Performative Prediction

Juan C. Perdomo* Tijana Zrnic* Celestine Mendler-Dünner Moritz Hardt
{jperdomo, 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
pbaljekas@google.com

D. Sculley
Google Research
dsculley@google.com

Yoni Halpern
Google Research
yhalpern@google.com

---

**Initial**

![Initial Distribution](image1.png)

**Max Utility**

![Max Utility Distribution](image2.png)

**Eq. Opportunity**

![Eq. Opportunity Distribution](image3.png)
Retraining

1. Train model
2. Observe distribution shift
3. Collect new data
4. Go back to step 1
What can we say theoretically?
Framework
\[ \theta \quad \mathcal{D}(\theta) \]

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

\[ \ell(Z; \theta) \]
Risk vs. Performative Risk

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

\[ PR(\theta) := \mathbb{E}_{Z \sim \mathcal{D}(\theta)}[\ell(Z; \theta)] \]
Definition 2.1 (performative optimality and risk). A model $f_{\theta_{PO}}$ is performatively optimal if the following relationship holds:

$$\theta_{PO} = \arg\min_{\theta} \mathbb{E}_{Z \sim D(\theta)} \ell(Z; \theta).$$
Example 2.2 (biased coin flip)

\[ X \in \{-1, 1\} \]
\[ \epsilon < 0.5 - \mu \quad \mu \in (0, 0.5) \]

\[ Y \mid X \sim \text{Bern}(0.5 + \mu X + \epsilon \theta X) \]

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

\[ \ell(z; \theta) := (y - f_\theta(x))^2 \]
Example 2.2 (biased coin flip)

\[ Y \mid X \sim \text{Bern}(0.5 + \mu X + \epsilon \theta X) \quad f_\theta(x) := \theta x + 0.5 \]

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

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

\[ \frac{\partial}{\partial \theta}(...) = 2X^2(\theta(1 - 2\epsilon) - \mu) \]

\[ \theta_{PO} = \frac{\mu}{1 - 2\epsilon} \]
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)

\[ Y \mid X \sim \text{Bern}(0.5 + \mu X + \epsilon \theta X) \]

\[ \theta_{PO} = \frac{\mu}{1 - 2\epsilon} \]
Can we actually find optimal points?
Problem!

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

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

\[ G(\theta) := \arg \min_{\theta', \theta} \mathbb{E}_{Z \sim D(\theta)}[\ell(Z; \theta')] \]
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 $f_{\theta_{PS}}$ is *performatively stable* if the following relationship holds:

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

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

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

\[ \theta_{PS} = \mu + \theta_{PS} \epsilon \]

\[ \theta_{PS} = \frac{\mu}{1 - \epsilon} \]
Example 2.2 (continued)

\[ \theta_{PS} = \frac{\mu}{1 - \epsilon} \quad \theta_{PO} = \frac{\mu}{1 - 2\epsilon} \]
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 $z$, $\gamma$-strongly convex (A2), and that the distribution map $\mathcal{D}(\cdot)$ is $\varepsilon$-sensitive. Then, for every performatively stable point $\theta_{PS}$ and every performative optimum $\theta_{PO}$:

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

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

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

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 $\beta$-jointly smooth if the gradient $\nabla_{\theta} \ell(z; \theta)$ is $\beta$-Lipschitz in $\theta$ and $z$, that is

$$
\|\nabla_{\theta} \ell(z; \theta) - \nabla_{\theta} \ell(z'; \theta)\|_2 \leq \beta \|\theta - \theta'\|_2, \quad \|\nabla_{\theta} \ell(z; \theta) - \nabla_{\theta} \ell(z'; \theta)\|_2 \leq \beta \|z - z'\|_2, \quad (A1)
$$

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

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

$$
\ell(z; \theta) \geq \ell(z; \theta') + \nabla_{\theta} \ell(z; \theta')^T (\theta - \theta') + \frac{\gamma}{2} \|\theta - \theta'\|_2^2, \quad (A2)
$$

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

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

$$W_1\left(\mathcal{D}(\theta), \mathcal{D}(\theta')\right) \leq \varepsilon \|\theta - \theta'\|_2,$$

where $W_1$ denotes the Wasserstein-1 distance, or earth mover’s distance.
Convergence to a stable point through RRM

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

**Theorem 3.5.** Suppose that the loss \( \ell(z; \theta) \) is \( \beta \)-jointly smooth (A1) and \( \gamma \)-strongly convex (A2). If the distribution map \( D(\cdot) \) is \( \varepsilon \)-sensitive, then the following statements are true:

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

(b) If \( \varepsilon < \frac{\gamma}{\beta} \), the iterates \( \theta_t \) of RRM converge to a unique performatively stable point \( \theta_{PS} \) at a linear rate: \( \|\theta_t - \theta_{PS}\|_2 \leq \delta \) for \( t \geq \left( 1 - \varepsilon \frac{\beta}{\gamma} \right)^{-1} \log \left( \frac{\|\theta_0 - \theta_{PS}\|_2}{\delta} \right) \).
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 $D(\cdot)$ is $\epsilon$-sensitive with $\epsilon > 0$. 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 $\beta$-jointly smooth and convex, but not strongly convex.

(b) The loss is $\gamma$-strongly convex, but not jointly smooth.

(c) The loss is $\beta$-jointly smooth and $\gamma$-strongly convex, but $\epsilon \geq \frac{\gamma}{\beta}$. 
Other interesting results

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

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

(b) The iterates $\theta_t$ of RGD converge and that there exist $\alpha > 1, \mu > 0$ such that $\xi_{\alpha, \mu} \overset{\text{def}}{=} \int_{\mathbb{R}^n} e^{\|x\|^\alpha} dD(\theta)$ 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 $D(\cdot)$ is $\varepsilon$-sensitive with $\varepsilon < \frac{\gamma}{2\beta}$, then with probability $1 - p$, RERM satisfies,

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

(b) If $D(\cdot)$ is $\varepsilon$-sensitive with $\varepsilon < \frac{\gamma}{(\beta + \gamma)(1 + 1.5\eta \beta)}$, then with probability $1 - p$, REGD with step size $\eta \leq \frac{2}{\beta + \gamma}$ satisfies,

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

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

\[ P_\theta(X) \neq P_{\theta'}(X) \quad P_\theta(Y \mid X) = P_{\theta'}(Y \mid X) \]

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

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

\[ \mathcal{D}_t(\theta) \neq \mathcal{D}_{t+1}(\theta) \]
Stability under different learning algorithms

Invariant Risk Minimization

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

$$\min_{\Phi : \mathcal{X} \to \mathcal{H}} \sum_{e \in \mathcal{E}_{tr}} R^e(w \circ \Phi)$$

subject to $$w \in \arg \min_{\bar{w} : \mathcal{H} \to \mathcal{Y}} R^e(\bar{w} \circ \Phi), \text{ for all } e \in \mathcal{E}_{tr}.$$
Questions?