, if  $a_t$  bounded by above by  $M$ , then  $a_t + b_t \rightarrow -\infty$ , then  $a_{t+1} - a_t \geq \eta(\frac{1}{2} - \sigma(a_t + b_t)) - \eta\sigma'(a_t + b_t)M \geq \eta(\frac{1}{2} - (M+1)\sigma(a_t + b_t)) \geq \frac{1}{4}\eta$  for  $t$  large enough. We thus reach a contradiction.
5. Claim  $a_t + b_t \rightarrow -\infty$ . If  $|a_t + b_t| \leq M$  for some absolute constant  $M$ , then (25) is a contradiction as we have proved  $ra_t - b_t \rightarrow +\infty$ .
6. Claim  $\lim_{t \rightarrow \infty} \frac{ra_t - b_t}{t} = \eta r$ . We have shown that  $\lim_{t \rightarrow \infty} a_t \rightarrow \infty$  and  $\lim_{t \rightarrow \infty} a_t + b_t = -\infty$ .

$$\lim_{t \rightarrow \infty} \sigma(a_t) - \sigma(a_t + b_t) = 1.$$

Therefore by (24) we can show  $\lim_{t \rightarrow \infty} \frac{ra_t - b_t}{t} = \eta r$ .

7. Claim  $\liminf_{t \rightarrow \infty} \sigma'(a_t + b_t)(ra_t - b_t) \geq \liminf_{t \rightarrow \infty} r\sigma'(a_t + b_t)a_t \geq \frac{r}{1+r}$ . Here the last inequality uses the fact  $\frac{\sigma'(a_t + b_t)a_t}{\sigma(a_t) - \sigma(a_t + b_t)} \in \left[ \frac{1 + \frac{1}{a_t}}{(1+r) + \frac{1}{a_t}}, 1 \right)$ , established in Lemma 11.

8. Claim  $\limsup_{t \rightarrow \infty} \frac{|a_t + b_t|}{\log(t)} \leq 1$ . This fact is immediate because of  $\sigma'(a_t + b_t)(ra_t - b_t) = \Omega(1)$  and  $ra_t - b_t = \Theta(t)$ .
9. Claim  $a_t = \frac{ra_t - b_t}{r+1} + \frac{a_t + b_t}{r+1} = \Theta(t) - O(\log(t)) = \Theta(t)$ . Similarly, we can show  $-b_t = \Theta(t)$ .

■

## A.2 Technical Lemmas for Section 2.1

This section contains the technical building blocks for proofs in Section 2.1. Recall  $\sigma(z) = 1/(1 + e^{-z})$  is the sigmoid function and  $\sigma'(z) = \sigma(z)(1 - \sigma(z))$ .

**Lemma 10** Define  $G_\delta(x, z) := -(1 + \delta)\sigma'(z)x + \sigma(x) - \sigma(z)$ , for  $\delta \geq 0$ ,  $z < 0$  and  $x \geq 0$ . Then,

- Fix  $z < 0$  and  $\delta \in [0, \infty)$ . If  $x > \frac{1}{1+\delta}(e^{-z} + 1)$ , then  $G_\delta(x, z) < 0$ .
- Fix  $z < 0$  and  $\delta \in [0, 1]$ , If  $x < \frac{1}{1+\delta}(e^{-z} - e^z)$ , then  $G_\delta(x, z) > 0$ .

**Proof** [Proof of Lemma 10]  $G_\delta(x, z)$  is a concave function in  $x$  for  $x \geq 0$  that only crosses the  $x$ -axis once. Denote  $x_0 > 0$  to be the unique solution to  $G_\delta(x_0, z) = 0$  for a fixed  $z < 0$ , then

$$(1 + \delta)\sigma'(z)x_0 = \sigma(x_0) - \sigma(z) \leq 1 - \sigma(z).$$

We know  $x_0 \leq \frac{1}{1+\delta} \frac{1 - \sigma(z)}{\sigma'(z)}$  and first claim is thus established.

The second claim follows as

$$G_\delta(-z, z) \geq 0, \quad \forall 0 \leq \delta \leq 1.$$

This is because  $\inf_{z \leq 0} \frac{\sigma(-z) - \sigma(z)}{-\sigma'(z)z} \geq 2 > 1 + \delta$ . Denote  $x_0 > 0$  to be the unique solution to  $G_\delta(x_0, z) = 0$ , we just proved  $x_0 > -z$  and therefore

$$\sigma(-z) - \sigma(z) \leq \sigma(x_0) - \sigma(z) = (1 + \delta)\sigma(z)[1 - \sigma(z)]x_0.$$

Therefore  $x_0 > \frac{1}{1+\delta} \frac{\sigma(-z) - \sigma(z)}{\sigma'(z)}$ . Rearrange the terms, we have proved  $x_0 > \frac{1}{1+\delta}(e^{-z} - e^z)$ . ■

**Lemma 11** Fix  $r > 0$  and  $\eta < 1/2(2 + r)$ . Consider the recursive relationship defined in Lemma 9. Assume there exists some  $t_0 \in \mathbb{N}$ , such that  $(a_{t_0}, b_{t_0})$  satisfy Assumption 1 with  $r > 0$  and  $c > 1$ . Then for  $t \geq t_0$ , we have

- $a_{t+1} > a_t$  is monotonic increasing ;
- $a_{t+1} + b_{t+1} < a_t + b_t$  is monotonic decreasing ;

and

$$\frac{\sigma'(a_t + b_t)a_t}{\sigma(a_t) - \sigma(a_t + b_t)} \in \left[ \frac{1 + \frac{1}{a_t}}{(1+r) + \frac{1}{a_t}}, 1 \right), \quad \forall t \geq t_0.$$

**Proof** [Proof of Lemma 11] The proof relies on induction. By assumption at  $t = t_0$ , we know

$$\frac{e^{a_t+b_t}a_t}{1 - e^{2(a_t+b_t)}} < 1, \quad \frac{e^{a_t+b_t}a_t}{1 + e^{a_t+b_t}} \geq \frac{1 + \frac{1}{a_t}}{1 + r + \frac{1}{a_t}}.$$

Therefore there exists some fixed small  $\epsilon > 0$  that satisfies

$$\frac{e^{a_t+b_t}a_t}{1 - e^{2(a_t+b_t)}} \leq \frac{1}{1 + \epsilon}, \quad \frac{e^{a_t+b_t}a_t}{1 + e^{a_t+b_t}} \geq \frac{1 + \frac{1}{a_t}}{1 + r + \frac{1}{a_t}}. \quad (26)$$

Apply Lemma 10, we know  $G_0(a_t, a_t + b_t) > G_\epsilon(a_t, a_t + b_t) \geq 0$ , and  $G_{\frac{r}{1+1/a_t}}(a_t, a_t + b_t) < 0$ , which implies

$$\frac{\sigma'(a_t + b_t)a_t}{\sigma(a_t) - \sigma(a_t + b_t)} \in \left[ \frac{1 + \frac{1}{a_t}}{1 + r + \frac{1}{a_t}}, \frac{1}{1 + \epsilon} \right] \subset \left[ \frac{1}{1 + r}, 1 \right), \quad \text{for } t = t_0.$$

Furthermore, monotonicity can be established for  $t = t_0$ ,

$$a_{t+1} = a_t + \eta G_0(a_t, a_t + b_t) > a_t > c, \quad (27)$$

$$a_{t+1} + b_{t+1} = a_t + b_t + \eta G_{\frac{r}{1+1/a_t}}(a_t, a_t + b_t) < a_t + b_t. \quad (28)$$

Now for the induction step from  $t$  to  $t + 1$ , we apply the Lemma 12. The assumptions for Lemma 12 are verified, and therefore

$$\frac{e^{a_{t+1}+b_{t+1}}a_{t+1}}{1 - e^{2(a_{t+1}+b_{t+1})}} < \frac{e^{a_t+b_t}a_t}{1 - e^{2(a_t+b_t)}} \leq \frac{1}{1 + \epsilon}, \quad \frac{e^{a_{t+1}+b_{t+1}}a_{t+1}}{1 + e^{a_{t+1}+b_{t+1}}} > \frac{1 + \frac{1}{a_{t+1}}}{1 + r + \frac{1}{a_{t+1}}}.$$

Repeat the induction step we can complete the proof. ■

**Lemma 12** Fix  $r > 0$  and  $\eta < 1/2(2 + r)$ . Consider the one-step recursive relationship defined in Lemma 9. Assume  $(a_t, b_t)$  satisfy the Assumption 1 with  $r > 0$  and  $c > 1$ . Then we have  $a_{t+1} + b_{t+1} < a_t + b_t$  and  $a_{t+1} > a_t$ , and furthermore satisfy the following inequality

$$\frac{e^{a_{t+1}+b_{t+1}}a_{t+1}}{1 - e^{2(a_{t+1}+b_{t+1})}} < \frac{e^{a_t+b_t}a_t}{1 - e^{2(a_t+b_t)}}, \quad (29)$$

$$\frac{e^{a_{t+1}+b_{t+1}}a_{t+1}}{1 + e^{a_{t+1}+b_{t+1}}} > \frac{1 + \frac{1}{a_{t+1}}}{1 + r + \frac{1}{a_{t+1}}}. \quad (30)$$

In other words,  $(a_{t+1}, b_{t+1})$  satisfy Assumption 1 as well.

**Proof** [Proof of Lemma 12] The proof consists of two parts: a recursive estimate for the upper bound to prove (29) and then a recursive estimate for the low bound in (30). First, due to (27) and (28), we know  $a_{t+1} + b_{t+1} < a_t + b_t$  and  $a_{t+1} > a_t$ .

**Upper bound recursion** Now let's establish the recursion for the upper bound. Note

$$\begin{aligned}
 \frac{e^{a_{t+1}+b_{t+1}}a_{t+1}}{1-e^{2(a_{t+1}+b_{t+1})}} &< \frac{e^{a_{t+1}+b_{t+1}}a_{t+1}}{1-e^{2(a_t+b_t)}} && \because a_{t+1}+b_{t+1} < a_t+b_t \\
 &= \frac{e^{a_t+b_t}a_t}{1-e^{2(a_t+b_t)}} e^{\eta G_r(a_t, a_t+b_t)} \left(1 + \eta \frac{G_0(a_t, a_t+b_t)}{a_t}\right) \\
 &\leq \frac{e^{a_t+b_t}a_t}{1-e^{2(a_t+b_t)}} e^{\eta G_r(a_t, a_t+b_t) + \eta \frac{G_0(a_t, a_t+b_t)}{a_t}} && \because 1+z \leq e^z \\
 &= \frac{e^{a_t+b_t}a_t}{1-e^{2(a_t+b_t)}} e^{\eta(1+\frac{1}{a_t})G_{\frac{r}{1+1/a_t}}(a_t, a_t+b_t)} && \text{by definition of } G_\theta(x, z) \\
 &< \frac{e^{a_t+b_t}a_t}{1-e^{2(a_t+b_t)}} < 1 && \because G_{\frac{r}{1+1/a_t}}(a_t, a_t+b_t) < 0.
 \end{aligned}$$

In addition, we can prove that there exists a constant  $\epsilon > 0$  as in (26) such that  $G_\epsilon(a_t, a_t+b_t) > 0$  and

$$a_{t+1} - a_t = \eta(\sigma(a_t) - \sigma(a_t+b_t)) \left[1 - \frac{\sigma'(a_t+b_t)a_t}{\sigma(a_t) - \sigma(a_t+b_t)}\right] > \eta(\sigma(c) - \sigma(0)) \frac{\epsilon}{1+\epsilon} = \Omega(1).$$

**Lower bound recursion** Define

$$\underline{r}_t := \frac{1 + \frac{1}{a_t}}{1 + r + \frac{1}{a_t}}, \quad \forall t.$$

Define  $\tilde{r}_t := \sigma(a_t+b_t)a_t$ , we are going to establish  $\tilde{r}_{t+1} > \underline{r}_{t+1}$ .

Now let's establish the recursion for the lower bound.

$$\begin{aligned}
 \frac{e^{a_{t+1}+b_{t+1}}a_{t+1}}{1+e^{a_{t+1}+b_{t+1}}} &> \frac{e^{a_{t+1}+b_{t+1}}a_{t+1}}{1+e^{a_t+b_t}} && \because a_{t+1}+b_{t+1} < a_t+b_t \\
 &= \frac{e^{a_t+b_t}a_t}{1+e^{a_t+b_t}} e^{\eta G_r(a_t, a_t+b_t)} e^{\log\left(1 + \frac{a_{t+1}-a_t}{a_t}\right)} \\
 &> \frac{e^{a_t+b_t}a_t}{1+e^{a_t+b_t}} e^{\eta G_r(a_t, a_t+b_t)} e^{\eta \frac{G_0(a_t, a_t+b_t)}{a_{t+1}}} && \because \log(1+z) > \frac{z}{1+z} \\
 &= \sigma(a_t+b_t)a_t \cdot e^{\eta\left(-\left(1+r+\frac{1}{a_{t+1}}\right)\sigma'(a_t+b_t)a_t + \left(1+\frac{1}{a_{t+1}}\right)[\sigma(a_t) - \sigma(a_t+b_t)]\right)}.
 \end{aligned}$$

Recall the definition of  $\tilde{r}_t := \sigma(a_t+b_t)a_t$ , we continue

$$\begin{aligned}
 &= \sigma(a_t+b_t)a_t \cdot e^{\eta\left(-\left(1+r+\frac{1}{a_{t+1}}\right)\sigma'(a_t+b_t)a_t + \left(1+\frac{1}{a_{t+1}}\right)[\sigma(a_t) - \sigma(a_t+b_t)]\right)} \\
 &= \tilde{r}_t \cdot e^{\eta\left(-\left(1+r+\frac{1}{a_{t+1}}\right)\sigma'(a_t+b_t)a_t + \left(1+\frac{1}{a_{t+1}}\right)[1 - \sigma(a_t+b_t)]\right)} e^{-\eta\left(1+\frac{1}{a_{t+1}}\right)(1-\sigma(a_t))} \\
 &= \tilde{r}_t \cdot e^{\eta[1-\sigma(a_t+b_t)]\left(-\left(1+r+\frac{1}{a_{t+1}}\right)\sigma'(a_t+b_t)a_t + \left(1+\frac{1}{a_{t+1}}\right)\right)} e^{-\eta\left(1+\frac{1}{a_{t+1}}\right)(1-\sigma(a_t))} \\
 &> \tilde{r}_t \cdot \left\{1 - \eta[1 - \sigma(a_t+b_t)]\left(1+r + \frac{1}{a_{t+1}}\right)\left(\tilde{r}_t - \frac{1 + \frac{1}{a_{t+1}}}{1 + r + \frac{1}{a_{t+1}}}\right)\right\} e^{-\eta\left(1+\frac{1}{a_{t+1}}\right)(1-\sigma(a_t))}.
 \end{aligned}$$

Here the last two lines follow from the facts  $\sigma'(z) = \sigma(z)(1 - \sigma(z))$  and  $e^z \geq 1 + z$ . We first control the last term on the RHS of the above equation,

$$e^{-\eta(1 + \frac{1}{a_{t+1}})(1 - \sigma(a_t))} \geq 1 - \eta(1 + \frac{1}{a_{t+1}})(1 - \sigma(a_t)) > 1 - \eta \frac{1 + c^{-1}}{1 + e^{a_t}} > 1 - \eta(1 + c^{-1})e^{-a_t}.$$

Define  $\tilde{\eta} := \eta(1 + r + c^{-1}) > \eta[1 - \sigma(a_t + b_t)](1 + r + \frac{1}{a_{t+1}})$ , we continue with the bound

$$\tilde{r}_{t+1} > \tilde{r}_t \left[ 1 - \tilde{\eta}(\tilde{r}_t - \underline{r}_{t+1}) \right] \left[ 1 - \eta(1 + c^{-1})e^{-a_t} \right]. \quad (31)$$

For  $a_t > c$  some absolute constant, we have  $e^{-a_t} < \underline{r}_t - \underline{r}_{t+1}$ , as the RHS  $\underline{r}_t - \underline{r}_{t+1}$  is of order  $(a_{t+1} - a_t)a_t^{-2} = \Theta(a_t^{-2})$  yet the LHS is exponential in  $a_t$ . Therefore  $\eta(1 + c^{-1})e^{-a_t} < \tilde{\eta}e^{-a_t} < \tilde{\eta}(\underline{r}_t - \underline{r}_{t+1}) < \tilde{\eta}(\tilde{r}_t - \underline{r}_{t+1})$ , and thus (31) reads

$$\tilde{r}_{t+1} > \tilde{r}_t \left[ 1 - \tilde{\eta}(\tilde{r}_t - \underline{r}_{t+1}) \right]^2.$$

One can subtract  $\underline{r}_{t+1}$  on both sides, and see

$$\tilde{r}_{t+1} - \underline{r}_{t+1} > \tilde{r}_t \left[ 1 - \tilde{\eta}(\tilde{r}_t - \underline{r}_{t+1}) \right]^2 - \underline{r}_{t+1} \geq (\tilde{r}_t - \underline{r}_{t+1})[1 - 2\tilde{\eta}\tilde{r}_t + \tilde{\eta}^2\tilde{r}_t(\tilde{r}_t - \underline{r}_{t+1})].$$

The lower bound on Eqn. (30) is true as long as the RHS of the above is positive

$$(\tilde{r}_t - \underline{r}_{t+1})[1 - 2\tilde{\eta}\tilde{r}_t + \tilde{\eta}^2\tilde{r}_t(\tilde{r}_t - \underline{r}_{t+1})] > 0.$$

Note  $\tilde{\eta} = \eta(1 + r + c^{-1}) < \eta(2 + r) < 1/2$ ,  $\tilde{r}_t < 1$  and  $\tilde{r}_t > \underline{r}_t > \underline{r}_{t+1}$ , we have concluded

$$\frac{e^{a_{t+1} + b_{t+1}} a_{t+1}}{1 + e^{a_{t+1} + b_{t+1}}} = \tilde{r}_{t+1} > \underline{r}_{t+1} = \frac{1 + \frac{1}{a_{t+1}}}{1 + r + \frac{1}{a_{t+1}}}.$$

■

### A.3 Proofs in Section 2.2

**Proof** [Proof of Theorem 6] Let  $y^i = \langle x^i, \theta^* \rangle + \epsilon^i$  with  $\epsilon^i$  i.i.d. Gaussian, then

$$\begin{aligned} \theta^{(1)} &= \theta^{(0)} - \eta \cdot \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} \ell(\langle x^i, \theta \rangle, y^i) |_{\theta = \theta^{(0)}}, \\ \theta^{(1)} - \theta^* &= \left( I - 2\eta \frac{1}{n} \sum_{i=1}^n x^i \otimes x^i \right) (\theta^{(0)} - \theta^*) + 2\eta \frac{1}{n} \sum_{i=1}^n \epsilon^i x^i. \end{aligned}$$

Fix  $n$ , first let  $T \rightarrow \infty$ . Theorem 1 shows that  $x_T^i \rightarrow \Delta_b$  (we denote the dependence on  $T$  explicitly) in the following sense

$$\langle x_T^i, \Delta_b \rangle^2 \geq 1 - O(e^{-cT}).$$

Choose  $\eta = 1/2$ , we have

$$\|\theta^{(1)} - \theta^*\| \leq \underbrace{\left\| \left( I - \frac{1}{n} \sum_{i=1}^n x^i \otimes x^i \right) (\theta^{(0)} - \theta^*) \right\|}_{:=\Gamma} + \left\| \frac{1}{n} \sum_{i=1}^n \epsilon^i x^i \right\|.$$

For the first term  $\Gamma$ , note  $\Gamma = \langle \Delta_b, \Gamma \rangle \Delta_b + \Pi_{\Delta_b}^\perp \Gamma$ . Along the direction  $\Delta_b$ ,

$$\langle \Delta_b, \Gamma \rangle = \|\theta^{(0)} - \theta^*\| \cdot \left( 1 - \frac{1}{n} \sum_{i=1}^n \langle x^i, \Delta_b \rangle^2 \right).$$

Projecting to the orthogonal complement to  $\Delta_b$ , we know

$$\|\Pi_{\Delta_b}^\perp \Gamma\| \leq \frac{1}{n} \sum_{i=1}^n \langle x^i, \Delta_b \rangle \|\Pi_{\Delta_b}^\perp x_i\| \leq \frac{1}{n} \sum_{i=1}^n \langle x^i, \Delta_b \rangle \sqrt{1 - \langle x^i, \Delta_b \rangle^2}.$$

Therefore

$$\lim_{T \rightarrow \infty} \|\Gamma\| \leq \lim_{T \rightarrow \infty} \left\{ \|\theta^{(0)} - \theta^*\| \cdot \left( 1 - \frac{1}{n} \sum_{i=1}^n \langle x_T^i, \Delta_b \rangle^2 \right) + \frac{1}{n} \sum_{i=1}^n \langle x_T^i, \Delta_b \rangle \sqrt{1 - \langle x_T^i, \Delta_b \rangle^2} \right\} = 0.$$

For the second term, we apply the Hanson–Wright inequality and recall  $\|x^i\| = 1$  to reach

$$\left\| \frac{1}{n} \sum_{i=1}^n \epsilon^i x^i \right\| \leq \sqrt{C_1 \cdot \frac{\text{Tr}(\frac{\sum_{i=1}^n x^i \otimes x^i}{n})(1 + \log^{0.5}(1/\delta)) + C_2 \|\frac{\sum_{i=1}^n x^i \otimes x^i}{n}\|_{\text{op}} \log(1/\delta)}{n}} = O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right)$$

with probability at least  $1 - \delta$ .  $\blacksquare$

**Proof** [Proof of Theorem 7] In what follows, we consider fixed  $n$ , and let  $T \rightarrow \infty$ .

$$\theta^{(1)} = \theta^{(0)} - \eta \cdot \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} \ell(\langle x^i, \theta \rangle, y^i) |_{\theta=\theta^{(0)}},$$

$$\theta^{(1)} - \theta^* = (\theta^{(0)} - \theta^*) - \eta \frac{1}{n} \sum_{i=1}^n (\sigma(\langle x^i, \theta^{(0)} \rangle) - y^i) x^i.$$

Theorem 1 shows that  $x_T^i \rightarrow \Delta_c$ , where the convergence means directional convergence, and we denote the dependence on  $T$  explicitly. Then as  $\lim_{T \rightarrow \infty} \langle \Delta_c, x_T^i \rangle = 1$  and  $\|x_T^i\| = 1$ , we know,

$$\lim_{T \rightarrow \infty} \langle \frac{\theta^*}{\|\theta^*\|}, x_T^i \rangle^2 = 1 - \lim_{T \rightarrow \infty} \langle \Delta_c, x_T^i \rangle^2 = 0,$$

and therefore

$$\lim_{T \rightarrow \infty} \left| \langle \theta^* - \theta_{n,T,\eta}^{(1)}, \frac{\theta^*}{\|\theta^*\|} \rangle - \langle \theta^* - \theta^{(0)}, \frac{\theta^*}{\|\theta^*\|} \rangle \right| \leq \eta \frac{1}{n} \sum_{i=1}^n \lim_{T \rightarrow \infty} |\sigma(\langle x_T^i, \theta^{(0)} \rangle) - y_T^i| \cdot \left| \langle \frac{\theta^*}{\|\theta^*\|}, x_T^i \rangle \right| \rightarrow 0.$$

The final claim is a direct fact from

$$\lim_{T \rightarrow \infty} \|\theta^* - \theta_{n,T,\eta}^{(1)}\| \geq \left| \langle \theta^* - \theta_{n,T,\eta}^{(1)}, \frac{\theta^*}{\|\theta^*\|} \rangle \right| = \left| \langle \theta^* - \theta^{(0)}, \frac{\theta^*}{\|\theta^*\|} \rangle \right|.$$

We conclude the proof by taking  $\liminf$  w.r.t. to  $n$ .  $\blacksquare$