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

Tengyuan Liang*[1]

[1]University of Chicago

February 14, 2023

Abstract

Covariate distribution shifts and adversarial perturbations present robustness challenges to the conventional statistical learning framework: mild inconceivable 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 fast 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 that the training data set is drawn from, for strong guarantees of learnability and generalization [1]. The reason is grounded since, if not, either (1) concept shift, namely the conditional distribution of $Y|X$ changes hence the underlying Bayes optimal prediction model $f^\star : X \to Y$ could shift,

*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. The author wishes to thank Alex Belloni, Denis Chetverikov, Max Farrell, Chris Hansen, Sendhil Mullainathan, and Ben Recht for the valuable discussions.

or (2) covariate shift, namely the covariate distribution $\mu \in \mathcal{P}(X)$ shifts so that the underlying evaluation metric[1] for the learned model f changes. Nevertheless, in practice, supervised learning is often deployed "in the wild," meaning the test data distribution typically extrapolates the training data distribution.

Concept shift is inherently a difficult problem as if shooting for a moving target. However, covariate shift may be less severe a problem as long as the concept does not vary. Specific statistical methods allow for mild extrapolation[2], say the (fixed-design) linear regression and the (local) nonparametric regression [2]. Recently, another notable line of work has arisen [3, 4], where covariate shift adaption is studied assuming knowledge of the density ratio of the covariate shift.

Akin to covariate shift, adversarial perturbations have recently revived interest in machine learning and robust optimization communities [5, 6]. Adversarial perturbation is motivated by the following observation: small local perturbations to the covariate distribution can significantly compromise the supervised learning performance [5, 7, 8]. 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 μ: the model f minimizes the risk from a model class, yet the covariate distribution μ maximizes the risk from a probability distribution class. Such a game perspective between the learning model f and data distribution μ has been influential since the seminal work of boosting [9, 10].

This paper studies a particular form of covariate shifts following adversarial perturbations, with the underlying concept (i.e., Bayes optimal model) held fixed. 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 learning and experimental design. Before diving into our problem setup, we first elaborate on the background and literature.

1.1 Background and Literature Review

To make concrete the discussions on covariate shift and adversarial perturbation, we first fix some notations. Let X be space for the covariates, and $Y \subset \mathbb{R}$ be the space for a real-valued response variable. When 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 $f : X \to Y$ be a statistical model and $\ell(\cdot,\cdot) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ 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 π, one can define the utility function that accesses the risk of model f on data π

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

Models, covariate distributions, and Bayes optimality Following the literature [3, 11], we consider the covariate shift but not the concept shift. For a valid marginal probability measure for the covariate $\mu \in \mathcal{P}(X)$, we define the induced joint measure for the covariate and response pair

$$\pi_\mu := \int \delta_x \otimes \pi_x^\star \, \mathrm{d}\mu(x) \in \mathcal{P}(X \times Y), \tag{1.1}$$

where $\pi_x^\star \in \mathcal{P}(Y)$ denotes a fixed conditional data generating process for $\mathbf{y}|\mathbf{x} = x$ that does not vary with μ. (1.1) should be read as disintegration of measure [12], meaning for all bounded continuous function $h \in C_b(X \times Y)$

$$\int_{X\times Y} h(x, y)\, \mathrm{d}\pi_\mu(x, y) = \int_X \Big[\int_Y u(x, y)\, \mathrm{d}\pi_x^\star(y) \Big] \mathrm{d}\mu(x)\ .$$

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

[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.

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

$$f^\star_{\text{Bayes}} : x \mapsto \arg\min_{y' \in Y} \int \ell(y', y) \mathrm{d}\pi_x^\star(y) \ . \tag{1.2}$$

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

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

$$\mathcal{U}(f, \mu) := \mathcal{R}(f, \pi_\mu) = \int_X \Big[\int_Y \ell(f(x), y) \mathrm{d}\pi_x^\star(y) \Big] \mathrm{d}\mu(x) \ . \tag{1.3}$$

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

Classical statistical learning theory studies $\mathcal{U}(\widehat{f_\mu}, \mu)$ where $\widehat{f_\mu}$ is learned based on empirical data set drawn from the same distribution π_μ. However, when the covariate distribution shift to another measure ν that is different from the training data distribution μ, the performance $\mathcal{U}(\widehat{f_\mu}, \nu)$ deteriorates, see [3, 11] for a review on covariate shifts. Assuming the knowledge of the density ratio $\mathrm{d}\nu/\mathrm{d}\mu$, 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 [6, 13, 7]. 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,

$$\mathcal{U}_\gamma(f, \mu) := \max_{\nu \,:\, d^2(\nu, \mu) \le \gamma} \mathcal{U}(f, \nu) \ , \tag{1.4}$$

where $\mathcal{U}(\cdot,\cdot)$ is defined in (1.3). When the metric is the Wasserstein metric W_2, the Lagrangian of (1.4) becomes

$$\max_{\lambda \ge 0} \min_{\nu \in \mathcal{P}(X)} -\mathcal{U}(f, \nu) + \lambda \Big[W_2^2(\nu, \mu) - \gamma \Big] = \max_{\lambda \ge 0} \mathop{\mathbb{E}}_{\mathbf{x} \sim \mu} \underbrace{\Big[\min_{x' \in X} \Big(-\mathcal{U}(f, \delta_{x'}) + \frac{\lambda}{2} \|x' - \mathbf{x}\|^2 \Big) \Big]}_{\text{Moreau-Yosida regularization}} - \gamma\lambda \ , \tag{1.5}$$

where δ_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 $-\mathcal{U}(f, \delta_x) : X \to \mathbb{R}$ with parameter λ^{-1}, thus serving as a smoothed regularization to the original loss.

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 ν differs from μ. It is worth noting that under a host of Gaussian covariate models, [14, 8] provided sharp high-dimensional asymptotic characterization of the adversarial risk $\mathcal{U}_\gamma(\widehat{f}, \mu)$ and the standard risk $\mathcal{U}(\widehat{f}, \mu)$ and thus shed light on the tradeoffs, for a range of models $\widehat{f}$ interpolating between empirical risk minimization and adversarial risk minimization.

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 $\gamma \in \mathbb{R}_+$,

$$\nu := \arg\min_{\nu \in \mathcal{P}(X)} -\mathcal{U}(f, \nu) + \frac{1}{\gamma} W_2^2(\nu, \mu) \ , \tag{1.6}$$

corresponding to (1.5). Denote the distribution shift map $\mathrm{Ds}_\gamma : x \mapsto \arg\min_{x' \in X} \left(-\mathcal{U}(f, \delta_{x'}) + \frac{1}{2\gamma}\|x' - x\|^2 \right)$ defined by the Moreau-Yosida envelope. Informally, such a map defines the worst-case covariate shift for the model f evaluated at measure μ, as the maximizer of the adversarial perturbation is attained at

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

where $\lambda > 0$ is the solution to the dual. In the infinitesimal limit $\gamma \asymp \lambda^{-1} \to 0$, one can show that [15, 16]

$$\frac{\left(\mathrm{Ds}_\gamma - \mathrm{Id}\right)[x]}{\gamma} \to \frac{\partial}{\partial x}\mathcal{U}(f, \delta_x) \,. \tag{1.7}$$

The distribution shift map Ds_γ presents a way of constructing adversarial examples [5, 17, 18], 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 (1.6) is called the Wasserstein gradient flow as $\gamma \to 0$, where the density ρ_t (associated with ν_t), w.r.t. the Lebesgue measure, evolves according to the following PDE [15]

$$\partial_t \rho_t + \nabla \cdot (\rho_t \mathrm{V}) = 0 \,, \text{ where } \mathrm{V} : x \mapsto \frac{\partial}{\partial x}\mathcal{U}(f, \delta_x) \,. \tag{1.8}$$

1.2 Problem Setup

This paper considers regression and classification problems in a possibly 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 a infinite-dimensional linear model class $\mathcal{F} := \{f_\theta \mid f_\theta(x) := \langle x, \theta \rangle, \theta \in \ell^2_{\mathbb{N}}\}$ where the inner-product corresponds to the Hilbert space $\ell^2_{\mathbb{N}}$. Slightly abusing the notation, we write for convenience the utility function

$$\mathcal{U}(\theta, \mu) = \mathop{\mathbb{E}}_{(x,y) \sim \pi_\mu}\left[\ell(f_\theta(x), y)\right] = \int_X \Big[\int_Y \ell(f_\theta(x), y)\, \mathrm{d}\pi^\star_x(y) \Big] \mathrm{d}\mu(x) \,. \tag{1.9}$$

We investigate two types of conditional relationships for $\pi^\star_x$ in (1.1), namely for some $\theta^\star \in \ell^2_{\mathbb{N}}$:

$$\begin{aligned}
&\text{Regression: } \mathbf{y}|\mathbf{x} = x \sim \text{Gaussian}\Big(\langle x, \theta^\star \rangle, 1\Big),\ \ell(f, y) = (f - y)^2 \,; \\
&\text{Classification: } \mathbf{y}|\mathbf{x} = x \sim \text{Bernoulli}\Big(\sigma(\langle x, \theta^\star \rangle)\Big),\ \ell(f, y) = -fy + \log\Big(1 + e^f\Big) \,.
\end{aligned}$$

Here $\sigma(z) = 1/(1 + e^{-z})$ is the sigmoid function. In both the regression and classification settings we study, the Bayes optimal model (1.2) is uniquely defined

$$f^\star_{\text{Bayes}}(x) = \langle x, \theta^\star \rangle \,.$$

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

$$\begin{aligned}
\min_\theta \max_\mu \ \mathcal{U}(\theta, \mu) \geq \max_\mu \min_\theta \ \mathcal{U}(\theta, \mu) &\geq \max_\mu \ \int_X \Big[\min_{\theta \in \ell^2_{\mathbb{N}}} \int_Y \ell(f_\theta(x), y)\, \mathrm{d}\pi^\star_x(y) \Big] \mathrm{d}\mu(x) \\
&= \max_\mu \ \mathcal{U}(\theta^\star, \mu) \geq \min_\theta \max_\mu \ \mathcal{U}(\theta, \mu) \,.
\end{aligned} \tag{1.10}$$

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

The game perspective is not new. The celebrated boosting literature [9, 10, 19, 20] 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 μ is perturbed as in (1.6) under the Kullback-Leibler metric, resulting in an exponentiated gradient step for covariate shift. The game perspective is also instrumental to the generative modeling and adversarial learning literature [21, 22, 23, 24, 25, 26], 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 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 $\mu^{(0)}$

$$\theta^{(0)} \in \text{BR}(\mu^{(0)}) := \operatorname*{arg\,min}_{\theta \in \ell_{\mathbb{N}}^2} \mathcal{U}(\theta, \mu^{(0)}) .$$

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

$$\theta^{(0)} \in \text{BR}(\mu^{(0)}) = \left\{ \Pi_{\text{supp}(\mu^{(0)})} \theta^\star + \Pi^\perp_{\text{supp}(\mu^{(0)})} \xi \mid \forall \xi \in \ell^2_{\mathbb{N}} \right\} ,$$

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 $\Pi_{\text{supp}(\mu^{(0)})}\theta^\star$, while on the orthogonal domain $\Pi^\perp_{\text{supp}(\mu^{(0)})}$ no information is learned.

The minimum Hilbert space norm solution in the over-identified set $\text{BR}(\mu^{(0)})$ is $\Pi_{\text{supp}(\mu^{(0)})}\theta^\star$. Clearly, $\theta^{(0)} = \Pi_{\text{supp}(\mu^{(0)})}\theta^\star$ is inconsistent with the Bayes optimal model $\theta^\star$ [3]. It is, therefore, natural to consider, for typical adversarial distribution shifts $\mu^{(0)} \to \mu^{(1)}$, 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 $\theta^\star$. 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 $\nu_0 := \mu^{(0)}$,

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

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

$$\partial_t \rho_t + \nabla \cdot (\rho_t \mathrm{V}) = 0 , \text{ where } \mathrm{V} : x \mapsto \frac{\partial}{\partial x} \mathcal{U}(\theta^{(0)}, \delta_x) . \tag{1.11}$$

Conceptually, the adversarial distribution shift is a gradient ascent flow on the Wasserstein space $(\mathcal{P}(X), W_2)$ defined in (1.11) with effective time γT. In this paper, we study a discretization of (1.11) with stepsize γ and iterations T, as follows

$$x_{t+1} = x_t + \gamma \cdot \frac{\partial}{\partial x} \mathcal{U}(\theta^{(0)}, \delta_x)|_{x = x_t} , \text{ for } t = 0, 1, \ldots, T , \text{ where } x_0 \sim \mu^{(0)} . \tag{1.12}$$

Now we state the main goal of this paper:

Adversarial covariate shifts move the current covariate domain to an extrapolation region. We provide a precise characterization of 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.

2 Main Results

2.1 Adversarial Covariate Shifts: Blessings and Curses

Let $\theta^{(0)} \in \ell^2_{\mathbb{N}}$ be the current learning model and $\theta^\star - \theta^{(0)}$ be the remaining signal to be identified. Define two unit-norm directions: the blessing direction $\Delta_{\mathrm{b}} \in \mathbb{R}^{\mathbb{N}}$ and the curse direction $\Delta_{\mathrm{c}} \in \mathbb{R}^{\mathbb{N}}$

$$\Delta_{\mathrm{b}} := \frac{\theta^\star - \theta^{(0)}}{\|\theta^\star - \theta^{(0)}\|} \in \ell^2_{\mathbb{N}}(1)\,, \tag{2.1}$$

$$\Delta_{\mathrm{c}} := -\frac{\|\theta^{(0)}\|}{\|\theta^\star\|} \cdot \frac{\theta^\star - \theta^{(0)}}{\|\theta^\star - \theta^{(0)}\|} + \frac{\|\theta^\star - \theta^{(0)}\|}{\|\theta^\star\|} \cdot \frac{\theta^{(0)}}{\|\theta^{(0)}\|} \in \ell^2_{\mathbb{N}}(1)\,. \tag{2.2}$$

The name blessing comes from the fact that $\Delta_{\mathrm{b}} /\!/ \theta^\star - \theta^{(0)}$, 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 $\Delta_{\mathrm{b}} \perp \theta^\star$, that is, the direction is perpendicular to the signal direction. Curiously, we will show that the adversarial learning dynamic collapses to a probability measure along the blessing direction Δ_{b} in the regression problem; in sharp contrast, the probability measure induced by the adversarial learning dynamic converges to the curse direction Δ_{c} in the classification problem. The formal directional convergence results are stated in the following Theorems 1 and 2. For the flow of the exposition, the primary assumption and intuition of the proof are deferred to Section 2.4. The detailed proof of all theorems can be found in Appendix A.

We first state the result for the infinite-dimensional regression problem. All the relevant notations in Theorems 1 and 2 were introduced in Section 1.2.

Theorem 1 (Regression: directional convergence). *Consider the regression setting where $\ell(y', y) = (y' - y)^2$ and $\mathbf{y}|\mathbf{x} = x \sim \mathrm{Gaussian}\big(\langle x, \theta^\star\rangle, 1\big)$. Let $x_0 \in \mathrm{supp}(\mu^{(0)})$ that satisfies $\langle x_0, \theta^\star - \theta^{(0)}\rangle \neq 0$. Then the induced adversarial distribution shift dynamic (1.12) satisfies*

$$\lim_{T\to\infty} \left|\left\langle \tfrac{x_T}{\|x_T\|}, \Delta_{\mathrm{b}} \right\rangle\right| = 1\,,\ \textit{where } \Delta_{\mathrm{b}} /\!/ \theta^\star - \theta^{(0)} \textit{ is defined in } (2.1)\,. \tag{2.3}$$

Moreover, the directional convergence is exponential in T,

$$\left|\left\langle \tfrac{x_T}{\|x_T\|}, \Delta_{\mathrm{b}} \right\rangle\right| \in \left[1 - O\!\left(\tfrac{1}{e^{cT}}\right),\ 1\right]. \tag{2.4}$$

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

Remark. This theorem concerns the case when the current model $\theta^{(0)}$ is imperfect—namely $\|\theta^\star - \theta^{(0)}\| \neq 0$—the only case when distribution shifts that vary $\mu \in \mathcal{P}(X)$ could impact learning the conditional relationship $\pi^\star_x$ and hence identifying the equilibrium $f^\star_{\mathrm{Bayes}}$ defined in (1.2). The current model could be imperfect due to (1) $\mathrm{supp}(\mu^{(0)}) \subsetneq X$ the covariate does not span the full infinite-dimensional space, and/or (2) 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 $\mu^{(0)} \to \mu^{(1)}$ 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 Δ_{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 $\theta^\star$. The impact of the distribution shifts on the next stage learner, in this sequential game perspective, is formally stated in Theorem 3. 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.

Theorem 2 (Classification: directional convergence)**.** *Consider the classification setting where* $\ell(y', y) = -y'y + \log\left(1 + e^{y'}\right)$ *and* $\mathbf{y}|\mathbf{x} = x \sim \mathrm{Bernoulli}\left(\sigma(\langle x, \theta^\star\rangle)\right)$*. Assume* $\theta^{(0)} \perp \theta^\star - \theta^{(0)}$*. Let* $x_0 \in \mathrm{supp}(\mu^{(0)})$ *and assume there exists a* $t_0 \in \mathbb{N}$ *such that* $(a_0, b_0) := \left(\langle x_{t_0}, \theta^{(0)}\rangle, \langle x_{t_0}, \theta^\star - \theta^{(0)}\rangle\right)$ *satisfying Assumption 1. Then the induced adversarial distribution shift dynamic (1.12) satisfies*

$$\lim_{T\to\infty} \left|\left\langle \tfrac{x_T}{\|x_T\|}, \Delta_{\mathrm{c}}\right\rangle\right| = 1\ , \ \textit{where } \Delta_{\mathrm{c}} \perp \theta^\star \textit{ is defined in (2.2)}\,. \tag{2.5}$$

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

$$\left|\left\langle \tfrac{x_T}{\|x_T\|}, \Delta_{\mathrm{c}}\right\rangle\right| \in \left[1 - O\Big(\tfrac{\log^2(T)}{T^2}\Big)\ ,\ 1\right]\,. \tag{2.6}$$

Remark. This theorem concerns also the case when the current model $\theta^{(0)}$ is imperfect—namely $\|\theta^\star - \theta^{(0)}\| \neq 0$. In particular, this theorem studies when $\mathrm{supp}(\mu^{(0)}) \subsetneq X$. As stated before, the best response model $\theta^{(0)}$ associated with the measure $\mu^{(0)}$ takes the form $\theta^{(0)} \in \mathrm{BR}(\mu^{(0)}) = \left\{\Pi_{\mathrm{supp}(\mu^{(0)})}\theta^\star + \Pi^\perp_{\mathrm{supp}(\mu^{(0)})}\xi \mid \forall \xi \in \ell^2_{\mathbb{N}}\right\}$. The minimum-norm solution for the best response set satisfies the assumption in the theorem, namely $\theta^{(0)} \perp \theta^\star - \theta^{(0)}$. The theorem is also about directional convergence: the adversarial distribution shift dynamics $\mu^{(0)} \to \mu^{(1)}$ asymptotically align all the mass of the covariates along a one-dimensional, "curse" direction $\Delta_{\mathrm{c}} \perp \theta^\star$, 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 $\theta^\star$. The impact of the distribution shifts on the next stage learner, in this sequential game perspective, is formally stated in Theorem 4. Quantitatively, the directional convergence to Δ_{c} in the classification setting is quadratic in T, much slower than the directional convergence to Δ_{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 2, visualized in Figs 2.3 and 2.3.

Tracking down the exact behavior of the distribution shift dynamic—non-convex and non-linear—and establishing a sharp directional convergence rate are the main technical innovation and difficulty in this paper. 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. Discussions regarding the initial condition on x_0 will also be found in Section 2.4.

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 2 can translate to a direct impact on the learner's subsequent learning towards the Bayes optimal model $f^\star_{\text{Bayes}}(x) = \langle x, \theta^\star \rangle$, the equilibrium. The impact is dramatic, and either comes in as a blessing or a curse.

The sequential game between model f_θ and covariate distribution μ 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 $\widehat{\mu}^{(0)} \to \widehat{\mu}^{(1)}$ 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 (1.12) 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 (1.1), and update

$$\theta^{(1)} = \theta^{(0)} - \eta \cdot \frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial \theta} \ell(\langle x^i, \theta \rangle, y^i)|_{\theta = \theta^{(0)}} \,. \tag{2.7}$$

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

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.

Theorem 3 (Regression: blessing to the learner)**.** *Consider the same setting as in Theorem 1. For all $\theta^{(0)}$ such that $\|\theta^\star - \theta^{(0)}\| \neq 0$, the learner's one-step reaction to the distribution shift as in (2.7) with $\eta = 1/2$ satisfies*

$$\lim_{n\to\infty} \lim_{T\to\infty} \|\theta^\star - \theta^{(1)}_{n,T,\eta}\| = 0 \;\; a.s. \,. \tag{2.8}$$

Theorem 4 (Classification: curse to the learner)**.** *Consider the same setting as in Theorem 2. For all $\theta^{(0)}$ such that $\langle \theta^\star - \theta^{(0)}, \theta^\star \rangle \neq 0$, the learner's one-step reaction to the distribution shift as in (2.7) with any fixed $\eta > 0$ satisfies*

$$\lim_{n\to\infty} \lim_{T\to\infty} \frac{\langle \theta^\star - \theta^{(1)}_{n,T,\eta}, \theta^\star \rangle}{\langle \theta^\star - \theta^{(0)}, \theta^\star \rangle} = 1 \,. \tag{2.9}$$

Moreover,

$$\liminf_{n\to\infty} \lim_{T\to\infty} \|\theta^\star - \theta^{(1)}_{n,T,\eta}\| > 0 \,. \tag{2.10}$$

Remark. The result in Corollary 4 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 $\widehat{\mu}^{(1)}$ in making progress towards $\theta^\star$. Comparing Theorems 3 and 4, we see that the adversarial distribution shifts in the regression setting make the learner's one-step subsequent move optimal! (2.8) shows that one-step improvement using gradient descent dynamic as in (2.7) will reach the Bayes optimal model, also the equilibrium to the minimax game as in (1.10). On the contrary, in the classification setting, (2.10) 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 3 and 4.

Experiment setup Both simulations are based on a high-dimensional setting where $X = \mathbb{R}^{200}$ and $\mathrm{supp}(\mu^{(0)}) = \mathbb{R}^{100}$ a randomly drawn subspace from the Haar measure (a uniformly drawn subspace of dimension 100). Specify a Bayes optimal model $\theta^\star = [1, 1/2, \ldots, 1/i, \ldots, 1/200]^\top$, denoted by the the Yellow $\star$ in the figures. The best response model restricted to the subspace $\mathrm{supp}(\mu^{(0)})$, $\theta^{(0)}$ is noted by Green $\blacktriangle$, and the remaining signal $\theta^\star - \theta^{(0)}$ is denoted by Red $\times$. The probability distribution of the covariates is shown by the Blue $\bullet$. To visualize the adversarial distribution shift on a two-dimensional plane, we project all points to the two-dimensional subspace spanned by $\{\theta^\star, \theta^\star - \theta^{(0)}\}$. 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 $\{\frac{\theta^\star}{\|\theta^\star\|}, \frac{\theta^\star - \theta^{(0)}}{\|\theta^\star - \theta^{(0)}\|}\}$.

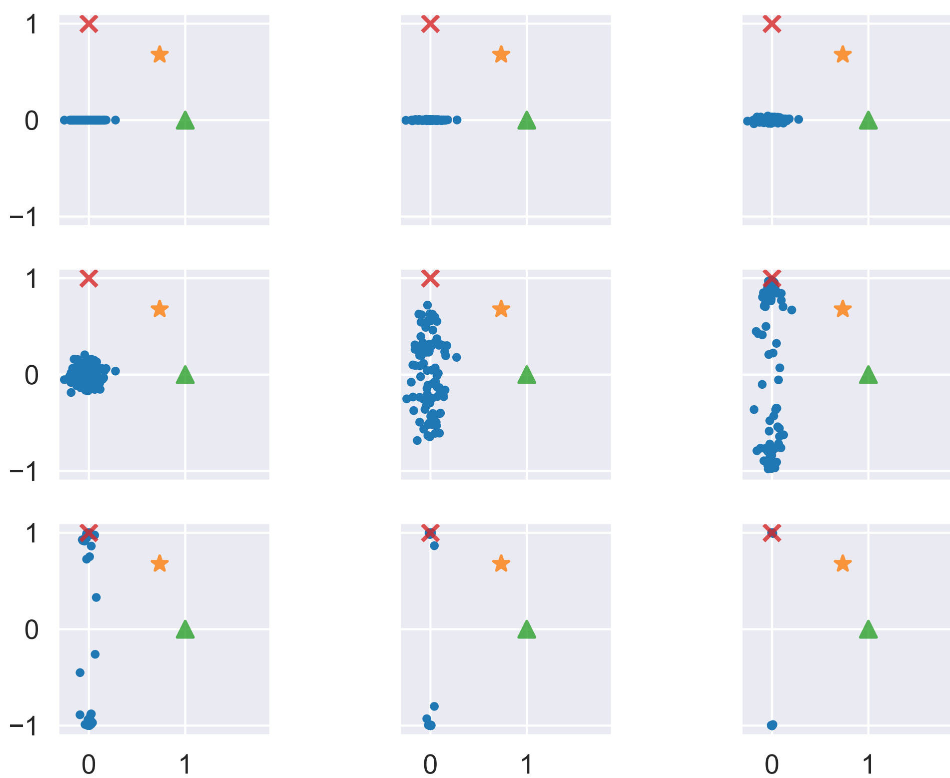

Figure 1: Regression setting, directional convergence. From left to right, top to bottom, we plot the directional information at timestamp $t = 0, 5, 10, \ldots, 40$, once every 5 iterations.

Directional convergence: regression Consider (1.9) with $\mathbf{y}|\mathbf{x} = x \sim \mathrm{Gaussian}\big(\langle x, \theta^\star \rangle, 1\big)$ and $\ell(y', y) = (y' - y)^2$. The adversarial distribution shift evolves according to (1.12). Here the blessing direction Δ_b defined in (2.1) is precisely marked by Red $\times$. As seen in Fig. 2.3, the Blue $\bullet$ data clouds collapsed to a align perfectly to Δ_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 $\mathbb{R}^{200}$.

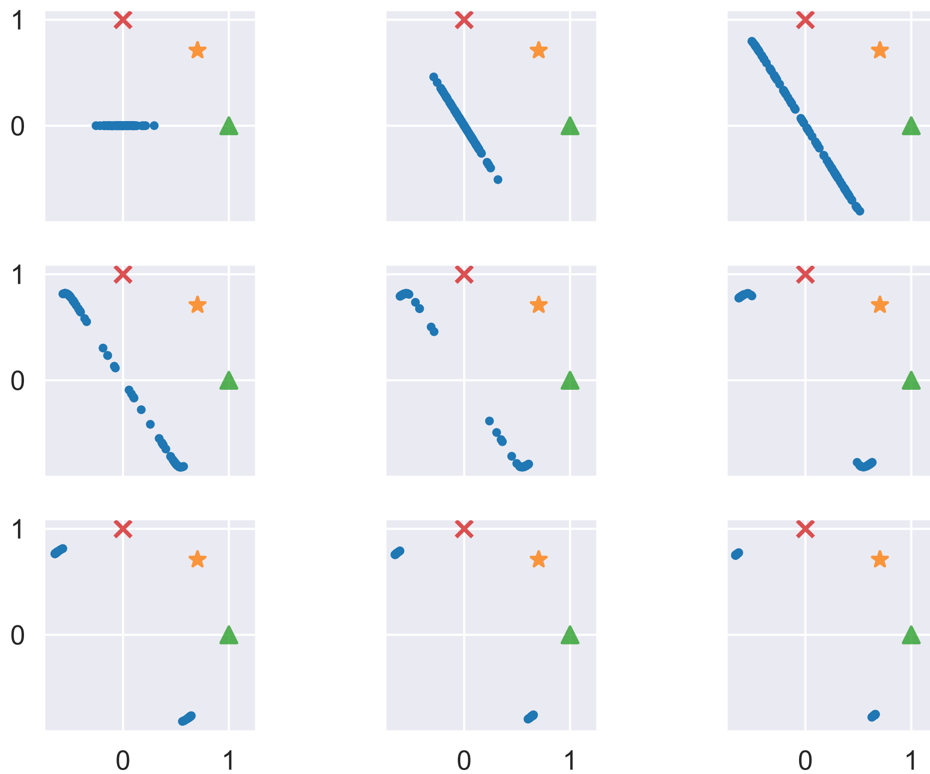

Figure 2: Classification setting, directional convergence. From left to right, top to bottom, we plot the directional information at timestamp $t = 0, 25, 50, \ldots, 200$, once every 25 iterations.

Directional convergence: classification Consider (1.9) with $\mathbf{y}|\mathbf{x} = x \sim \text{Bernoulli}\big(\sigma(\langle x, \theta^\star \rangle)\big)$ and $\ell(y', y) = -y'y + \log\big(1 + e^{y'}\big)$. The adversarial distribution shift evolves according to (1.12). Here the curse direction Δ_c defined in (2.2) is perpendicular to Yellow ★. Seen in Fig. 2.3, the Blue • data clouds eventually land on the directions $\{\pm\Delta_c\}$. Compared to Fig. 2.3, 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 3 and 4. 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 $|\langle \frac{x_t}{\|x_t\|}, \Delta_c \rangle| = 1$.

2.4 Assumptions and Intuition of the Proof

In this section, we first state the initial condition for Theorem 2 and discuss the assumption. Then we lay out the intuition of the technical proof behind Theorem 2.

Assumption 1 (Initial condition)**.** *Given fixed $\eta, r > 0$, we call two numbers (a_0, b_0) satisfy the initial condition, if*

$$\frac{e^{a_0+b_0}a_0}{1-e^{2(a_0+b_0)}} < 1\,, \tag{2.11}$$

$$\frac{e^{a_0+b_0}a_0}{1+e^{a_0+b_0}} \geq \frac{1+\frac{1}{a_0}}{1+r+\frac{1}{a_0}}\,, \tag{2.12}$$

and $a_0 > c$, $a_0 + b_0 < 0$, for some large enough constant $c > 0$.

Remark. When the initialization $a_0 > c$ with c not too small, the assumption holds for a range of b_0: (2.11) and (2.12) are equivalent to

$$a_0\frac{1+\sqrt{1+4a_0^{-2}}}{2} < e^{-(a_0+b_0)} < a_0\frac{1+r+a_0^{-1}}{1+a_0^{-1}} - 1\,. \tag{2.13}$$

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.

Now we elaborate on the intuition and technical innovation behind the proof of Theorem 2. 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

$$x_{t+1} = x_t + \gamma\cdot\Big[-\sigma(\langle x_t, \theta_\star\rangle)\Big(1-\sigma(\langle x_t,\theta_\star\rangle)\Big)\langle x_t, \theta^{(0)}\rangle\cdot\theta^\star + \Big(\sigma(\langle x_t,\theta^{(0)}\rangle) - \sigma(\langle x_t,\theta^\star\rangle)\Big)\cdot\theta^{(0)}\Big]$$

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

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

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

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

Recall that $b_0 = \langle x_0, \theta^\star - \theta^{(0)}\rangle = \langle x_0, \theta^\star - \Pi_{\mathrm{supp}(\mu^{(0)})}\theta^\star\rangle = 0$ since $\theta^{(0)} \in \mathrm{BR}(\mu^{(0)})$ and $x_0 \in \mathrm{supp}(\mu^{(0)})$. 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 3.

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)

$$L_t := \frac{\sigma'(a_t+b_t)a_t}{\sigma(a_t)-\sigma(a_t+b_t)} < 1 \tag{2.14}$$

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 $t \to \infty$, the quantity never cross below a threshold

$$L_t \geq \frac{1}{1+r},\ \forall t \geq t_0\,. \tag{2.15}$$

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 1.

The above intuition is educative but hard to directly operate upon over iterations, due to the non-linear form of (2.14) 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 4. We define

$$L_t^{\text{env-U}} := \frac{e^{a_t+b_t} a_t}{1-e^{2(a_t+b_t)}}, \text{ and } L_t^{\text{env-L}} := \frac{e^{a_t+b_t} a_t}{1+e^{a_t+b_t}} . \tag{2.16}$$

We show that these two envelopes are related to L_t in the following sense (Lemma 2), but not sandwiching it (we only have $L_t^{\text{env-L}} \le \min\{L_t, L_t^{\text{env-U}}\}$)

$$L_t^{\text{env-U}} < 1 \implies L_t < 1, \text{ and } L_t^{\text{env-L}} > \frac{1}{1+r} \implies L_t > \frac{1}{1+r} . \tag{2.17}$$

It turns out that the envelopes intervene cleanly with the non-linear recursions for $t \to t+1$. The crux of the argument lies in a strengthened version of the "rough observation" outlined in the previous paragraph, specifically done in Lemma 4. 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

$$L_t^{\text{env-L}} > \frac{1+a_t^{-1}}{1+r+a_t^{-1}} \implies L_{t+1}^{\text{env-L}} > \frac{1+a_{t+1}^{-1}}{1+r+a_{t+1}^{-1}} > \frac{1}{1+r} . \tag{2.18}$$

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.

References

[1] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[2] Charles J Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, pages 1348–1360, 1980.

[3] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[4] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

[5] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[6] Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

[7] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[8] Adel Javanmard and Mahdi Soltanolkotabi. Precise statistical analysis of classification accuracies for adversarial training. *The Annals of Statistics*, 50(4):2127–2156, August 2022. ISSN 0090-5364, 2168-8966. doi: 10.1214/22-AOS2180.

[9] Yoav Freund and Robert 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.

[10] Yoav Freund and Robert E Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103, 1999.

[11] Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

[12] Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

[13] Andrew Ross and Finale 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.

[14] Adel Javanmard, Mahdi Soltanolkotabi, and Hamed Hassani. Precise tradeoffs in adversarial training for linear regression. In *Conference on Learning Theory*, pages 2034–2078. PMLR, 2020.

[15] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

[16] Wenxuan Guo, YoonHaeng Hur, Tengyuan Liang, and Chris Ryan. Online learning to transport via the minimal selection principle. In *Proceedings of thirty fifth conference on learning theory*, volume 178 of *Proceedings of machine learning research*, pages 4085–4109. PMLR, July 2022.

[17] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

[18] Sébastien Bubeck, Yeshwanth Cherapanamjeri, Gauthier Gidel, and Remi 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.

[19] Matus Telgarsky. Margins, shrinkage, and boosting. In *International Conference on Machine Learning*, pages 307–315. PMLR, 2013.

[20] Tengyuan Liang and Pragya 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.

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144, 2020.

[22] Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*, 2015.

[23] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

[24] Tengyuan Liang. How Well Generative Adversarial Networks Learn Distributions. *Journal of Machine Learning Research*, 22(228):1–41, 2021. ISSN 1533-7928.

[25] Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 907–915. PMLR, 2019.

[26] Aryan Mokhtari, Asuman Ozdaglar, and Sarath 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.

[27] Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.

A Proofs

A.1 Proofs in Section 2.1

Proof of Theorem 1. In the regression setting, the utility evaluated at delta measure δ_x takes the form

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

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

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

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

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

It is then clear to see that

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

and thus, we can conclude the proof by noting

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

□

Proof of Theorem 2. Let $\sigma(z) = 1/(1+e^{-z})$ be the sigmoid function and $\sigma'(z) = \sigma(z)(1-\sigma(z))$. In the classification setting,

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

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

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

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

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

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

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

where $q := \langle \frac{\theta^\star - \theta^{(0)}}{\|\theta^\star - \theta^{(0)}\|}, \frac{\theta^{(0)}}{\|\theta^{(0)}\|} \rangle = 0$ and $r := \frac{\|\theta^\star - \theta^{(0)}\|^2}{\|\theta^{(0)}\|^2}$. 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 1,

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

Recall that the direction Δ_{c} can be written as

$$\Delta_{\mathrm{c}} = -\frac{1}{\sqrt{1+r}} \cdot \frac{\theta^\star - \theta^{(0)}}{\|\theta^\star - \theta^{(0)}\|} + \frac{\sqrt{r}}{\sqrt{1+r}} \cdot \frac{\theta^{(0)}}{\|\theta^{(0)}\|} \,.$$

Therefore

$$\begin{aligned}
\langle x_T, \Delta_{\mathrm{c}} \rangle &= -\frac{1}{\sqrt{1+r}} \cdot \frac{b_T}{\|\theta^\star - \theta^{(0)}\|} + \frac{\sqrt{r}}{\sqrt{1+r}} \cdot \frac{a_T}{\|\theta^{(0)}\|} \\
\|x_T\| &= \left[\frac{b_T^2}{\|\theta^\star - \theta^{(0)}\|^2} + \frac{a_T^2}{\|\theta^{(0)}\|^2} + \Big\langle x_0, \Pi^\perp_{\{\theta^{(0)}, \theta^\star - \theta^{(0)}\}} x_0 \Big\rangle \right]^{1/2} \,,
\end{aligned}$$

and thus we conclude the proof recalling Lemma 1

$$\begin{aligned}
\frac{\langle x_T, \Delta_{\mathrm{c}} \rangle}{\|x_T\|} &= \frac{(1+r) a_T - (a_T + b_T)}{\sqrt{1+r} \sqrt{b_T^2 + r a_T^2 + \|\theta^\star - \theta^{(0)}\|^2 \Big\langle x_0, \Pi^\perp_{\{\theta^{(0)}, \theta^\star - \theta^{(0)}\}} x_0 \Big\rangle}} \\
&= \frac{1 - \frac{1}{1+r} \frac{a_T + b_T}{a_T}}{\sqrt{1 - \frac{2}{1+r} \frac{a_T + b_T}{a_T} + \frac{1}{1+r} \left[\left(\frac{a_T + b_T}{a_T} \right)^2 + \frac{\langle x_0, \Pi^\perp_{\{\theta^{(0)}, \theta^\star - \theta^{(0)}\}} x_0 \rangle}{a_T^2} \right]}} \,.
\end{aligned}$$

□

Lemma 1 (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 $b_0 = 0$*

$$a_{t+1} = a_t - \eta \cdot \sigma'(a_t + b_t) a_t + \eta \cdot \Big(\sigma(a_t) - \sigma(a_t + b_t) \Big) \,, \tag{A.1}$$

$$b_{t+1} = b_t - \eta \cdot r \sigma'(a_t + b_t) a_t \,. \tag{A.2}$$

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

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

Proof of Lemma 1. a_{t_0} and b_{t_0} satisfy the conditions in Lemma 3, 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

$$ra_{t+1} - b_{t+1} = ra_t - b_t + \eta \cdot r\Big(\sigma(a_t) - \sigma(a_t + b_t)\Big)\,, \tag{A.3}$$

$$a_{t+1} + b_{t+1} = \Big(1 - \eta\sigma'(a_t + b_t)\Big)(a_t + b_t) - \eta\sigma'(a_t + b_t)(ra_t - b_t) + \eta\Big(\sigma(a_t) - \sigma(a_t + b_t)\Big)\,. \tag{A.4}$$

Below we use the Bachmann–Landau notation: we say two sequences of positive real numbers $x_t = O(z_t)$ iff $\limsup_{t\to\infty} x_t/z_t < \infty$, $x_t = \Omega(z_t)$ iff $\liminf_{t\to\infty} x_t/z_t > 0$, 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 (A.1)-(A.2).

2. Claim $ra_t - b_t \to +\infty$. Proof by contradiction. If the monotonic sequence $ra_t - b_t$ is uniformly upper bounded by M. Then $\sigma(a_t) - \sigma(a_t + b_1) \leq \sigma(a_t) - \sigma(a_t + b_t) = o(1)$, which implies $a_t \to \infty$, which contradicts $ra_t \leq ra_t - b_t \leq M$.

3. Claim $b_t \to -\infty$. Proof by contradiction. If b_t is bounded below by $-M$, then $\eta \cdot r\sigma'(a_t + b_t)a_t \to 0$. Note $a_t \geq a_{t_0} > 0$, then $\sigma'(a_t + b_t) \to 0$. Recall that $a_t + b_t \leq a_{t_0} + b_{t_0} < 0$ and is monotonic decreasing, then $a_t + b_t \to -\infty$ and thus b_t must go to negative infinity. Contradiction.

4. Claim $a_t \to +\infty$. Proof by contradiction, if a_t bounded by above by M, then $a_t + b_t \to -\infty$, then $a_{t+1} - a_t \geq \eta\Big(\frac{1}{2} - \sigma(a_t + b_t)\Big) - \eta\sigma'(a_t + b_t)M \geq \eta\Big(\frac{1}{2} - (M+1)\sigma(a_t + b_t)\Big) \geq \frac{1}{4}\eta$ for t large enough. Contradiction.

5. Claim $a_t + b_t \to -\infty$. If $|a_t + b_t| \leq M$ for some absolute constant M, then (A.4) is contradicted as we have proved $ra_t - b_t \to +\infty$.

6. Claim $\lim\limits_{t\to\infty} \frac{ra_t - b_t}{t} = \eta r$. We have shown that $\lim_{t\to\infty} a_t \to \infty$ and $\lim_{t\to\infty} a_t + b_t = -\infty$.

$$\lim_{t\to\infty} \sigma(a_t) - \sigma(a_t + b_t) = 1\,. \tag{A.5}$$

   Therefore by (A.3) we can show $\lim\limits_{t\to\infty} \frac{ra_t - b_t}{t} = \eta r$.

7. Claim $\liminf\limits_{t\to\infty} \sigma'(a_t + b_t)(ra_t - b_t) \geq \liminf\limits_{t\to\infty} r\sigma'(a_t + b_t)a_t \geq \frac{r}{1+r}$. 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 3.

8. Claim $\limsup\limits_{t\to\infty} \frac{|a_t+b_t|}{\log(t)} \leq 1$. 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 2. *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 of Lemma 2. $G_\delta(x,z)$ 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

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

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 3. *Fix $\eta, r > 0$. Consider the recursive relationship defined in Lemma 1. Assume there exists some $t_0 \in \mathbb{N}$, such that (a_{t_0}, b_{t_0}) satisfy Assumption 1. Then for $t \geq t_0$*

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

and

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

Proof of Lemma 3. The proof relies on induction. By assumption at $t = t_0$, we know

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

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

$$\frac{e^{a_t+b_t}a_t}{1-e^{2(a_t+b_t)}} \leq \frac{1}{1+\epsilon}\,,\ \frac{e^{a_t+b_t}a_t}{1+e^{a_t+b_t}} \geq \frac{1+\frac{1}{a_t}}{1+r+\frac{1}{a_t}}\,. \tag{A.6}$$

Apply Lemma 2, 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

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

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

$$a_{t+1} = a_t + \eta G_0(a_t, a_t+b_t) > a_t > c\,, \tag{A.7}$$

$$a_{t+1} + b_{t+1} = a_t + b_t + \eta G_r(a_t, a_t+b_t) < a_t + b_t + \eta G_{\frac{r}{1+1/a_t}}(a_t, a_t+b_t) < a_t + b_t\,. \tag{A.8}$$

Now for the induction step from t to $t+1$, we apply the Lemma 4. The assumptions for Lemma 4 are verified, and therefore

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

Repeat the induction step we can complete the proof. □

Lemma 4. *Fix $\eta, r > 0$. Consider the one-step recursive relationship defined in Lemma 1. Assume (a_t, b_t) satisfy the Assumption 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*

$$\frac{e^{a_{t+1}+b_{t+1}}a_{t+1}}{1-e^{2(a_{t+1}+b_{t+1})}} < \frac{e^{a_t+b_t}a_t}{1-e^{2(a_t+b_t)}}, \tag{A.9}$$

$$\frac{e^{a_{t+1}+b_{t+1}}a_{t+1}}{1+e^{a_{t+1}+b_{t+1}}} > \frac{1+\frac{1}{a_{t+1}}}{1+r+\frac{1}{a_{t+1}}}. \tag{A.10}$$

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

Proof of Lemma 4. The proof consists of two parts: a recursive estimate for the upper bound to prove (A.9) and then a recursive estimate for the low bound in (A.10). First, due to (A.7) and (A.8), 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

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

In addition, we can prove that there exists a constant ϵ as in (A.6) such that

$$G_\epsilon(a_t, a_t+b_t) > 0$$

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

Lower bound recursion Define

$$\underline{r_t} := \frac{1+\frac{1}{a_t}}{1+r+\frac{1}{a_t}}$$

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.

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

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

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

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

$$
\tilde{r}_{t+1} > \tilde{r}_t \Big[1 - \tilde{\eta}\Big(\tilde{r}_t - \underline{r_{t+1}}\Big)\Big]\Big[1 - \eta(1+c^{-1})e^{-a_t}\Big] . \tag{A.11}
$$

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 $\eta(1+c^{-1})e^{-a_t} < \tilde{\eta}e^{-a_t} < \tilde{\eta}\Big(\underline{r_t} - \underline{r_{t+1}}\Big) < \tilde{\eta}\Big(\tilde{r}_t - \underline{r_{t+1}}\Big)$, and thus (A.11) reads

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

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

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

The lower bound on Eqn. (A.10) is true as long as the RHS of the above is positive

$$
(\tilde{r}_t - \underline{r_{t+1}})[1 - 2\tilde{\eta}\tilde{r}_t + \tilde{\eta}^2\tilde{r}_t(\tilde{r}_t - \underline{r_{t+1}})] > 0 . \tag{A.12}
$$

Note $\tilde{\eta} < 1/2$, $\tilde{r}_t < 1$ and $\tilde{r}_t > \underline{r_t} > \underline{r_{t+1}}$, we have concluded

$$
\frac{e^{a_{t+1}+b_{t+1}}a_{t+1}}{1+e^{a_{t+1}+b_{t+1}}} = \tilde{r}_{t+1} > \underline{r_{t+1}} = \frac{1+\frac{1}{a_{t+1}}}{1+r+\frac{1}{a_{t+1}}} . \tag{A.13}
$$

□

A.3 Proofs in Section 2.2

Proof of Theorem 3.

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

Fix n, first let $T \to \infty$. Theorem 1 shows that $x_T^i \to \Delta_\mathrm{b}$ (we denote the dependence on T explicitly) in the following sense

$$\langle x_T^i, \Delta_\mathrm{b} \rangle^2 \geq 1 - O(e^{-cT}) \,. \tag{A.14}$$

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

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

For the first term (i), along the direction Δ_b,

$$\langle \Delta_\mathrm{b}, \Gamma \rangle = \|\theta^{(0)} - \theta_\star\| \cdot (1 - \frac{1}{n}\sum_{i=1}^n \langle x^i, \Delta_\mathrm{b}\rangle^2) \,.$$

One the orthogonal complement to Δ_b,

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

Therefore

$$\lim_{T\to\infty} \|\Gamma\| \leq \lim_{T\to\infty} \left\{ \|\theta^{(0)} - \theta_\star\| \cdot (1 - \frac{1}{n}\sum_{i=1}^n \langle x_T^i, \Delta_\mathrm{b}\rangle^2) + \frac{1}{n}\sum_{i=1}^n \langle x_T^i, \Delta_\mathrm{b}\rangle \sqrt{1 - \langle x_T^i, \Delta_\mathrm{b}\rangle^2} \right\} = 0 \,.$$

For the second term, due to the Hanson-Wright inequality,

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

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

Proof of Theorem 4.

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

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

Fixed n, first let $T \to \infty$. Theorem 1 shows that $x_T^i \to \Delta_\mathrm{c}$, where the convergence means directional convergence, and we denote the dependence on T explicitly. Then as $\lim_{T\to\infty} \langle \Delta_\mathrm{c}, x_T^i \rangle = 1$ and $\|x_T^i\| = 1$, we know,

$$\lim_{T\to\infty} \langle \tfrac{\theta^\star}{\|\theta^\star\|}, x_T^i \rangle^2 = 1 - \lim_{T\to\infty} \langle \Delta_\mathrm{c}, x_T^i \rangle^2 = 0 \,,$$

and therefore

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

The final claim is a direct fact from

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

Take lim inf w.r.t. to n, and we conclude the proof. □