Performative Prediction

Juan C. Perdomo* Tijana Zrnic* Celestine Mendler-Dünner Moritz Hardt
{jcperdomo, tijana.zrnic, mendler, hardt}@berkeley.edu

Motivation

Distribution Shift

Fairness Is Not Static:
Deeper Understanding of Long Term Fairness
via Simulation Studies

Alexander D'Amour*
Google Research
alexdamour@google.com

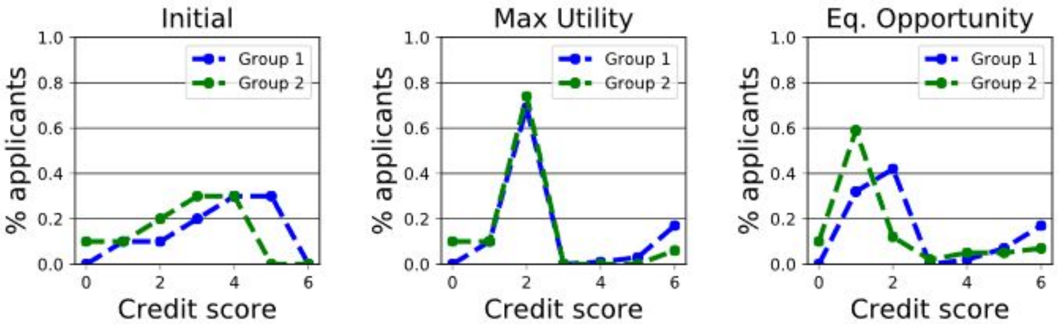
Hansa Srinivasan*
Google Research
hansas@google.com

James Atwood
Google Research
atwoodj@google.com

Pallavi Baljekar
Google Research
pbaljeka@google.com

D. Sculley
Google Research
dsculley@google.com

Yoni Halpern
Google Research
yhalpern@google.com



Retraining

1. Train model
2. Observe distribution shift
3. Collect new data
4. Go back to step 1

What can we say theoretically?

Framework

Notation

$$\theta \qquad \mathcal{D}(\theta)$$

$$Z = (X, Y) \sim \mathcal{D}(\theta)$$

$$\ell(Z; \theta)$$

Risk vs .Performative Risk

$$R(\theta) := \mathbb{E}_{Z \sim \mathcal{D}}[\ell(Z;\theta)]$$

$$PR(\theta) := \mathbb{E}_{Z \sim \mathcal{D}(\theta)}[\ell(Z;\theta)]$$

Optimality

Definition 2.1 (performative optimality and risk). A model $f_{\theta_{\mathrm{PO}}}$ is *performatively optimal* if the following relationship holds:

$$\theta_{\mathrm{PO}} = \arg\min_{\theta} \mathop{\mathbb{E}}_{Z \sim \mathcal{D}(\theta)} \ell(Z;\theta).$$

Example 2.2 (biased coin flip)

$$X \in \{-1, 1\}$$

$$\epsilon < 0.5 - \mu \quad \mu \in (0, 0.5)$$

$$Y \mid X \sim \mathrm{Bern}(0.5 + \mu X + \epsilon\theta X)$$

$$f_\theta(x) := \theta x + 0.5 \qquad \theta \in [0, 1]$$

$$\ell(z; \theta) := (y - f_\theta(x))^2$$

Example 2.2 (biased coin flip)

$$Y \mid X \sim \mathrm{Bern}(0.5 + \mu X + \epsilon\theta X) \qquad f_\theta(x) := \theta x + 0.5$$

$$\mathbb{E}_{Z\sim\mathcal{D}(\theta)}[\ell(Z;\theta)] = \mathbb{E}_X \mathbb{E}_{Y|X}[(y - f_\theta(x))^2 \mid X]$$

$$\mathbb{E}_{Y|X}[(y - f_\theta(x))^2 \mid X] = X^2(\theta^2 - 2\theta\mu - 2\theta^2\epsilon) + 0.25$$

$$\frac{\partial}{\partial\theta}(\ldots) = 2X^2(\theta(1-2\epsilon) - \mu)$$

$$\theta_{PO} = \frac{\mu}{1-2\epsilon}$$

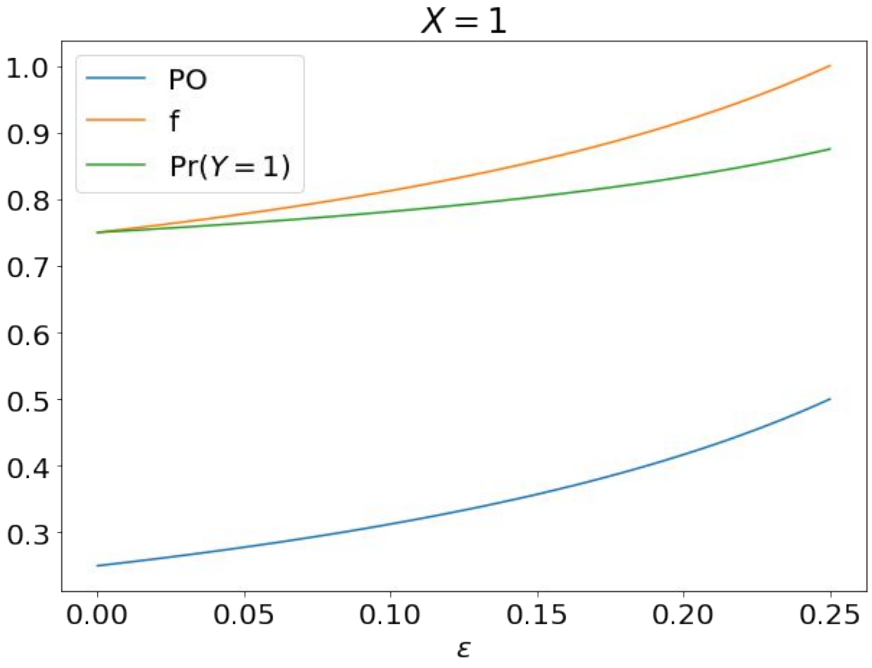
Example 2.2 (biased coin flip)

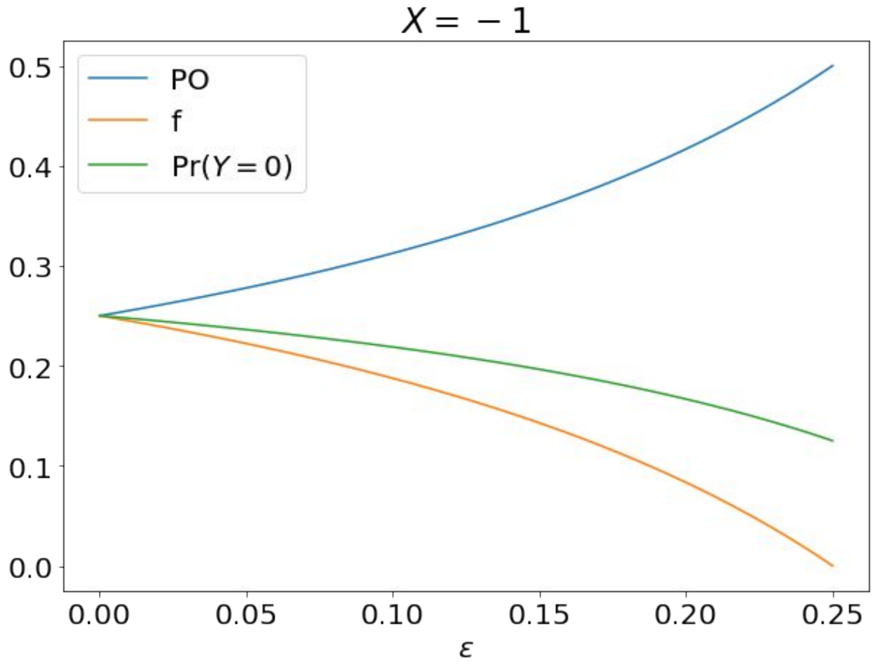
$$\epsilon = 0 \implies \theta_{PO} = \mu$$

$$\implies f_{\theta_{PO}}(x) = \mu x + 0.5 = \mathbb{E}[Y \mid X = x]$$

Example 2.2 (biased coin flip)

$$Y \mid X \sim \mathrm{Bern}(0.5 + \mu X + \epsilon\theta X) \qquad \theta_{PO} = \frac{\mu}{1 - 2\epsilon}$$





Can we actually find optimal points?

Problem!

$$PR(\theta) := \mathbb{E}_{Z \sim \mathcal{D}(\theta)}[\ell(Z;\theta)]$$

$$\theta_{t+1} := \arg\min_{\theta} \mathbb{E}_{Z \sim \mathcal{D}(\theta_t)}[\ell(Z;\theta)]$$

$$G(\theta) := \arg\min_{\theta'} \mathbb{E}_{Z \sim \mathcal{D}(\theta)}[\ell(Z;\theta')]$$

Decoupling risk

$$DPR(\theta, \theta') := \mathbb{E}_{Z \sim \mathcal{D}(\theta)}[\ell(Z; \theta')]$$

Stability

Definition 2.3 (performative stability and decoupled risk). A model $f_{\theta_{\mathrm{PS}}}$ is *performatively stable* if the following relationship holds:

$$\theta_{\mathrm{PS}} = \arg\min_{\theta} \mathop{\mathbb{E}}_{Z \sim \mathcal{D}(\theta_{\mathrm{PS}})} \ell(Z;\theta).$$

$$\theta_{PS} = \arg\min_{\theta} DPR(\theta_{PS}, \theta)$$

Example 2.2 (continued)

$$\theta_{PS} = \arg\min_{\theta} \mathbb{E}_{Z \sim \mathcal{D}(\theta_{PS})}[\ell(Z;\theta)]$$

$$\mathbb{E}_{Z \sim \mathcal{D}(\theta)}[\ell(Z;\theta)] = \mathbb{E}_X \mathbb{E}_{Y|X}[(y - f_\theta(x))^2 \mid X]$$

$$\mathbb{E}_{Y|X}[(y - f_\theta(x))^2 \mid X] = X^2(-2\theta_{PS}\theta\epsilon + \theta^2 - 2\theta\mu) + 0.25$$

$$\frac{\partial}{\partial\theta}(\ldots) = X^2(-2\theta_{PS}\epsilon + 2\theta - 2\mu)$$

$$\arg\min_{\theta} \mathbb{E}_{Z \sim \mathcal{D}(\theta_{PS})}[\ell(Z;\theta)] = \mu + \theta_{PS}\epsilon$$

Example 2.2 (continued)

$$\arg\min_\theta \mathbb{E}_{Z\sim\mathcal{D}(\theta_{PS})}[\ell(Z;\theta)] = \mu + \theta_{PS}\epsilon$$

$$\theta_{PS} = \mu + \theta_{PS}\epsilon$$

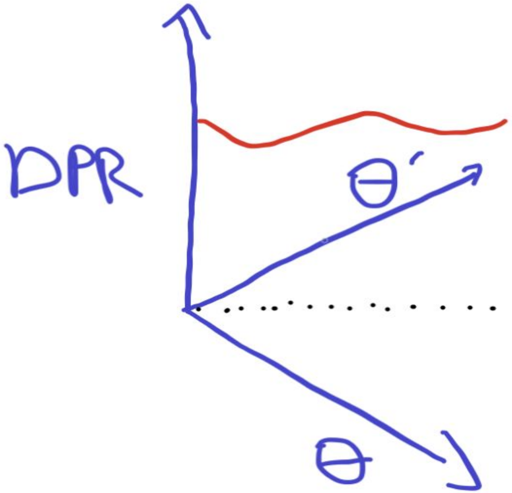
$$\theta_{PS} = \frac{\mu}{1-\epsilon}$$

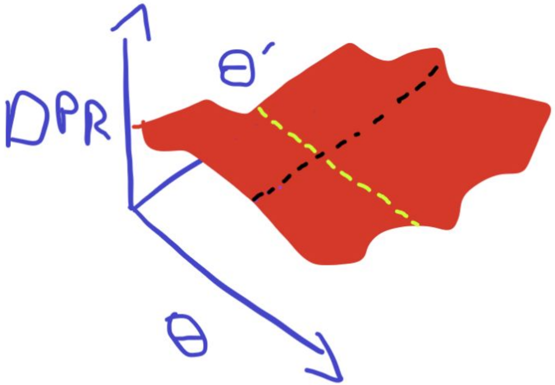
Example 2.2 (continued)

$$\theta_{PS} = \frac{\mu}{1-\epsilon} \qquad \theta_{PO} = \frac{\mu}{1-2\epsilon}$$

Stability vs. Optimality

$$DPR(\theta, \theta') := \mathbb{E}_{Z \sim \mathcal{D}(\theta)}[\ell(Z; \theta')]$$





Stability vs. Optimality

Theorem 4.3. *Suppose that the loss $\ell(z;\theta)$ is L_z-Lipschitz in z, γ-strongly convex (A2), and that the distribution map $\mathcal{D}(\cdot)$ is ε-sensitive. Then, for every performatively stable point θ_{PS} and every performative optimum θ_{PO}:*

$$\|\theta_{\mathrm{PO}} - \theta_{\mathrm{PS}}\|_2 \leqslant \frac{2L_z\varepsilon}{\gamma}.$$

Definition 3.1 (ε-sensitivity)**.** We say that a distribution map $\mathcal{D}(\cdot)$ is *ε-sensitive* if for all $\theta, \theta' \in \Theta$:

$$W_1\Big(\mathcal{D}(\theta), \mathcal{D}(\theta')\Big) \leqslant \varepsilon\|\theta - \theta'\|_2,$$

where W_1 denotes the Wasserstein-1 distance, or earth mover's distance.

Theoretical Results

Assumptions

(*joint smoothness*) We say that a loss function $\ell(z;\theta)$ is β-jointly smooth if the gradient $\nabla_\theta \ell(z;\theta)$ is β-Lipschitz in θ *and* z, that is

$$\left\|\nabla_\theta \ell(z;\theta) - \nabla_\theta \ell(z;\theta')\right\|_2 \leqslant \beta \left\|\theta - \theta'\right\|_2, \quad \left\|\nabla_\theta \ell(z;\theta) - \nabla_\theta \ell(z';\theta)\right\|_2 \leqslant \beta \left\|z - z'\right\|_2, \tag{A1}$$

for all $\theta, \theta' \in \Theta$ and $z, z' \in \mathcal{Z}$.

(*strong convexity*) We say that a loss function $\ell(z;\theta)$ is γ-strongly convex if

$$\ell(z;\theta) \geqslant \ell(z;\theta') + \nabla_\theta \ell(z;\theta')^\top (\theta - \theta') + \frac{\gamma}{2} \left\|\theta - \theta'\right\|_2^2, \tag{A2}$$

for all $\theta, \theta' \in \Theta$ and $z \in \mathcal{Z}$. If $\gamma = 0$, this assumption is equivalent to convexity.

Assumptions

Definition 3.1 (ε-sensitivity)**.** We say that a distribution map $\mathcal{D}(\cdot)$ is ε-*sensitive* if for all $\theta, \theta' \in \Theta$:

$$W_1\big(\mathcal{D}(\theta), \mathcal{D}(\theta')\big) \leqslant \varepsilon \|\theta - \theta'\|_2,$$

where W_1 denotes the Wasserstein-1 distance, or earth mover's distance.

Convergence to a stable point through RRM

$$G(\theta) := \arg\min_{\theta'} \mathbb{E}_{Z\sim\mathcal{D}(\theta)}[\ell(Z;\theta')]$$

Theorem 3.5. *Suppose that the loss $\ell(z;\theta)$ is β-jointly smooth* (A1) *and γ-strongly convex* (A2). *If the distribution map $\mathcal{D}(\cdot)$ is ε-sensitive, then the following statements are true:*

(a) $\|G(\theta) - G(\theta')\|_2 \leqslant \varepsilon\frac{\beta}{\gamma}\|\theta - \theta'\|_2$, *for all* $\theta, \theta' \in \Theta$.

(b) If $\varepsilon < \frac{\gamma}{\beta}$, *the iterates θ_t of RRM converge to a unique performatively stable point θ_{PS} at a linear rate:* $\|\theta_t - \theta_{\mathrm{PS}}\|_2 \leqslant \delta$ *for* $t \geqslant \left(1 - \varepsilon\frac{\beta}{\gamma}\right)^{-1}\log\left(\frac{\|\theta_0-\theta_{\mathrm{PS}}\|_2}{\delta}\right)$.

Proof idea

1. Part (b) follows from (a) by the Banach fixed-point theorem
2. Focus on showing that G is a contraction mapping
    a. Strong convexity upper bounds squared G-distance
    b. Sensitivity and smoothness lower bound G- and param-distance
    c. Combine resulting inequalities

Do we need these assumptions?

Proposition 3.6. *Suppose that the distribution map $\mathcal{D}(\cdot)$ is ε-sensitive with $\varepsilon > 0$. RRM can fail to converge at all in any of the following cases, for any choice of parameters $\beta, \gamma > 0$:*

(a) *The loss is β-jointly smooth and convex, but not strongly convex.*

(b) *The loss is γ-strongly convex, but not jointly smooth.*

(c) *The loss is β-jointly smooth and γ-strongly convex, but $\varepsilon \geqslant \frac{\gamma}{\beta}$.*

Other interesting results

Theorem 3.8. *Suppose that the loss $\ell(z;\theta)$ is β-jointly smooth (A1) and γ-strongly convex (A2). If the distribution map $\mathcal{D}(\cdot)$ is ε-sensitive with $\varepsilon < \frac{\gamma}{(\beta+\gamma)(1+1.5\eta\beta)}$, then RGD with step size $\eta \leqslant \frac{2}{\beta+\gamma}$ satisfies the following:*

(a) $\|G_{gd}(\theta) - G_{gd}(\theta')\|_2 \leqslant \big(1 - \eta$

(b) *The iterates θ_t of RGD con*
$\|\theta_t - \theta_{\mathrm{PS}}\|_2 \leqslant \delta$ *for* $t \geqslant \frac{1}{\eta}\big(\frac{\beta\gamma}{\beta+}$

Theorem 3.10. *Suppose that the loss $\ell(z;\theta)$ is β-jointly smooth (A1) and γ-strongly convex (A2), and that there exist $\alpha > 1, \mu > 0$ such that $\xi_{\alpha,\mu} \overset{\text{def}}{=} \int_{\mathbb{R}^m} e^{\mu|x|^\alpha} d\mathcal{D}(\theta)$ is finite $\forall \theta \in \Theta$. Let $\delta \in (0,1)$ be a radius of convergence. Consider running RERM or RGD with $n_t = O\left(\frac{1}{(\varepsilon\delta)^m}\log\left(\frac{t}{p}\right)\right)$ samples at time t.*

(a) *If $\mathcal{D}(\cdot)$ is ε-sensitive with $\varepsilon < \frac{\gamma}{2\beta}$, then with probability $1-p$, RERM satisfies,*

$$\|\theta_t - \theta_{\mathrm{PS}}\|_2 \leqslant \delta, \text{ for all } t \geqslant \frac{\log\left(\frac{1}{\delta}\|\theta_0 - \theta_{\mathrm{PS}}\|_2\right)}{\left(1 - \frac{2\varepsilon\beta}{\gamma}\right)}.$$

(b) *If $\mathcal{D}(\cdot)$ is ε-sensitive with $\varepsilon < \frac{\gamma}{(\beta+\gamma)(1+1.5\eta\beta)}$, then with probability $1-p$, REGD with satisfies,*

$$\|\theta_t - \theta_{\mathrm{PS}}\|_2 \leqslant \delta, \text{ for all } t \geqslant \frac{\log\left(\frac{1}{\delta}\|\theta_0 - \theta_{\mathrm{PS}}\|_2\right)}{\eta\left(\frac{\beta\gamma}{\beta+\gamma} - \varepsilon(3\eta\beta^2 + 2\beta)\right)},$$

for a constant choice of step size $\eta \leqslant \frac{2}{\beta+\gamma}$.

Remaining Issues

SGD analysis?

Stochastic Optimization for Performative Prediction

Celestine Mendler-Dünner* Juan C. Perdomo* Tijana Zrnic* Moritz Hardt†
{mendler, jcperdomo, tijana.zrnic, hardt}@berkeley.edu

University of California, Berkeley

Types of Distribution Shift

$$P_\theta(X) \neq P_{\theta'}(X) \qquad P_\theta(Y \mid X) = P_{\theta'}(Y \mid X)$$

$$P_\theta(Y \mid X) \neq P_{\theta'}(Y \mid X)$$

$$P_\theta(Y \mid \Phi(X)) = P_{\theta'}(Y \mid \Phi(X))$$

$$\mathcal{D}_t(\theta) \neq \mathcal{D}_{t+1}(\theta)$$

Stability under different learning algorithms

Invariant Risk Minimization

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, David Lopez-Paz

$$\begin{aligned} \min_{\substack{\Phi:\mathcal{X}\to\mathcal{H} \\ w:\mathcal{H}\to\mathcal{Y}}} \quad & \sum_{e\in\mathcal{E}_{\mathrm{tr}}} R^e(w\circ\Phi) \\ \text{subject to} \quad & w \in \operatorname*{arg\,min}_{\bar{w}:\mathcal{H}\to\mathcal{Y}} R^e(\bar{w}\circ\Phi), \ \text{for all } e\in\mathcal{E}_{\mathrm{tr}}. \end{aligned}$$

Questions?