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

Tengyuan Liang

University of Chicago

December 5, 2022

Abstract

Covariate distribution shifts and adversarial perturbations present robustness challenges to the conventional statistical learning framework: seemingly small unconceivable 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, more needs to be studied 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, an unspoken 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 \( Y \mid X \) changes hence the underlying Bayes optimal prediction model \( f^\ast : X \to Y \) could shift,
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 extrapolates the training data distribution.

Concept shift is inherently a difficult problem as if shooting for a moving target. However, covariate shift may not be as severe a problem as long as the concept does not vary. Specific statistical methods work with mild extrapolation\(^2\), say (fixed-design) linear regression and (local) nonparametric regression [2]. Recently, another notable line of work is [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 revived interest in machine learning and robust optimization recently [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 \( \mu \): the model \( f \) minimizes the risk from a model class, yet the covariate distribution \( \mu \) maximizes the risk from a probability distribution class.

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 discussion 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 \( \pi \), one can define the utility function that accesses the risk of model \( f \) on data \( \pi \)

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

**Models, covariate distributions, and Bayes optimality** Following the literature [3, 9], 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

\[
\pi_\mu := \int \delta_x \otimes \pi_\mu^* \, d\mu(x) \in \mathcal{P}(X \times Y),
\]

where \( \pi_\mu^* \in \mathcal{P}(Y) \) denotes a fixed conditional data generating process for \( y|x = x \) that does not vary with \( \mu \). \((1.1)\) should be read as disintegration of measure [10], meaning for all bounded continuous function \( h \in C_b(X \times Y) \)

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

\(^1\)One typical evaluation metric is \( \| f - f^* \|_{L^2(\mu)} \), 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.

---

2
Given a fixed conditional distribution $\pi^\star_x$ and a loss function $\ell(\cdot, \cdot)$, one can thus define the Bayes optimal model (suppose for now that this map is well-defined)

$$f^\star_{\text{Bayes}}: x \mapsto \arg\min_{y' \in Y} \int \ell(y', y) d\pi^\star_x(y).$$

(1.2)

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 $f: X \to Y$ and the covariate distribution $\mu \in P(X)$,

$$\mathcal{U}(f, \mu) := \mathcal{R}(f, \pi^\mu) = \int_X \left[ \int_Y \ell(f(x), y) d\pi^\star_x(y) \right] d\mu(x).$$

(1.3)

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

Adversarial perturbation The theoretical insights toward understanding adversarial perturbations have so far centered around robustness and regularization [6, 11, 7]. Given a metric measure space $(P(X), d)$, a covariate distribution $\mu \in P(X)$, and a current model $f$, consider the following population version of the adversarial perturbation,

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

(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

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

(1.5)

Moreau-Yosida regularization

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 $\lambda^{-1}$, 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 $\nu$ differs from $\mu$. It is worth noting that under a host of Gaussian covariate models, [12, 8] provided sharp high-dimensional asymptotic characterization of the adversarial risk $\mathcal{U}_\gamma(\hat{f}, \mu)$ and the standard risk $\mathcal{U}(\hat{f}, \mu)$ and thus shed light on the tradeoffs, for a range of models $\hat{f}$ 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 $\gamma \in \mathbb{R}_+$,

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

(1.6)
corresponding to (1.5). Denote the distribution shift map \( Ds_\gamma : x \mapsto \arg\min_{x' \in \mathcal{X}} \left( -\mathcal{U}(f, \delta_{x'}) + \frac{1}{\gamma} \| x' - x \|^2 \right) \) 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 
\[
\nu = (Ds_{\lambda^{-1}}) \rho
\]
where \( \lambda > 0 \) is the solution to the dual. In the infinitesimal limit \( \gamma \sim \lambda^{-1} \to 0 \), one can show that [13, 14]
\[
\frac{Ds_\gamma - \text{Id}}{\gamma} \rightarrow \frac{\partial}{\partial x} \mathcal{U}(f, \delta_x). \quad (1.7)
\]
The distribution shift map \( Ds_\gamma \) presents a way of constructing adversarial examples [5, 15, 16], 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 \( \rho_t \) (associated with \( \nu_t \)) evolves according to the following PDE [13]
\[
\partial_t \rho_t + \nabla \cdot (\rho_t V) = 0, \text{ where } V : x \mapsto \frac{\partial}{\partial x} \mathcal{U}(f, \delta_x). \quad (1.8)
\]

### 1.2 Problem Setup

This paper considers regression and classification problems in a possibly infinite-dimensional setting. Let \( \mathcal{X} \subset \mathbb{R}^N \) be a subset of a possibly infinite-dimensional space for the covariates, and \( \mathcal{Y} \subset \mathbb{R} \) be the space for a real-valued response variable. We concern a infinite-dimensional linear model class \( \mathcal{F} := \{ f_\theta \mid f_\theta(x) := \langle x, \theta \rangle, \theta \in \ell_2^N \} \) where the inner-product corresponds to the Hilbert space \( \ell_2^N \). Slightly abusing the notation, we write for convenience the utility function
\[
\mathcal{U}(\theta, \mu) = \mathbb{E}_{(x, y) \sim \mu} \left[ \ell(f_\theta(x), y) \right] = \int \int \ell(f_\theta(x), y) \, d\pi^*_x(y) \, d\mu(x). \quad (1.9)
\]
We investigate two types of conditional relationships for \( \pi^*_x \) in (1.1), namely for some \( \theta^* \in \ell_2^N \):

- **Regression:** \( y | x \sim \text{Gaussian}((x, \theta^*), 1) \), \( \ell(f, y) = (f - y)^2 \);  
- **Classification:** \( y | x \sim \text{Bernoulli}(\sigma((x, \theta^*)) \), \( \ell(f, y) = -f \cdot y + \log(1 + e^f) \).

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

**Game and equilibria** In the game between the model \( \theta \in \ell_2^N \) and the covariate distribution \( \mu \in \mathcal{P}(\mathcal{X}) \),
\[
\min_{\theta} \max_{\mu} \mathcal{U}(\theta, \mu) \geq \max_{\mu} \min_{\theta} \mathcal{U}(\theta, \mu) \geq \max_{\mu} \int_X \left[ \min_{\theta \in \ell_2^N} \int_Y \ell(f_\theta(x), y) \, d\pi^*_x(y) \right] \, d\mu(x) \quad (1.10)
\]
\[
= \max_{\mu} \mathcal{U}(\theta^*, \mu) \geq \min_{\mu} \max_{\theta} \mathcal{U}(\theta, \mu). 
\]

Therefore the min-max theorem holds in this infinite-dimensional context when the \( f_{\text{Bayes}}^*(x) = \langle x, \theta^* \rangle \) is well-defined. A Nash equilibrium of the \( \mathcal{U}(\cdot, \cdot) \) 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. The celebrated boosting literature [17, 18, 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 \( \mu \) 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 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 \( \text{supp}(\mu^{(0)}) \subset X = \ell_2^N \) (so that extrapolation is meaningful), the best response model \( f_{\theta^{(0)}} \in \mathcal{F} \) solves the following risk minimization associated with measure \( \mu^{(0)} \)

\[
\theta^{(0)} \in \text{BR}(\mu^{(0)}) := \arg\min_{\theta \in \ell_2^N} \mathcal{U}(\theta, \mu^{(0)}).
\]

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

\[
\theta^{(0)} \in \text{BR}(\mu^{(0)}) = \left\{ \Pi_{\text{supp}(\mu^{(0)})} \theta^* + \Pi_{\perp \text{supp}(\mu^{(0))}} \xi \mid \forall \xi \in \ell_2^N \right\},
\]

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 \( \Pi_{\text{supp}(\mu^{(0)})} \theta^* \), 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 \( \Pi_{\text{supp}(\mu^{(0)})} \theta^* \). Clearly, \( \theta^{(0)} = \Pi_{\text{supp}(\mu^{(0)})} \theta^* \) is inconsistent with the Bayes optimal model \( \theta^* \) [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 \( \theta^* \). 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 \( v_0 := \mu^{(0)} \),

\[
v_{t+1} := \arg\min_{v \in \mathcal{P}(X)} -\mathcal{U}(\theta^{(0)}, v) + \frac{1}{\gamma} W_2^2(v, v_t), \text{ for } t = 0, 1, \ldots, T,
\]

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

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

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 \( \gamma T \). In this paper, we study a discretization of (1.11) with stepsize \( \gamma \) and iterations \( T \), as follows

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

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 fast to the hardest experimental design trapping subsequent learning.

2 Main Results

2.1 Adversarial Covariate Shifts: Blessings and Curses

Let $\theta^{(0)} \in \ell^2_N$ be the current learning model and $\theta^* - \theta^{(0)}$ be the remaining signal to be identified. Define two unit-norm directions: the blessing direction $\Delta_b \in \mathbb{R}^N$ and the curse direction $\Delta_c \in \mathbb{R}^N$

\begin{align}
\Delta_b &:= \frac{\theta^* - \theta^{(0)}}{\|\theta^* - \theta^{(0)}\|} \in \ell^2_N(1), \\
\Delta_c &:= -\frac{\|\theta^{(0)}\|}{\|\theta^*\|} \cdot \frac{\theta^* - \theta^{(0)}}{\|\theta^* - \theta^{(0)}\|} + \frac{\|\theta^* - \theta^{(0)}\|}{\|\theta^{(0)}\|} \cdot \frac{\theta^{(0)}}{\|\theta^{(0)}\|} \in \ell^2_N(1).
\end{align}

The name blessing comes from the fact that $\Delta_b//\theta^* - \theta^{(0)}$, 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 $\Delta_b \perp \theta^*$, 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 $\Delta_b$ 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 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 $y|x = x \sim \text{Gaussian}(x, \theta^*)$, $1$. 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 (1.12) satisfies

$$\lim_{T \to \infty} \left| \left\langle \frac{x_T}{\|x_T\|}, \Delta_b \right\rangle \right| = 1,$$

where $\Delta_b//\theta^* - \theta^{(0)}$ is defined in (2.1).

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

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

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

Remark. 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 $\pi^*_\theta$ and hence identifying the equilibrium $\hat{f}^{\text{Bayes}}_{\text{opt}} : x \mapsto \arg\min_{y \in Y} \int \ell(y', y) d\pi^*_\theta(y)$. The current model could be imperfect due to (1) 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
Moreover, the directional convergence is quadratic in $T/t$ the equilibrium of learning, the Bayes optimal model for the next stage of learning: making the current model $\theta$ along a one-dimensional “blessing” direction $\Delta_{\theta}$, 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 $\theta$ 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 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 $\ell(y', y) = -y'y + \log(1 + e^{y'y})$ and $y|x = x \sim \text{Bernoulli}((x, \theta^*))$. Assume $\theta^{(0)} \perp \theta^* - \theta^{(0)}$. Let $x_0 \in \text{supp}(\mu^{(0)})$ and assume there exists a $t_0 \in \mathbb{N}$ such that $(a_0, b_0) := \left((x_{i_0}, \theta^{(0)}), (x_{i_0}, \theta^* - \theta^{(0)})\right)$ satisfying Assumption 1. Then the induced adversarial distribution shift dynamic (1.12) satisfies

$$\lim_{T \to \infty} \left| \left\langle \frac{\chi_T}{\|x\|}, \Delta_{\theta} \right\rangle \right| = 1, \text{ where } \Delta_{\theta} \perp \theta^* \text{ is defined in (2.2)}. \quad (2.5)$$

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

$$\left| \left\langle \frac{\chi_T}{\|x\|}, \Delta_{\theta} \right\rangle \right| \in \left[ 1 - O\left(\frac{\log^2(T)}{T^2}\right), 1 \right]. \quad (2.6)$$

**Remark.** This theorem concerns also the case when the current model $\theta^{(0)}$ is imperfect—namely $\|\theta^* - \theta^{(0)}\| \neq 0$. In particular, this theorem studies when $\text{supp}(\mu^{(0)}) \subseteq X$. As stated before, the best response model $\theta^{(0)}$ associated with the measure $\mu^{(0)}$ takes the form $\theta^{(0)} \in \text{BR}(\mu^{(0)}) = \left\{ \prod_{\text{supp}(\mu^{(0)})} \theta^* + \prod_{\text{supp}(\mu^{(0))}} \xi \mid \forall \xi \in \ell^2_N \right\}$. 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 $\mu^{(0)} \to \mu^{(1)}$ asymptotically align all the mass of the covariates along a one-dimensional, “curse” direction $\Delta_{\theta} \perp \theta^*$, 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. 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 4. Quantitatively, the directional convergence to $\Delta_{\theta}$ in the classification setting is quadratic in $T$, much slower than the directional convergence to $\Delta_{\theta}$ 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, c.f. 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 difficulties 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 the directional convergence results established
in Theorems 1 and 2 can translate to a direct impact on the learner’s subsequent learning progress towards the Bayes optimal model \(f_{\text{Bayes}}^*(x) = \langle x, \theta^* \rangle\), the equilibrium. The impact is dramatic, and either comes in as a blessing or a curse.

The sequential game between model \(f_0\) 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 \(\tilde{\mu}(0) \rightarrow \tilde{\mu}(1)\) given the previous model \(\theta(0)\). Here \(\tilde{\mu}(0) = \frac{1}{n} \sum_{i=1}^n \delta_{x_i/\|x_i\|}\) where \(x_i \sim \mu(0)\), and \(\tilde{\mu}(1) := \frac{1}{n} \sum_{i=1}^n \delta_{x_i'/(\|x_i\|)}\) where \(x_i'\) 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 \(\tilde{\mu}(1)\). Draw \(n\)-i.i.d. samples \((x^i, y^i) \sim \tilde{\pi}(\tilde{\mu}(1))\), defined in (1.1), and update
\[
\theta(1) = \theta(0) - \eta \cdot \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} \ell((x^i, \theta), y^i)|_{\theta = \theta(0)}.
\]

(2.7)

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 \(\tilde{\mu}(1)\) 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 \(\|\theta^* - \theta(0)\| > 0\), the learner’s one-step reaction to the distribution shift as in (2.7) with \(\eta = 1/2\) satisfies
\[
\lim_{n \to \infty} \lim_{T \to \infty} \|\theta^* - \theta_{n,T,\eta}(1)\| = 0 \text{ a.s.}.
\]

(2.8)

**Theorem 4** (Classification: curse to the learner). Consider the same setting as in Theorem 2. For all \(\theta(0)\) such that \(\langle \theta^* - \theta(0), \theta^* \rangle \neq 0\), the learner’s one-step reaction to the distribution shift as in (2.7) with any fixed \(\eta > 0\) satisfies
\[
\lim_{n \to \infty} \lim_{T \to \infty} \frac{\langle \theta^* - \theta_{n,T,\eta}(1), \theta^* \rangle}{\langle \theta^* - \theta(0), \theta^* \rangle} = 1.
\]

(2.9)

Moreover,
\[
\liminf_{n \to \infty} \lim_{T \to \infty} \|\theta^* - \theta_{n,T,\eta}(1)\| > 0.
\]

(2.10)

**Remark.** The result in Corollary 4 holds valid for any fixed number of gradient descent steps for the learner
\[
\theta(t+1) = \theta(t) - \eta \cdot \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} \ell((x^i, \theta), y^i)|_{\theta = \theta(t)}.
\]

Therefore, the learner gets stuck with the adversarial distribution \(\tilde{\mu}(1)\) in making progress towards \(\theta^\ast\). 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 $\text{supp}(\mu(0)) = \mathbb{R}^{100}$ a randomly drawn subspace from the Haar measure (a uniformly drawn subspace of dimension 100). Specify a Bayes optimal model $\theta^\star = [1, 1/2, \ldots, 1/i, \ldots, 1/200]^\top$, denoted by the the **Yellow star** in the figures. The best response model restricted to the subspace $\text{supp}(\mu(0))$, $\theta(0)$ is noted by **Green triangle**, and the remaining signal $\theta^\star - \theta(0)$ is denoted by **Red cross**. The probability distribution of the covariates is shown by the **Blue dot**. 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 movement of the **Blue dot** 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 $\{\frac{\theta^\star}{\|\theta^\star\|}, \frac{\theta^\star - \theta(0)}{\|\theta^\star - \theta(0)\|}\}$.

**Directional convergence: regression** Consider (1.9) with $y | x = x \sim \text{Gaussian}(\langle x, \theta^\star \rangle, 1)$ and $\ell(y', y) = (y' - y)^2$. The adversarial distribution shift evolves according to (1.12). Here the blessing direction $\Delta_b$ defined in (2.1) is precisely marked by **Red cross**. As seen in Fig. 2.3, the **Blue dot** 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 dot** 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 (1.9) with $y | x = x \sim \text{Bernoulli}(\sigma(\langle x, \theta^\star \rangle))$ and $\ell(y', y) = -y'y + \log(1 + e^{y'})$. 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 {±Δ_c}. Compared to Fig. 2.3, the direction {±Δ_c} is different from the regression case {±Δ_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 \(\langle \frac{x_t}{\|x_t\|}, \Delta_c \rangle = 1\).

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.

### 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

$$1 + \frac{\sqrt{1 + 4a_0^{-2}}}{2} < e^{-a_0^2} < 1 + \frac{1 + a_0^{-1}}{1 + a_0^{-1}} - 1. \quad (2.13)$$

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

Now we elaborate on intuitions 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

$$x_{t+1} = x_t + \gamma \cdot \left[ -\sigma'(x_t, \theta) \left( 1 - \sigma(x_t, \theta) \right) \langle x_t, \theta(0) \rangle \cdot \theta - \left( \sigma(x_t, \theta(0)) - \sigma(x_t, \theta^*) \right) \cdot \theta(0) \right]$$

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

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

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

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

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 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)

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

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

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

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

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

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

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

11
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 4. On the one hand, if the lower envelope function $L_{t+1}^{\text{env-L}} > \frac{1 + a_t^{-1}}{1 + r + a_t^{-1}} \in \left[ \frac{1}{1 + r}, 1 \right]$, 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

$$L_{t}^{\text{env-L}} > \frac{1 + a_t^{-1}}{1 + r + a_t^{-1}} \Rightarrow L_{t+1}^{\text{env-L}} > \frac{1 + a_{t+1}^{-1}}{1 + r + a_{t+1}^{-1}} > \frac{1}{1 + r}. \quad (2.18)$$

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 nothing but a mild condition requiring the dynamic of $L_t$ to reach below 1. All the detailed proof can be found in Appendix A.

References


A Proofs

A.1 Proofs in Section 2.1

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

$$\mathcal{U}(\theta, \delta_x) = \int_Y \ell(f_0(x), y) \, d\pi_x^*(y) = \langle x, \theta^* - \theta \rangle^2 + 1.$$  

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

$$x_{t+1} = x_t + \gamma \cdot \frac{\partial}{\partial x} \mathcal{U}(\theta^{(0)}, \delta_{x_t}),$$

$$= \left[ I + 2\gamma(\theta^* - \theta^{(0)})(\theta^* - \theta^{(0)})^\top \right] x_t,$$

$$= (I + \tilde{\gamma} \Delta_b \Delta_b^\top) x_t,$$

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

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

It is then clear to see that

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

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

and thus, we can conclude the proof by noting

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

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,

$$\mathcal{U}(\theta, \delta_x) = \int_Y \ell(f_0(x), y) \, d\pi_x^*(y) = \frac{1}{1 + e^{-(x, \theta^*)}} \log \left( 1 + e^{-(x, \theta^*)} \right) + \frac{1}{1 + e^{(x, \theta^*)}} \log \left( 1 + e^{(x, \theta^*)} \right).$$

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

$$x_{t+1} = x_t + \gamma \cdot \frac{\partial}{\partial x} \mathcal{U}(\theta^{(0)}, \delta_{x_t}),$$

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

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

$$a_{t+1} = a_t + \eta \cdot (1 + \sqrt{\eta}) \sigma'(a_t + b_t) a_t + \eta \cdot \left( \sigma(a_t) - \sigma(a_t + b_t) \right),$$

$$b_{t+1} = b_t + \eta \cdot (\sqrt{\eta} + r) \sigma'(a_t + b_t) a_t + \eta \cdot \sqrt{\eta} \left( \sigma(a_t) - \sigma(a_t + b_t) \right).$$
where \( q := \langle \theta^r - \theta^{(0)} \rangle, \frac{\theta^{(0)}}{\|\theta^{(0)}\|} \rangle = 0 \) and \( r := \|\theta^r - \theta^{(0)}\|^2 \). 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,

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

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

\[
\langle x_T, \Delta_c \rangle = -\frac{1}{\sqrt{1 + r}} \cdot \frac{b_T}{\|\theta^* - \theta^{(0)}\|^2} + \frac{\sqrt{r}}{\sqrt{1 + r}} \cdot \frac{a_T}{\|\theta^{(0)}\|}
\]

\[
\|x_T\| = \left[ \frac{b_T^2}{\|\theta^* - \theta^{(0)}\|^2} + \frac{a_T^2}{\|\theta^{(0)}\|^2} + \left( x_{0,t, \Pi_{\theta^* \theta^{(0)} \theta^{(0)}}} \right) \right]^{1/2},
\]

and thus we conclude the proof recalling Lemma 1

\[
\frac{\langle x_T, \Delta_c \rangle}{\|x_T\|} = \frac{(1 + )a_T - (a_T + b_T)}{\sqrt{1 + r} \cdot \left( b_T + ra_T \right) + \|\theta^* - \theta^{(0)}\|^2 \left( x_{0,t, \Pi_{\theta^* \theta^{(0)} \theta^{(0)}}} \right)} = \frac{1 - \frac{a_T + b_T}{a_T}}{1 + \frac{1}{1 + r} \left( \frac{a_T + b_T}{a_T} \right)^2 + \left( x_{0,t, \Pi_{\theta^* \theta^{(0)} \theta^{(0)}}} \right)} \]

\[\square\]

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 \( b_0 = 0 \)

\[
a_{t+1} = a_t - \eta \cdot \sigma'(a_t + b_t) a_t + \eta \cdot \left( \sigma(a_t) - \sigma(a_t + b_t) \right), \quad (A.1)
\]

\[
b_{t+1} = b_t - \eta \cdot r \sigma'(a_t + b_t) a_t \quad \text{.} \quad (A.2)
\]

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

\[
|a_T + b_T| = O(\log(T)), \quad a_T = \Theta(T).
\]

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

\[
a_{t+1} + b_{t+1} < a_t + b_t < 0 \text{ for } t \geq t_0 \text{.}
\]

Given the monotonicity, the proof proceeds in the following steps. First, note that

\[
ra_{t+1} - b_{t+1} = ra_t - b_t + \eta \cdot \left( \sigma(a_t) - \sigma(a_t + b_t) \right), \quad (A.3)
\]

\[
a_{t+1} + b_{t+1} = \left( 1 - \eta \sigma'(a_t + b_t) (a_t + b_t) - \eta \sigma'(a_t + b_t) (ra_t - b_t) + \eta \left( \sigma(a_t) - \sigma(a_t + b_t) \right) \right). \quad (A.4)
\]

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 \( ra_t - b_t \rightarrow +\infty\). 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 \( a_t \rightarrow \infty \), which 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 \( a_t \geq a_t > 0 \), then \( \sigma'(a_t + b_t) \rightarrow 0 \). Recall that \( a_t + b_t \leq a_t + b_t < 0 \) and is monotonic decreasing, then \( a_t + b_t \rightarrow -\infty \) and thus \( b_t \) must go to negative infinity. 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 \( a_t + a_t - a_t \geq \eta(\frac{1}{2} - \sigma(a_t + b_t)) - \eta \sigma'(a_t + b_t)M \geq \eta(\frac{1}{2} - (M + 1)\sigma(a_t + b_t)) \geq \frac{1}{4} \eta \) for \( t \) large enough. Contradiction.

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

6. Claim \( \lim_{t \to -\infty} \frac{ra_t - b_t}{t} = \eta r \). 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 \( \lim_{t \to -\infty} \frac{ra_t - b_t}{t} = \eta r \).

7. Claim \( \liminf_{t \to -\infty} \frac{\sigma'(a_t + b_t)(ra_t - b_t)}{t} \geq \liminf_{t \to -\infty} r \sigma'(a_t + b_t)a_t \geq \frac{\eta}{1+r} \). Here the last inequality uses the fact \( \frac{\sigma'(a_t + b_t)a_t}{\sigma(a_t) - \sigma(a_t + b_t)} \in \left[ \frac{1+\frac{2}{1+r}}{1-r} \right] \), established in Lemma 3.

8. Claim \( \limsup_{t \to -\infty} \frac{a_t + b_t}{\log(t)} \leq 1 \). 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 \( \eta = \frac{ra_t - b_t}{t + 1} + \Theta(t) = \Theta(t) - O(\log(t)) = \Theta(t) \). Similarly, we can show \( -b_t = \Theta(t) \).

\( \square \)

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 \( x > \frac{1}{1+t}(e^{-z} + 1) \), then \( G_\delta(x, z) < 0 \).
- Fix \( z < 0 \) and \( \delta \in [0, 1] \). If \( x < \frac{1}{1+t}(e^{-z} - e^z) \), 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 \( x_0 \leq \frac{1}{1+t} \frac{1 - \sigma(z)}{\sigma'(z)} \) and first claim is thus established.

The second claim follows as

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

16
This is because $\inf_{z \leq 0} \frac{\sigma(-z) - \sigma(z)}{\sigma(z)} \geq 2 > 1 + \delta$. Denote $x_0 > 0$ to be the unique solution to $G_0(x_0, z) = 0$, 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 $x_0 > \frac{1}{1 + \delta} \frac{\sigma(0) - \sigma(0)}{\sigma(0)}$. 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 $t_0, b, a \in \mathbb{N}$, such that $(a_t, b_t)$ 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

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

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

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

Apply Lemma 2, we know $G_0(a_t, a_t + b_t) > G_e(a_t, a_t + b_t) \geq 0$, and $G_{\frac{r}{1+\delta}}(a_t, a_t + b_t) < 0$, which implies

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

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

$$a_{t+1} = a_t + \eta G_0(a_t, a_t + b_t) > a_t > c, \quad (A.7)$$
$$a_{t+1} + b_{t+1} = a_t + b_t + \eta G_e(a_t, a_t + b_t) < a_t + b_t + \eta G_{\frac{r}{1+\delta}}(a_t, a_t + b_t) < a_t + b_t. \quad (A.8)$$

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

$$\frac{e^{a_{t+1} + b_{t+1}}a_{t+1}}{1 - e^{2(a_{t+1} + b_{t+1})}} < \frac{e^{a_t + b_t}a_t}{1 - e^{2(a_t + b_t)}} \leq \frac{1 + \frac{1}{a_t}}{1 + r + \frac{1}{a_t}} \frac{e^{a_t + b_t}a_t}{1 + e^{a_t + b_t}} > \frac{1 + \frac{1}{a_t}}{1 + r + \frac{1}{a_t}}. \quad (A.9)$$

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

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

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

\[
\begin{align*}
\frac{e^{a_{t+1} + b_{t+1}} a_{t+1}}{1 - e^{2(a_{t+1} + b_{t+1})}} &< \frac{e^{a_{t+1} + b_{t+1}} a_{t+1}}{1 - e^{2(a_t + b_t)}} \quad \because a_{t+1} + b_{t+1} < a_t + b_t \\
&= \frac{e^{a_t + b_t} a_t}{1 - e^{2(a_t + b_t)}} e^{\eta G_t(a_t, a_t + b_t)} (1 + \eta \frac{G_t(a_t, a_t + b_t)}{a_t}) \\
&\leq \frac{e^{a_t + b_t} a_t}{1 - e^{2(a_t + b_t)}} e^{\eta (1 + \frac{1}{r}) G_t(\frac{a_t + a_t + b_t}{a_t})} \\
&< \frac{e^{a_t + b_t} a_t}{1 - e^{2(a_t + b_t)}} < 1 \quad \because \frac{G_t(\frac{a_t + a_t + b_t}{a_t})}{G_t(x, z)} < 0.
\end{align*}
\]

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

\[
G(\frac{a_t + a_t + b_t}{a_t}) > 0
\]

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

**Lower bound recursion**  
Define

\[
r_t := \frac{1 + \frac{1}{a_t}}{1 + r + \frac{1}{a_t}}.
\]

Define $\bar{r}_t := \sigma(a_t + b_t) a_t$, we are going to establish $\bar{r}_{t+1} > r_{t+1}$. Now let’s establish the recursion for the lower bound.

\[
\begin{align*}
\frac{e^{a_{t+1} + b_{t+1}} a_{t+1}}{1 + e^{a_{t+1} + b_{t+1}}} &< \frac{e^{a_{t+1} + b_{t+1}} a_{t+1}}{1 + e^{a_t + b_t}} \quad \because a_{t+1} + b_{t+1} < a_t + b_t \\
&= \frac{e^{a_t + b_t} a_t}{1 + e^{a_t + b_t}} e^{\eta G_t(a_t, a_t + b_t)} e^{\frac{1}{r} \log(1 + \frac{a_t + a_t}{a_t})} \\
&> \frac{e^{a_t + b_t} a_t}{1 + e^{a_t + b_t}} e^{\eta G_t(a_t, a_t + b_t)} e^{\frac{G_t(a_t + a_t + b_t)}{a_t}} \\
&= \sigma(a_t + b_t) a_t e^{\eta \left[\left(1 + \frac{1}{a_t + 1}\right) \sigma'(a_t + b_t) a_t + \left(1 + \frac{1}{a_t + 1}\right) \left(\frac{a_t}{a_t + b_t} - \sigma(a_t + b_t)\right)\right]}.
\end{align*}
\]

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

\[
\begin{align*}
&= \sigma(a_t + b_t) a_t e^{\eta \left[\left(1 + \frac{1}{a_t + 1}\right) \sigma'(a_t + b_t) a_t + \left(1 + \frac{1}{a_t + 1}\right) \left(\frac{a_t}{a_t + b_t} - \sigma(a_t + b_t)\right)\right]} \\
&= \bar{r}_t e^{\eta \left[\left(1 + \frac{1}{a_t + 1}\right) \sigma'(a_t + b_t) a_t + \left(1 + \frac{1}{a_t + 1}\right) \left(\frac{a_t}{a_t + b_t} - \sigma(a_t + b_t)\right)\right]} e^{-\eta \left(1 + \frac{1}{a_t + 1}\right) (1 - \sigma(a_t))} \\
&= \bar{r}_t e^{\eta \left(1 - \sigma(a_t + b_t)\right) \left(\frac{a_t}{a_t + b_t} - \sigma(a_t + b_t)\right)} e^{-\eta \left(1 + \frac{1}{a_t + 1}\right) (1 - \sigma(a_t))} \\
&> \bar{r}_t \left(1 - \frac{1}{a_t + 1} \left(\frac{1 + \frac{1}{a_t + 1}}{1 + r + \frac{1}{a_t + 1}}\right)\right) e^{-\eta \left(1 + \frac{1}{a_t + 1}\right) (1 - \sigma(a_t))} \quad \because \sigma'(z) = \sigma(z) (1 - \sigma(z)) \\
&\because \epsilon^2 \geq 1 + z.
\end{align*}
\]
We first control the last term on the RHS of the above equation,
\[ e^{-\eta(1 + \frac{1}{\eta a_{t+1}})(1 - \sigma(a_t))} \geq 1 - \eta(1 + \frac{1}{\eta a_{t+1}})(1 - \sigma(a_t)) > 1 - \eta \frac{1 + c^{-1}}{1 + e^{a_t}} > 1 - \eta(1 + c^{-1})e^{-a_t}. \]

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
\[ \tilde{r}_{t+1} > \tilde{r}_t \left[ 1 - \tilde{\eta}(\tilde{r}_t - \tilde{r}_{t+1}) \right] \left[ 1 - \eta(1 + c^{-1})e^{-a_t} \right]. \] (A.11)

For \( a_t > c \) some absolute constant, we have \( e^{-a_t} < r_t - r_{t+1} \), as the RHS \( r_t - r_{t+1} \) is of order \( (a_{t+1} - a_t) a_t^{-2} = \Theta(a_t^{-2}) \) yet the LHS is exponential in \( a_t \). Therefore \( \eta(1 + c^{-1})e^{-a_t} < \tilde{\eta} e^{-a_t} < \tilde{\eta}(r_t - r_{t+1}) < \tilde{\eta}(\tilde{r}_t - r_{t+1}) \), and thus (A.11) reads
\[ \tilde{r}_{t+1} > \tilde{r}_t \left[ 1 - \tilde{\eta}(\tilde{r}_t - \tilde{r}_{t+1}) \right]^2. \]

One can subtract \( r_{t+1} \) on both sides, and see
\[ \tilde{r}_{t+1} - r_{t+1} > \tilde{r}_t \left[ 1 - \tilde{\eta}(\tilde{r}_t - \tilde{r}_{t+1}) \right]^2 - r_{t+1} \geq (\tilde{r}_t - r_{t+1})[1 - 2\tilde{\eta}\tilde{r}_t + \tilde{\eta}^2(\tilde{r}_t - r_{t+1})]. \]

The lower bound on Eqn. (A.10) is true as long as the RHS of the above is positive
\[ (\tilde{r}_t - r_{t+1})[1 - 2\tilde{\eta}\tilde{r}_t + \tilde{\eta}^2(\tilde{r}_t - r_{t+1})] > 0. \] (A.12)

Note \( \tilde{\eta} < 1/2, \tilde{r}_t < 1 \) and \( \tilde{r}_t > r_t > r_{t+1} \), we have concluded
\[ \frac{e^{a_t_{t+1} + b_{t+1}} r_{t+1}}{1 + e^{a_{t+1} + b_{t+1}}} = \tilde{r}_{t+1} > \frac{r_{t+1}}{1 + r + a_{t+1}}. \] (A.13)

\[ \square \]

### A.3 Proofs in Section 2.2

#### Proof of Theorem 3.

\[ \theta^{(1)} = \theta^{(0)} - \eta \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ell(\langle x^i, \theta \rangle, y^i)|_{\theta = \theta^{(0)}}, \]
\[ \theta^{(1)} - \theta_* = \left( I - 2\eta \frac{1}{n} \sum_{i=1}^{n} x^i \otimes x^i \right) (\theta^{(0)} - \theta_*) + 2\eta \frac{1}{n} \sum_{i=1}^{n} e^i x^i. \]

Fix \( n \), first let \( T \to \infty \). Theorem 1 shows that \( x^i_T \to \Delta_b \) (we denote the dependence on \( T \) explicitly) in the following sense
\[ \langle x^i_T, \Delta_b \rangle^2 \geq 1 - O(e^{-cT}). \] (A.14)

Choose \( \eta = 1/2 \), we have
\[ \|\theta^{(1)} - \theta_*\| \leq \| (I - \frac{1}{n} \sum_{i=1}^{n} x^i \otimes x^i)(\theta^{(0)} - \theta_*) \| + \| \frac{1}{n} \sum_{i=1}^{n} e^i x^i \|. \]
\[ := \Gamma \]
For the first term (i), along the direction $\Delta_b$,

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

One the orthogonal complement to $\Delta_b$,

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

Therefore

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

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

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

with probability at least $1 - \delta$. 

**Proof of Theorem 4.**

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

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

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

$$\lim_{T \to \infty} \langle \theta^* - \theta^{(1)}_T, x^i_T \rangle^2 = 1 - \lim_{T \to \infty} \langle \Delta_c, x^i_T \rangle^2 = 0 ,$$

and therefore

$$\lim_{T \to \infty} \left| \langle \theta^* - \theta^{(1)}_T, \theta^* \rangle - \langle \theta^* - \theta^{(0)}_T, \theta^* \rangle \right| \leq \eta \cdot \frac{1}{n} \sum_{i=1}^{n} \lim_{T \to \infty} \left| \sigma \left( \langle x^i_T, \theta^{(0)} \rangle \right) - y^i \right| \cdot \left| \langle \theta^*, x^i_T \rangle \right| ,$$

$$\leq \frac{1}{n} \sum_{i=1}^{n} \lim_{T \to \infty} \left| \langle \theta^*, x^i_T \rangle \right| = 0 .$$

The final claim is a direct fact from

$$\lim_{T \to \infty} \| \theta^* - \theta^{(1)}_T \| \geq \left| \langle \theta^* - \theta^{(1)}_T, \theta^* \rangle \right| = \left| \langle \theta^* - \theta^{(0)}_T, \theta^* \rangle \right| \cdot \left| \langle \theta^*, \theta^* \rangle \right| .$$

Take $\lim \inf \text{w.r.t. to } n$, and we conclude the proof. 

20