Blessings and Curses of Covariate Shifts

Adversarial Learning Dynamics, Directional Convergence, and Equilibria

Tengyuan Liang

The University of Chicago, Booth School of Business

Table of contents

Motivation

Learning vs. domain adaptation

Two environments: source/training and target/testing

X covariate, Y response or label

Learn a **statistical model/hypothesis** $\widehat{f} : X \to Y$ using **data** collected from source/training dom

Deploy $\widehat{f}$ to target/testing dom

Learning vs. domain adaptation

Two environments: source/training and target/testing

X covariate, Y response or label

Learn a **statistical model/hypothesis** $\widehat{f} : X \to Y$ using **data** collected from source/training dom

Deploy $\widehat{f}$ to target/testing dom

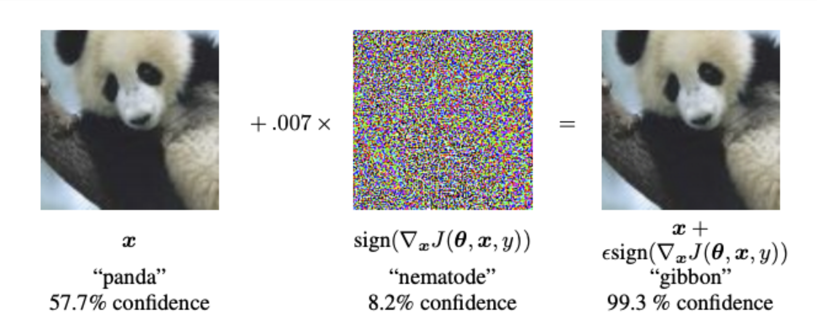
Question

Small discrepancy between source/training and target/testing makes domain adaptation possible?

Adversarial examples

Question

Small discrepancy between source/training and target/testing makes domain adaptation possible?



Goodfellow, Shlens, and Szegedy, 2014

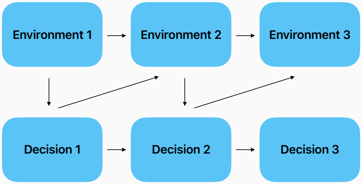
Quest for robustness

How to make the learned model $\widehat{f}$ robust to distributional shift, domain extrapolation or adversarial perturbation?

- robust features/representations (invariance)
- robust learning procedure (adversarial)
- regularization perspective
- loss function perspective
- other notions of robustness? boosting?

Environment shift and learning

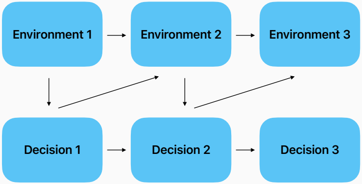
Chicken and egg problem:



- model environment change:
  RL, control, time series analysis
- study equilibria:
  game-theoretic, dynamics

Environment shift and learning

Chicken and egg problem:



- model environment change:
  RL, control, time series analysis
- study equilibria:
  game-theoretic, dynamics

Lucas Critique, 1976

> *"Given that the structure of an econometric model consists of optimal decision rules of economic agents, and that optimal decision rules vary systematically with changes in the structure of series relevant to the decision maker, it follows that any change in policy will systematically alter the structure of econometric models."*

This paper: game-theoretic and dynamic viewpoints

to study covariate shift and adversarial learning

Literature and Background

Learning vs. domain adaptation

Learnability π_{source} Vapnik, 1999

$$\mathcal{R}(\widehat{f}, \pi_{\text{source}}) \leq \widehat{\mathcal{R}}(\widehat{f}, \pi_{\text{source}}) + \text{VC bound}$$
$$\mathcal{R}(\widehat{f}, \pi_{\text{source}}) \leq \inf_{f} \mathcal{R}(f, \pi_{\text{source}}) + \text{excess risk}$$

Learning vs. domain adaptation

Learnability π_{source} Vapnik, 1999

$$\mathcal{R}(\widehat{f}, \pi_{\text{source}}) \leq \widehat{\mathcal{R}}(\widehat{f}, \pi_{\text{source}}) + \text{VC bound}$$

$$\mathcal{R}(\widehat{f}, \pi_{\text{source}}) \leq \inf_{f} \mathcal{R}(f, \pi_{\text{source}}) + \text{excess risk}$$

Dom Adaptation π_{source} vs. π_{target} Ben-David, Blitzer, Crammer, and Pereira, 2006

$$\mathcal{R}(\widehat{f}, \pi_{\text{target}}) \leq \ldots ?$$

Learning vs. domain adaptation

Learnability π_{source} Vapnik, 1999

$$\mathcal{R}(\widehat{f}, \pi_{\text{source}}) \leq \widehat{\mathcal{R}}(\widehat{f}, \pi_{\text{source}}) + \text{VC bound}$$

$$\mathcal{R}(\widehat{f}, \pi_{\text{source}}) \leq \inf_f \mathcal{R}(f, \pi_{\text{source}}) + \text{excess risk}$$

Dom Adaptation π_{source} vs. π_{target} Ben-David, Blitzer, Crammer, and Pereira, 2006

$$\mathcal{R}(\widehat{f}, \pi_{\text{target}}) \leq \widehat{\mathcal{R}}(\widehat{f}, \pi_{\text{source}}) + \text{VC bound} + \boxed{\mathrm{disc}(\pi_{\text{target}}, \pi_{\text{source}})}$$

$$\mathcal{R}(\widehat{f}, \pi_{\text{target}}) \leq \inf_f \{\mathcal{R}(f, \pi_{\text{target}}) + \mathcal{R}(f, \pi_{\text{source}})\} + \text{VC bound} + \boxed{\mathrm{disc}(\pi_{\text{target}}, \pi_{\text{source}})}$$

Learning vs. domain adaptation

Learnability π_{source} Vapnik, 1999

$$\mathcal{R}(\widehat{f}, \pi_{\text{source}}) \leq \widehat{\mathcal{R}}(\widehat{f}, \pi_{\text{source}}) + \text{VC bound}$$

$$\mathcal{R}(\widehat{f}, \pi_{\text{source}}) \leq \inf_f \mathcal{R}(f, \pi_{\text{source}}) + \text{excess risk}$$

Dom Adaptation π_{source} vs. π_{target} Ben-David, Blitzer, Crammer, and Pereira, 2006

$$\mathcal{R}(\widehat{f}, \pi_{\text{target}}) \leq \widehat{\mathcal{R}}(\widehat{f}, \pi_{\text{source}}) + \text{VC bound} + \boxed{\text{disc}(\pi_{\text{target}}, \pi_{\text{source}})}$$

$$\mathcal{R}(\widehat{f}, \pi_{\text{target}}) \leq \inf_f \{\mathcal{R}(f, \pi_{\text{target}}) + \mathcal{R}(f, \pi_{\text{source}})\} + \text{VC bound} + \boxed{\text{disc}(\pi_{\text{target}}, \pi_{\text{source}})}$$

Common **hypothesis** f with small error in both π_{source} and π_{target}

Small **discrepancy between measures** $\text{disc}(\pi_{\text{target}}, \pi_{\text{source}})$

Learning vs. domain adaptation

Common **hypothesis** f with small error in both π_{source} and π_{target}

Small **discrepancy between measures** $\operatorname{disc}(\pi_{\text{target}}, \pi_{\text{source}})$

Hypothesis shift $P_{\text{source}}(Y|X) \neq P_{\text{target}}(Y|X)$

concept (Bayes optimal) is a moving target, hard problem

Ben-David, Lu, et al., 2010

Covariate shift $P_{\text{source}}(X) \neq P_{\text{target}}(X)$

same concept, different evaluation metric, feasible problem

Stone, 1980; Shimodaira, 2000; Sugiyama, Krauledat, and Müller, 2007

Discrepancy and perturbation

What **discrepancy** $\mathrm{disc}(\pi_{\mathrm{target}}, \pi_{\mathrm{source}})$? How to encourage small **discrepancy**, or directly, small target error?

- IPM: $\mathrm{disc}(\pi_{\mathrm{t}}, \pi_{\mathrm{s}}) = \sup_f \left| \int \ell(f, z) \, \mathrm{d}\pi_{\mathrm{t}}(z) - \int \ell(f, z) \, \mathrm{d}\pi_{\mathrm{s}}(z) \right|$
total variation, adversarial metric Ben-David, Blitzer, Crammer, Kulesza, et al., 2010; Mansour, Mohri, and Rostamizadeh, 2009

Discrepancy and perturbation

What **discrepancy** $\mathrm{disc}(\pi_{\mathrm{target}}, \pi_{\mathrm{source}})$? How to encourage small **discrepancy**, or directly, small target error?

- IPM: $\mathrm{disc}(\pi_{\mathrm{t}}, \pi_{\mathrm{s}}) = \sup_f \left| \int \ell(f,z)\,\mathrm{d}\pi_{\mathrm{t}}(z) - \int \ell(f,z)\,\mathrm{d}\pi_{\mathrm{s}}(z) \right|$
  total variation, adversarial metric Ben-David, Blitzer, Crammer, Kulesza, et al., 2010; Mansour, Mohri, and Rostamizadeh, 2009
- KL and likelihood ratio: $\int \ell(f,z)\,\mathrm{d}\pi_{\mathrm{t}}(z) = \int \ell(f,z)\frac{\mathrm{d}\pi_{\mathrm{t}}(z)}{\mathrm{d}\pi_{\mathrm{s}}(z)}\,\mathrm{d}\pi_{\mathrm{s}}(z)$
  reweighted ERM: weight data by likelihood ratio $\frac{\mathrm{d}\pi_{\mathrm{t}}}{\mathrm{d}\pi_{\mathrm{s}}}$ Sugiyama, Krauledat, and Müller, 2007; Sugiyama and Mueller, 2005

Discrepancy and perturbation

What **discrepancy** $\mathrm{disc}(\pi_{\mathrm{target}}, \pi_{\mathrm{source}})$? How to encourage small **discrepancy**, or directly, small target error?

- IPM: $\mathrm{disc}(\pi_{\mathrm{t}}, \pi_{\mathrm{s}}) = \sup_f \left| \int \ell(f, z)\, \mathrm{d}\pi_{\mathrm{t}}(z) - \int \ell(f, z)\, \mathrm{d}\pi_{\mathrm{s}}(z) \right|$
  total variation, adversarial metric Ben-David, Blitzer, Crammer, Kulesza, et al., 2010; Mansour, Mohri, and Rostamizadeh, 2009
- KL and likelihood ratio: $\int \ell(f, z)\, \mathrm{d}\pi_{\mathrm{t}}(z) = \int \ell(f, z) \frac{\mathrm{d}\pi_{\mathrm{t}}(z)}{\mathrm{d}\pi_{\mathrm{s}}(z)}\, \mathrm{d}\pi_{\mathrm{s}}(z)$
  reweighted ERM: weight data by likelihood ratio $\frac{\mathrm{d}\pi_{\mathrm{t}}}{\mathrm{d}\pi_{\mathrm{s}}}$ Sugiyama, Krauledat, and Müller, 2007; Sugiyama and Mueller, 2005
- Regularization: $\ell(f, z) - \ell(f, z')$ is small if z, z' close

Discrepancy and perturbation

What **discrepancy** $\mathrm{disc}(\pi_{\mathrm{target}}, \pi_{\mathrm{source}})$? How to encourage small **discrepancy**, or directly, small target error?

- IPM: $\mathrm{disc}(\pi_{\mathrm{t}}, \pi_{\mathrm{s}}) = \sup_f |\int \ell(f,z)\,\mathrm{d}\pi_{\mathrm{t}}(z) - \int \ell(f,z)\,\mathrm{d}\pi_{\mathrm{s}}(z)|$
  total variation, adversarial metric Ben-David, Blitzer, Crammer, Kulesza, et al., 2010; Mansour, Mohri, and Rostamizadeh, 2009
- KL and likelihood ratio: $\int \ell(f,z)\,\mathrm{d}\pi_{\mathrm{t}}(z) = \int \ell(f,z)\frac{\mathrm{d}\pi_{\mathrm{t}}(z)}{\mathrm{d}\pi_{\mathrm{s}}(z)}\,\mathrm{d}\pi_{\mathrm{s}}(z)$
  reweighted ERM: weight data by likelihood ratio $\frac{\mathrm{d}\pi_{\mathrm{t}}}{\mathrm{d}\pi_{\mathrm{s}}}$ Sugiyama, Krauledat, and Müller, 2007; Sugiyama and Mueller, 2005
- Regularization: $\ell(f,z) - \ell(f,z')$ is small if z, z' close

Closely connected! **optimal transport**/**W**asserstein metric

$$\begin{aligned}
|\mathcal{R}(f,\pi_{\mathrm{t}}) - \mathcal{R}(f,\pi_{\mathrm{s}})| &= \left| \int \ell(f,z)\,\mathrm{d}\pi_{\mathrm{t}}(z) - \int \ell(f,z)\,\mathrm{d}\pi_{\mathrm{s}}(z) \right| \\
&\leq \inf_{\gamma\in\Pi(\pi_{\mathrm{t}},\pi_{\mathrm{s}})} \int \omega_f(\|z-z'\|)\,\mathrm{d}\gamma(z,z') \\
&\leq \mathrm{disc}^{\mathbf{W}}(\pi_{\mathrm{t}},\pi_{\mathrm{s}})
\end{aligned}$$

Adversarial learning and robust optimization

Learn the best model within small **discrepancy** perturbation

- Adversarial learning and examples Goodfellow, Shlens, and Szegedy, 2014; Ilyas et al., 2019; Madry et al., 2017
- Distributionally robust optimization Delage and Ye, 2010

$$\min_{f \in \mathcal{F}} \max_{\pi:\ \mathrm{d}^{\mathsf{W}}(\pi, \pi_{\mathrm{s}}) \leq \gamma} \mathcal{R}(f, \pi)$$

A long list: Bartlett, Bubeck, and Cherapanamjeri, 2021; Bubeck et al., 2021; Javanmard and Soltanolkotabi, 2022; Javanmard, Soltanolkotabi, and Hassani, 2020; Ross and Doshi-Velez, 2018...

Adversarial learning vs. robust optimization

$$\boxed{\min_{f \in \mathcal{F}} \max_{\pi:\ \mathrm{d}^{\mathsf{W}}(\pi, \pi_{\mathrm{s}}) \leq \gamma}} \mathcal{R}(f, \pi)$$

Two views:

- Game-theoretic view: finding equilibria
- Dynamic view: finding adversarial examples

Adversarial learning vs. robust optimization

$$\boxed{\min_{f \in \mathcal{F}} \max_{\pi:\ \mathrm{d}^{\mathsf{W}}(\pi, \pi_{\mathrm{s}}) \leq \gamma}} \mathcal{R}(f, \pi)$$

Two views:

- Game-theoretic view: finding equilibria
  - game between *learner (statistical model, f)* and *nature (data distribution, π)*
  - infinite dimensional game: does minimax theorem hold?
  - what notions of equilibria? Stackelberg, Nash
- Dynamic view: finding adversarial examples

Adversarial learning vs. robust optimization

$$\boxed{\min_{f \in \mathcal{F}} \max_{\pi:\ \mathrm{d}^{\mathsf{W}}(\pi, \pi_{\mathrm{s}}) \leq \gamma}} \mathcal{R}(f, \pi)$$

Two views:

- Game-theoretic view: finding equilibria
- Dynamic view: finding adversarial examples
  - for a given model f, gradient ascent on data finds adversarial examples
  - equiv. Wasserstein gradient flow on π

$$\pi := \arg\min_{\pi} \; -\mathcal{R}(f, \pi) + \frac{1}{\gamma} \mathrm{d}^{\mathsf{W}}(\pi, \pi_{\mathrm{s}})$$

  - hardest extrapolation domain for a given model f?

Game-theoretic and dynamic views: boosting

Boosting and *KL discrepancy* Freund and Schapire, 1997, 1999

Learner *linear* $f_\theta(x) = \langle \theta, x \rangle$, $x \in \mathbb{R}^p$ prediction of weak learners

Nature *finitely supported on fixed points* $\pi_w = \sum_{i=1}^n w_i \delta_{(x_i, y_i)}$, $w \in \Delta_n$ prob. simplex

Dynamics exponentiated gradient step

$$w^{t+1} := \operatorname*{arg\,min}_{w \in \Delta_n} \; -\mathcal{R}(\theta, w) + \frac{1}{\gamma} \mathrm{d}^{\mathbf{KL}}(w, w^t)$$

Equilibria max-min margin solution

$$\theta^\star = \operatorname*{arg\,min}_{\theta:\ \|\theta\| \le 1} \max_{w \in \Delta_n} \mathcal{R}(\theta, w), \quad \text{where } \mathcal{R}(\theta, w) := -\sum_{i=1}^n w_i y_i \langle x_i, \theta \rangle$$

Game-theoretic and dynamic views: boosting

Boosting and *KL discrepancy* Freund and Schapire, 1997, 1999

Learner *linear* $f_\theta(x) = \langle \theta, x \rangle$, $x \in \mathbb{R}^p$ prediction of weak learners

Nature *finitely supported on fixed points* $\pi_w = \sum_{i=1}^n w_i \delta_{(x_i, y_i)}$, $w \in \Delta_n$ prob. simplex

Dynamics exponentiated gradient step

$$w^{t+1} := \underset{w \in \Delta_n}{\arg\min} \; -\mathcal{R}(\theta, w) + \frac{1}{\gamma} \mathrm{d}^{\mathsf{KL}}(w, w^t)$$

Equilibria max-min margin solution

$$\theta^\star = \underset{\theta\colon \|\theta\| \le 1}{\arg\min} \max_{w \in \Delta_n} \mathcal{R}(\theta, w), \quad \text{where } \mathcal{R}(\theta, w) := -\sum_{i=1}^n w_i y_i \langle x_i, \theta \rangle$$

No extrapolation: nature shift weights, same domain/support

Kullback-Leibler as discrepancy measure, exponentiated gradient

Hypothesis shift: no. $P_{\mathrm{source}}(Y|X) \equiv P_{\mathrm{target}}(Y|X)$

Covariate shift: yes. $P_{\mathrm{source}}(X) \neq P_{\mathrm{target}}(X)$

This paper

Extrapolation: nature change domain, move outside support

Wasserstein as discrepancy measure, Wasserstein gradient

This paper

Extrapolation: nature change domain, move outside support

Wasserstein as discrepancy measure, Wasserstein gradient

Hypothesis shift: no. $P_{\mathrm{source}}(Y|X) \equiv P_{\mathrm{target}}(Y|X)$

Covariate shift: yes. $P_{\mathrm{source}}(X) \neq P_{\mathrm{target}}(X)$

This paper

Extrapolation: nature change domain, move outside support

Wasserstein as discrepancy measure, Wasserstein gradient

Hypothesis shift: no. $P_{\text{source}}(Y|X) \equiv P_{\text{target}}(Y|X)$

Covariate shift: yes. $P_{\text{source}}(X) \neq P_{\text{target}}(X)$

Our goal: game-theoretic and dynamic

`Adversarial covariate shifts move the current covariate domain to an extrapolation region. We precisely characterize the region driven by the adversarial dynamics, and subsequent implications to subsequent learning of equilibrium of the game.`

Main Results

Problem setup

- Model class: infinite-dimensional linear model

$$\mathcal{F} := \{f_\theta \mid f_\theta(x) := \langle x, \theta \rangle, \theta \in \ell^2_{\mathbb{N}}\}$$

Problem setup

- Model class: infinite-dimensional linear model

$$\mathcal{F} := \{ f_\theta \mid f_\theta(x) := \langle x, \theta \rangle, \theta \in \ell^2_{\mathbb{N}} \}$$

- Risk or utility:

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

Problem setup

- Model class: infinite-dimensional linear model

$$\mathcal{F} := \{f_\theta \mid f_\theta(x) := \langle x, \theta \rangle, \theta \in \ell^2_{\mathbb{N}}\}$$

- Risk or utility:

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

- Conditional concept $Y|X$: $\pi^\star_x(y)$

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

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

Equilibrum and dynamics

- **Game**: model $\theta \in \ell^2_{\mathbb{N}}$ and covariate distribution $\mu \in \mathcal{P}(X)$, competing for the risk

Equilibrum and dynamics

- **Game**: model $\theta \in \ell^2_{\mathbb{N}}$ and covariate distribution $\mu \in \mathcal{P}(X)$, competing for the risk
- Equilibrium: Bayes optimal model $f^\star_{\text{Bayes}}(x) = \langle x, \theta^\star \rangle$ is a Nash equilibrium of $\mathcal{U}(\cdot, \cdot)$

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

Equilibrum and dynamics

- **Game**: model $\theta \in \ell^2_{\mathbb{N}}$ and covariate distribution $\mu \in \mathcal{P}(X)$, competing for the risk
- Equilibrium: Bayes optimal model $f^\star_{\mathrm{Bayes}}(x) = \langle x, \theta^\star \rangle$ is a Nash equilibrium of $\mathcal{U}(\cdot, \cdot)$

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

- Dynamics: given model $\theta^{(0)}$, the covariate distribution $\mu^{(0)}$ is adversarially perturbed incrementally with Wasserstein as disc.

set stepsize $\gamma \in \mathbb{R}_+$, initialize $\nu_0 := \mu^{(0)}$,

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

and then set $\mu^{(1)} := \nu_{T+1}$

Main contributions

We show two directional convergence results that exhibit distinctive phenomena:

Contributions

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

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 Δ_{b} and the curse direction Δ_{c}

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

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

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 Δ_{b} and the curse direction Δ_{c}

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

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

blessing direction: $\boxed{\Delta_{\mathrm{b}} // \theta^\star - \theta^{(0)}}$

curse direction: $\boxed{\Delta_{\mathrm{c}} \perp \theta^\star}$

Regression

Regression: directional convergence

Theorem (L., 2022)

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 a mild initialization condition, then the adversarial distribution shift dynamic satisfies

$$\lim_{T\to\infty} \left|\left\langle \frac{x_T}{\|x_T\|}, \Delta_{\mathrm{b}} \right\rangle\right| = 1 \text{ , where } \Delta_{\mathrm{b}} // \theta^\star - \theta^{(0)}$$

Moreover, the directional convergence is exponential in T,

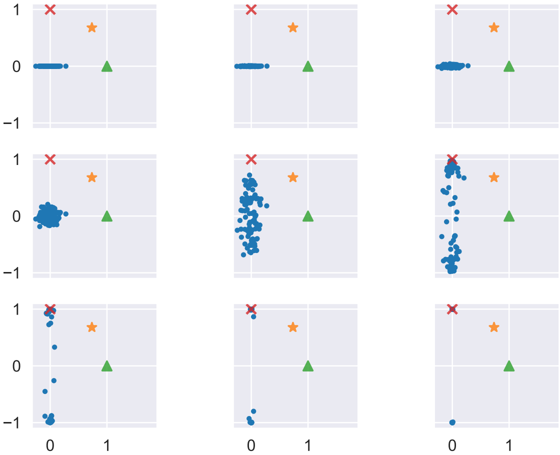
$$\left|\left\langle \frac{x_T}{\|x_T\|}, \Delta_{\mathrm{b}} \right\rangle\right| \in \left[1 - O\left(\frac{1}{e^{cT}}\right) \text{ , } 1\right] \text{ .}$$

Regression: directional convergence

Some remarks

- 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: a one-dimensional "blessing" direction Δ_{b}
- 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 $\theta^{\star}$
- directional alignment is fast, **exponential**!

Regression: numeric study



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.

Regression: impact on subsequent learning

Theorem (L., 2022)

The learner's one-step reaction to the distribution shift with $\eta = 1/2$ satisfies

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

Regression: impact on subsequent learning

Theorem (L., 2022)

The learner's one-step reaction to the distribution shift with $\eta = 1/2$ satisfies

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

For regression, the adversarial distribution shifts make the learner's one-step subsequent move optimal! The above shows that one-step improvement using gradient descent dynamic will reach the Bayes optimal model.

Classification

Classification: directional convergence

Theorem (L., 2022)

Consider the classification setting where
$\ell(y', y) = -y'y + \log\left(1 + e^{y'}\right)$ and $\mathbf{y}|\mathbf{x} = x \sim \mathrm{Bernoulli}\big(\sigma(\langle x, \theta^\star \rangle)\big)$.

Let $x_0 \in \mathrm{supp}(\mu^{(0)})$ that satisfies a mild initialization condition, then the adversarial distribution shift dynamic satisfies

$$\lim_{T\to\infty} \left|\left\langle \tfrac{x_T}{\|x_T\|}, \Delta_{\mathrm{c}} \right\rangle\right| = 1 \ , \ \text{where } \Delta_{\mathrm{c}} \perp \theta^\star$$

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

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

Classification: directional convergence

Theorem (L., 2022)

Consider the classification setting where
$\ell(y', y) = -y'y + \log\left(1 + e^{y'}\right)$ and $\mathbf{y}|\mathbf{x} = x \sim \mathrm{Bernoulli}\big(\sigma(\langle x, \theta^\star\rangle)\big)$.

Let $x_0 \in \mathrm{supp}(\mu^{(0)})$ that satisfies a mild initialization condition, then the adversarial distribution shift dynamic satisfies

$$\lim_{T\to\infty} \left|\left\langle \frac{x_T}{\|x_T\|}, \Delta_{\mathrm{c}} \right\rangle\right| = 1 \text{ , where } \Delta_{\mathrm{c}} \perp \theta^\star$$

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

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

Proof non-trivial: a non-convex non-linear system for covariate shift dynamics.

Classification: directional convergence

Some remarks

- 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.
- adversarial shift (under the logistic loss) asymptotically constructs the **hardest covariate design** under the $0-1$ loss, for the next stage of learning. Namely, the $(x, y) \sim \pi_{\mu^{(1)}}$ where y is a Bernoulli coin-flip that is independent of x, **impossible to predict**!
- directional alignment is slower, **sub-quadratic**!

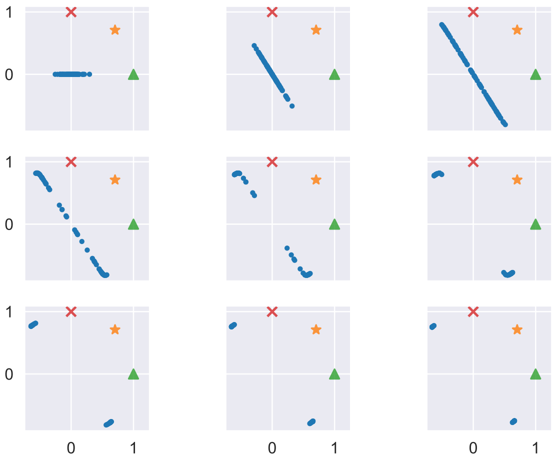
Classification: directional convergence

Some remarks

- 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.
- adversarial shift (under the logistic loss) asymptotically constructs the **hardest covariate design** under the $0-1$ loss, for the next stage of learning. Namely, the $(x, y) \sim \pi_{\mu^{(1)}}$ where y is a Bernoulli coin-flip that is independent of x, **impossible to predict**!
- directional alignment is slower, **sub-quadratic**!

Contrasts sharply with the phenomenon in the regression setting.

Classification: numeric study



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

Classification: impact on subsequent learning

Theorem (L., 2022)

The learner's one-step reaction to the distribution shift with any fixed $\eta > 0$ satisfies

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

Moreover,

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

Classification: impact on subsequent learning

Theorem (L., 2022)

The learner's one-step reaction to the distribution shift with any fixed $\eta > 0$ satisfies

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

Moreover,

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

For classification, the above 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.

Future Directions and Discussions

- Connections: adversarial learning and (automated) experiment design
    - regression: optimal design
    - classification: hardest design
- Tradeoffs: myopic learning vs. eventual learning
    - adversarial perturbation makes the current model suffer
    - yet, it may be beneficial to subsequent learning
- Lucas Critique:

  > *"Given that the structure of an econometric model consists of optimal decision rules of economic agents, and that optimal decision rules vary systematically with changes in the structure of series relevant to the decision maker, it follows that any change in policy will systematically alter the structure of econometric models."*

- Finite-sample bounds and nonlinear models
- Insights to understand GANs, domain adaptation, and more...

Thank you.

Liang, Tengyuan (2022). "Blessings and Curses of Covariate Shifts: Adversarial Learning Dynamics, Directional Convergence, and Equilibria". *In: arXiv preprint arXiv:2212.02457.*

Intuition of the Proof

- Regression: PCA-based analysis
- Classification: novel proof technique
  - dynamic of the distribution shift is non-convex and non-linear
  - two summary statistics a_t, b_t to keep track of the directional convergence
  - rough intuition: after a finite time t_0, a key quantity (L for Lyapunov)

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

    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

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

  - still hard to operate with recursions, define two envelope functions to confine the flow

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

Intuition of the Proof

- $L_t^{\mathrm{env-U}} < 1 \implies L_t < 1,$ and $L_t^{\mathrm{env-L}} > \frac{1}{1+r} \implies L_t > \frac{1}{1+r}$
- if the lower envelope function $L_t^{\mathrm{env-L}} > \frac{1+a_t^{-1}}{1+r+a_t^{-1}} \in [\frac{1}{1+r}, 1]$, then the upper envelope function decreases in the recursion, $L_{t+1}^{\mathrm{env-U}} < L_t^{\mathrm{env-U}} < 1$
- the lower envelope function cannot decrease too much

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

References i

- Bartlett, Peter, Sébastien Bubeck, and Yeshwanth Cherapanamjeri (2021). "Adversarial examples in multi-layer random relu networks". In: *Advances in Neural Information Processing Systems* 34, pp. 9241–9252.
- Ben-David, Shai, John Blitzer, Koby Crammer, Alex Kulesza, et al. (2010). "A theory of learning from different domains". In: *Machine learning* 79, pp. 151–175.
- Ben-David, Shai, John Blitzer, Koby Crammer, and Fernando Pereira (2006). "Analysis of representations for domain adaptation". In: *Advances in neural information processing systems* 19.
- Ben-David, Shai, Tyler Lu, et al. (2010). "Impossibility theorems for domain adaptation". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, pp. 129–136.

- Bubeck, Sébastien et al. (2021). "A single gradient step finds adversarial examples on random two-layers neural networks". In: *Advances in Neural Information Processing Systems* 34, pp. 10081–10091.
- Delage, Erick and Yinyu Ye (2010). "Distributionally robust optimization under moment uncertainty with application to data-driven problems". In: *Operations research* 58.3, pp. 595–612.
- Freund, Yoav and Robert E Schapire (1997). "A decision-theoretic generalization of on-line learning and an application to boosting". In: *Journal of computer and system sciences* 55.1, pp. 119–139.
- — (1999). "Adaptive game playing using multiplicative weights". In: *Games and Economic Behavior* 29.1-2, pp. 79–103.
- Goodfellow, Ian J, Jonathon Shlens, and Christian Szegedy (2014). "Explaining and harnessing adversarial examples". In: *arXiv preprint arXiv:1412.6572*.

References iii

Ilyas, Andrew et al. (2019). “Adversarial examples are not bugs, they are features”. In: *Advances in neural information processing systems* 32.

Javanmard, Adel and Mahdi Soltanolkotabi (Aug. 2022). “Precise statistical analysis of classification accuracies for adversarial training”. In: *The Annals of Statistics* 50.4, pp. 2127–2156. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/22-AOS2180. (Visited on 09/23/2022).

Javanmard, Adel, Mahdi Soltanolkotabi, and Hamed Hassani (2020). “Precise tradeoffs in adversarial training for linear regression”. In: *Conference on Learning Theory*. PMLR, pp. 2034–2078.

Madry, Aleksander et al. (2017). “Towards deep learning models resistant to adversarial attacks”. In: *arXiv preprint arXiv:1706.06083*.

Mansour, Yishay, Mehryar Mohri, and Afshin Rostamizadeh (2009). “Domain adaptation: Learning bounds and algorithms”. In: *arXiv preprint arXiv:0902.3430*.

References iv

- Ross, Andrew and Finale Doshi-Velez (2018). “Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1.
- Shimodaira, Hidetoshi (2000). “Improving predictive inference under covariate shift by weighting the log-likelihood function”. In: *Journal of statistical planning and inference* 90.2, pp. 227–244.
- Stone, Charles J (1980). “Optimal rates of convergence for nonparametric estimators”. In: *The Annals of Statistics*, pp. 1348–1360.
- Sugiyama, Masashi, Matthias Krauledat, and Klaus-Robert Müller (2007). “Covariate shift adaptation by importance weighted cross validation.”. In: *Journal of Machine Learning Research* 8.5.

References v

Sugiyama, Masashi and K Mueller (2005). "Generalization error estimation under covariate shift". In: *Workshop on Information-Based Induction Sciences*. Citeseer, pp. 21–26.

Vapnik, Vladimir (1999). *The nature of statistical learning theory*. Springer science & business media.