Performative Prediction

Andrew Braun Joonas Soininen

University of Victoria

March 11, 2025

Motivation

The Standard Assumption

In supervised learning, the data distribution $\mathcal{D}$ is **fixed** and independent of the deployed model.

Credit Default Risk: A Self-Fulfilling Prophecy

1. Bank trains model f_θ, predicts applicant has **high default risk**
2. Bank acts on prediction $\Rightarrow$ assigns a **high interest rate**
3. High interest rate **increases** the applicant's actual default risk

$\Rightarrow$ The model's prediction *caused* the outcome it aimed to predict

Motivation

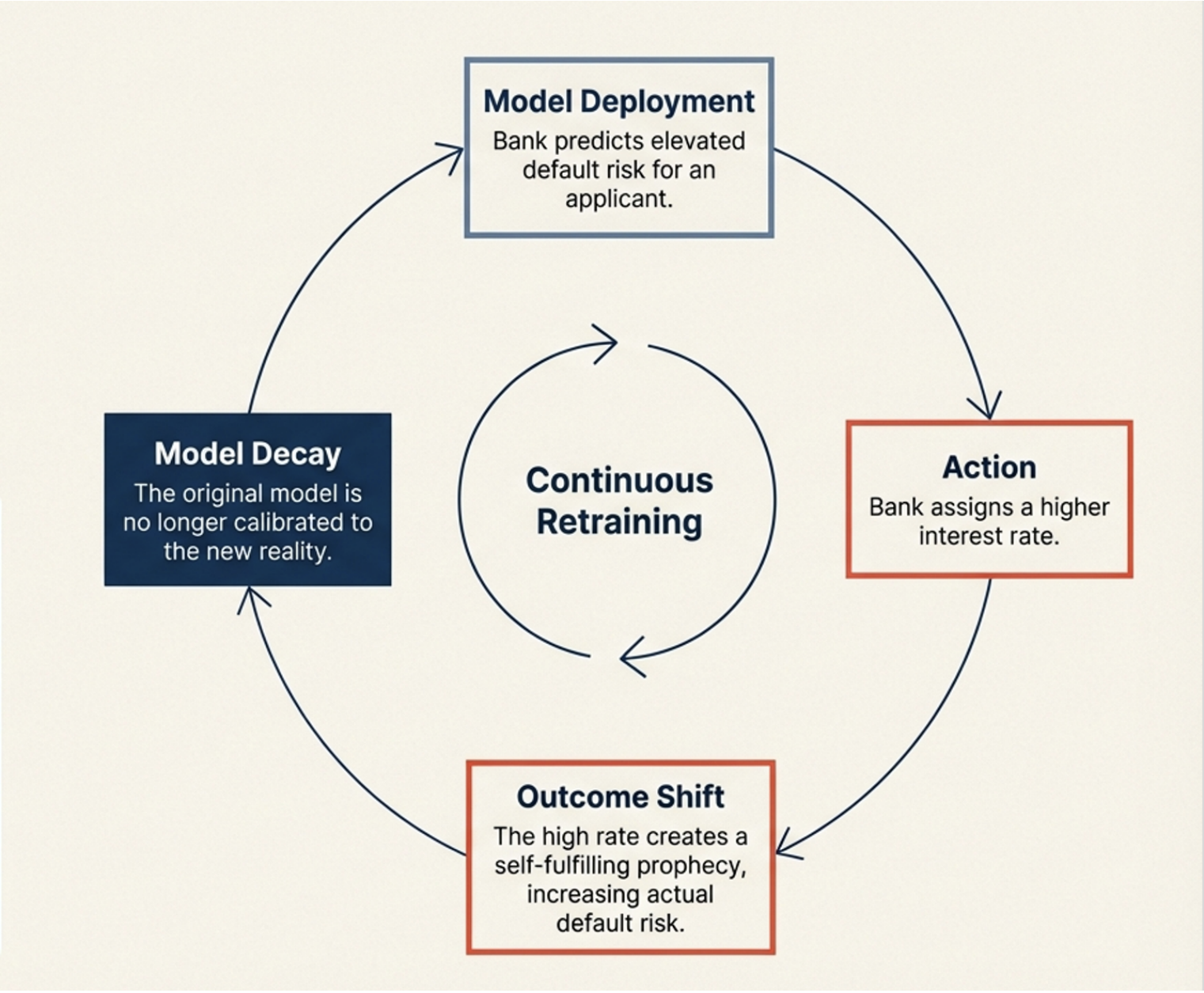
Performativity is ubiquitous:

- Traffic predictions → traffic patterns
- Crime prediction → police allocations
- Recommendations → preferences
- Stock prediction → trading prices

The Problem with Ignoring Performativity

The distribution $\mathcal{D}$ shifts as the model is deployed. Practitioners retrain — a *"cat and mouse game"* of chasing an ever moving target.

Motivation


Model Deployment
Bank predicts elevated default risk for an applicant.
Action
Bank assigns a higher interest rate.
Outcome Shift
The high rate creates a self-fulfilling prophecy, increasing actual default risk.
Model Decay
The original model is no longer calibrated to the new reality.
Continuous Retraining

Setup: The Performative Prediction Framework

	Classical Supervised Learning	Performative Prediction
Core Assumption	Static World	Reactive World
Data Distribution	Fixed D	Induced D(θ)
The Objective	Empirical Risk	Performative Risk PR(θ)
The Ultimate Goal	Finding the True Optimum	Finding Performative Stability

Setup: The Performative Prediction Framework

The Distribution Map $D(\cdot)$

We introduce a map $D : \Theta \to \Delta(\mathcal{Z})$ where $D(\theta)$ is the distribution over instances **that results from deploying** f_θ.

Notation:

- $\theta \in \Theta \subseteq \mathbb{R}^d$ — model parameters (closed, convex set)
- $Z = (X, Y)$ — feature-outcome pairs, $x \in \mathbb{R}^{m-1}$, $y \in \mathbb{R}$
- $\ell(z; \theta)$ — loss of model f_θ at instance z
- $D(\theta)$ — the **induced distribution** when θ is deployed

Crucially, the map $D(\cdot)$ is assumed **unknown** *to the decision-maker.*

Bank Credit Example

Notation in context:

- $\theta \in \Theta \subseteq \mathbb{R}^d$: logistic regression parameters (the bank's decision boundary)
- X features: demographic data, credit history, monthly income, number of open credit lines
- $Y \in \{0, 1\}$:whether the applicant defaults (1) or not (0)
- $\ell(z; \theta)$:logistic loss measuring prediction quality of default
- $D(\theta)$:distribution over applicants **after** strategic adaptation:

The bank does not know how applicants will react — yet deploying θ changes the very data it will observe next, creating a feedback loop.

Performative Optimality

Definition (2.1 Performative Risk & Optimality)

A model $f_{\theta_{PO}}$ is performatively optimal if the following relationship holds:

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

Where $\mathrm{PR}(\theta) \overset{def}{=} \mathbb{E}_{Z \sim \mathcal{D}(\theta)} \ell(Z; \theta)$ is the performative risk; then, $\theta_{PO} = \arg\min_{\theta} \mathrm{PR}(\theta)$

Sweep over all θ. For each one, imagine deploying it and the world shifting to $D(\theta)$. Measure your loss in that world, and pick the best. θ^{PO} is the winner of that sweep.

Performative Stability

Definition (2.3 Decoupled Risk & Stability)

A model $f_{\theta_{PS}}$ is performatively stable if the following relationship holds:

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

Where $\mathrm{DPR}(\theta, \theta') \overset{def}{=} \mathbb{E}_{Z \sim \mathcal{D}(\theta)} \ell(Z; \theta')$ is the decoupled performative risk; then, $\theta_{PS} = \arg\min_{\theta} \mathrm{DPR}(\theta_{PS}, \theta)$

If I deploy θ^{PS}, the world becomes $D(\theta^{\mathrm{PS}})$. I retrain on that world, and the best model on that world is again θ^{PS}. It's a fixed point where retraining doesn't move you.

Epsilon Sensitivity

Definition (3.1 Epsilon Sensitivity)

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

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

Where W_1 denotes the Wasserstein-1 distance or earth movers distance

- "Earth movers distance" is a metric of distributions. It essentially tells you how much work is needed to move mass from one distribution to another.
- ϵ-sensitivity is then saying: if I make a small change to my model, the world can only change by a proportional amount.

Joint Smoothness

Let $\mathcal{Z} \sim \cup_{\theta \in \Theta} supp\mathcal{D}(\theta)$ where *supp* denotes the support.

(A1) Joint Smoothness

$\ell(z;\theta)$ is β-jointly smooth if $\nabla_\theta \ell(z;\theta)$ is β-Lipschitz in both θ and z:

$$\|\nabla_\theta \ell(z;\theta) - \nabla_\theta \ell(z;\theta')\|_2 \leq \beta \|\theta - \theta'\|_2$$

$$\|\nabla_\theta \ell(z;\theta) - \nabla_\theta \ell(z';\theta)\|_2 \leq \beta \|z - z'\|_2$$

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

Strong Convexity

(A2) Strong Convexity

$\ell(z;\theta)$ is γ-strongly convex if for all $\theta, \theta' \in \Theta$, $z \in \mathcal{Z}$:

$$\ell(z;\theta) \geq \ell(z;\theta') + \nabla_\theta \ell(z;\theta')^\top (\theta - \theta') + \frac{\gamma}{2}\|\theta - \theta'\|_2^2$$

Repeated Risk Minimization

Definition 3.3: Repeated Risk Minimization (RRM)

Starting from an initial model f_{θ_0}, perform the following update for every $t \geq 0$:

$$\theta_{t+1} = G(\theta_t) \stackrel{\text{def}}{=} \arg\min_{\theta \in \Theta} \mathbb{E}_{Z \sim D(\theta_t)} \ell(Z; \theta)$$

$\Rightarrow$ Deploy Model

$\Rightarrow$ Observe Induced World

$\Rightarrow$ Retrain Model

$\Rightarrow$ Repeat

The question is: when does this process settle down rather than chase its own tail forever?

Convergence of Retraining

Theorem (3.5)

Suppose that the lose $\ell(z;\theta)$ is β-jointly smooth and γ-strongly convex. If the distribution map $D(\cdot)$ is ϵ-sensitive, then the following statements are true:

- $||G(\theta) - G(\theta')||_2 \leq \epsilon\frac{\beta}{\gamma}||\theta - \theta'||_2$, *for all* $\theta, \theta' \in \Theta$
- *If* $\epsilon < \frac{\gamma}{\beta}$, *the iterates* θ_t *converge to a unique performatively stable point* θ_{PS} *at a linear rate:*

$$||\theta_t - \theta_{PS}||_2 \leq \delta \text{ for } t \geq (1 - \epsilon\frac{\beta}{\gamma})^{-1} log(\frac{||\theta_0 - \theta_{PS}||_2}{\delta})$$

If the loss function is sufficiently "nice" and the distribution map is sufficiently (in)sensitive, then one need only retrain a model a small number of times before it converges to a unique stable point.

Caveats

The convergence result is **tight**. Removing any single assumption is enough to construct a counterexample where RRM diverges.

Proposition 3.6

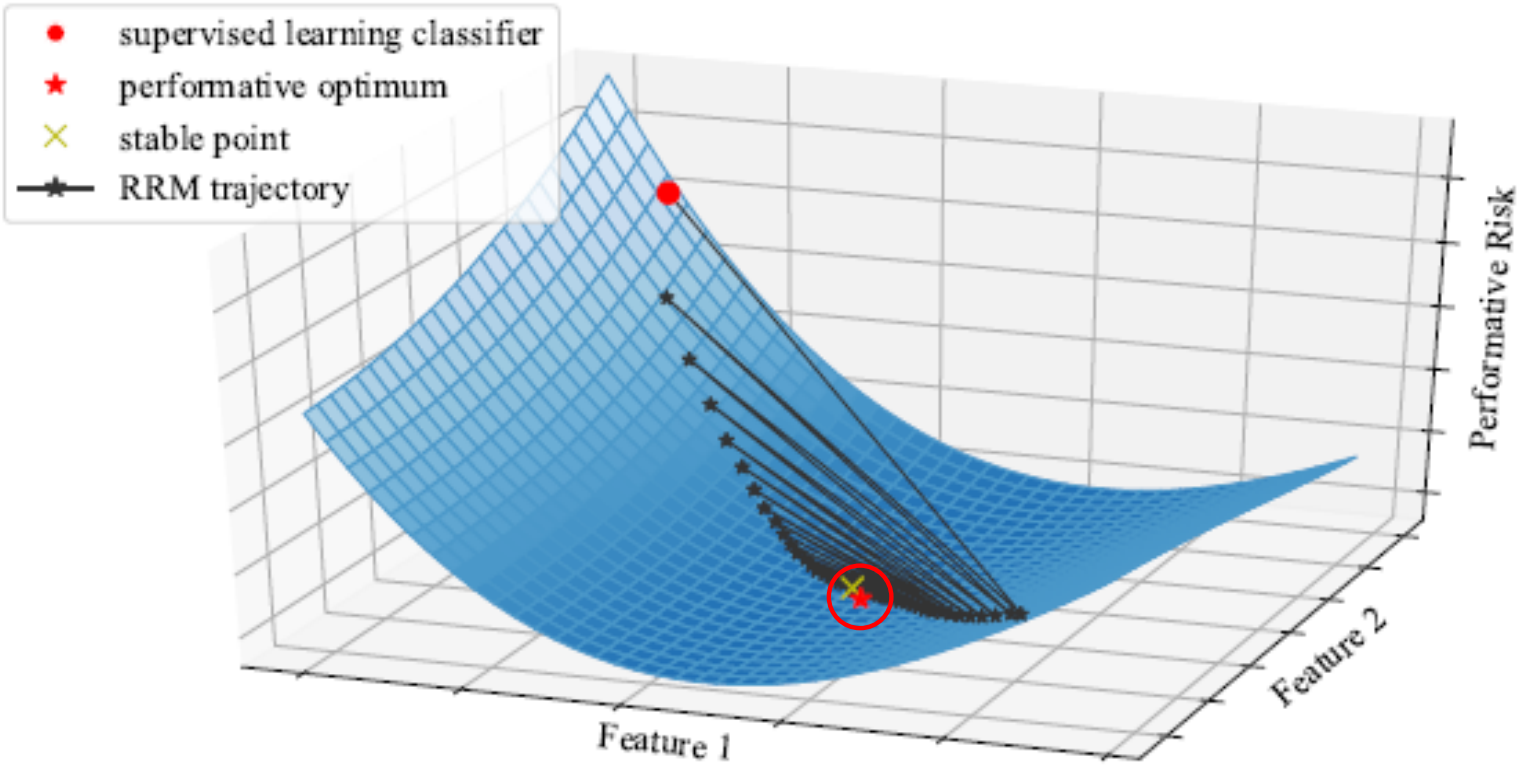
Suppose that the distribution map $\mathcal{D}(\cdot)$ is ϵ-sensitive with $\epsilon > 0$. RRM can fail to converge at all in any of the following cases, for any choice of parameters $\beta, \gamma > 0$:

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

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

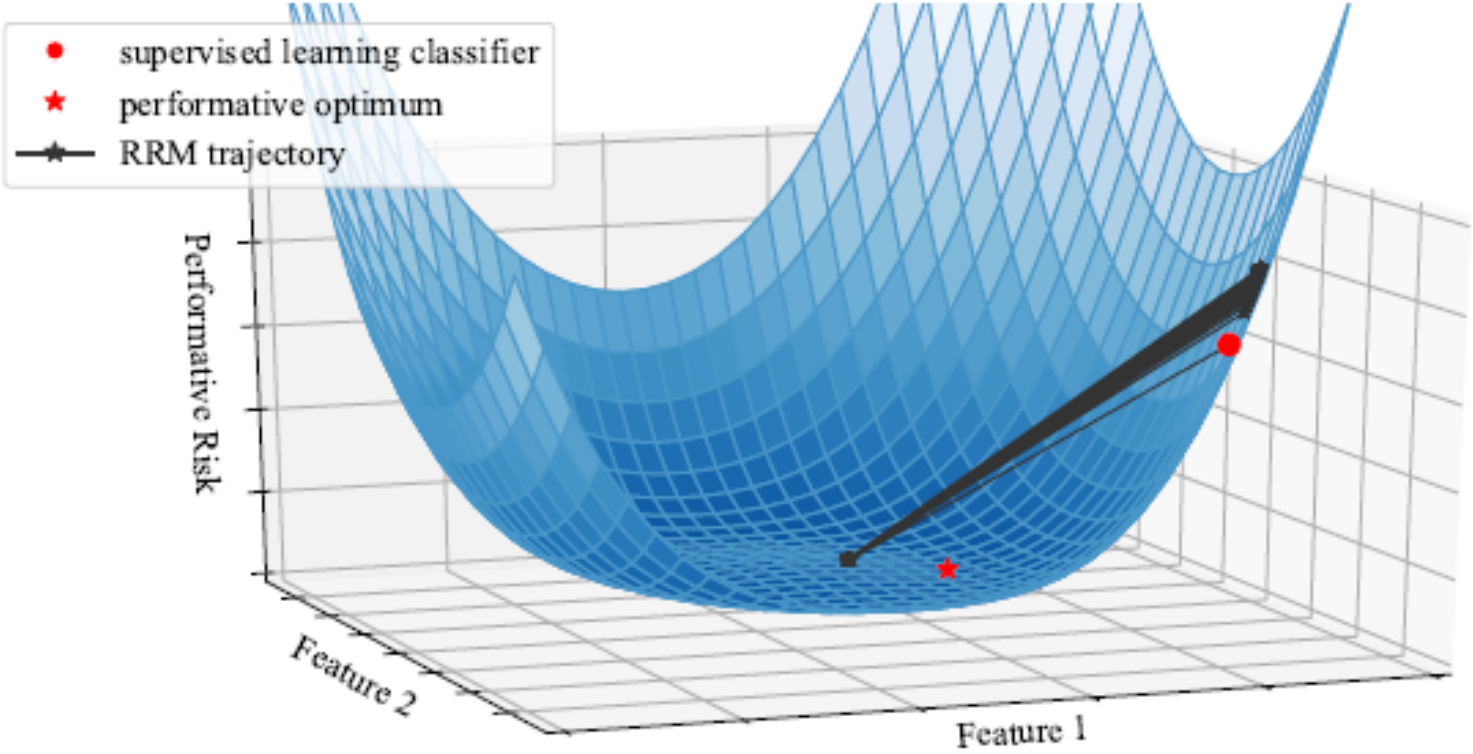
(c) The loss is β-jointly smooth and γ-strongly convex, but $\epsilon \geq \frac{\gamma}{\beta}$.

Caveats



(a) $\varepsilon = 25$

Caveats



(b) $\varepsilon = 100$

Repeated Gradient Descent (RGD)

Definition 3.7: Repeated Gradient Descent (RGD)

Starting from an initial model f_{θ_0}, perform the following update for every $t \geq 0$:

$$\theta_{t+1} = G_{\mathrm{gd}}(\theta_t) \stackrel{\mathrm{def}}{=} \Pi_\Theta \left(\theta_t - \eta \, \mathbb{E}_{Z \sim D(\theta_t)} \nabla_\theta \ell(Z; \theta_t) \right)$$

where $\eta > 0$ is a fixed step size and Π_Θ is the Euclidean projection onto Θ.

- RGD only requires ℓ to be **differentiable** in θ — no need to solve a full optimization problem.
- We do **not** take gradients of the performative risk $\mathrm{PR}(\theta)$ (which would require differentiating through $D(\theta)$). We only take the gradient of the loss at the current distribution.

Convergence of RGD (Theorem 3.8)

Theorem 3.8

Suppose $\ell(z;\theta)$ is β-jointly smooth and γ-strongly convex. If $D(\cdot)$ is ϵ-sensitive with $\epsilon < \frac{\gamma}{(\beta+\gamma)(1+1.5\eta\beta)}$, then RGD with $\eta \leq \frac{2}{\beta+\gamma}$ satisfies:

(a) $||G_{gd}(\theta) - G_{gd}(\theta')||_2 \leq (1 - \eta(\frac{\beta\gamma}{\beta+\gamma} - \epsilon(1.5\eta\beta^2 + \beta)))||\theta - \theta'||_2$

(b) The iterates θ_t of RGD converge to a unique performatively stable point θ_{PS} at a linear rate

RGD converges under a **stricter** condition on ϵ than RRM ($\frac{\gamma}{(\beta+\gamma)(1+1.5\eta\beta)}$ vs. $\frac{\gamma}{\beta}$). So, ϵ must be less than $\frac{\gamma}{(\beta+\gamma)(1+1.5\eta\beta)}$

RRM vs. RGD: Key Differences

RRM

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

- **Full** optimization each round
- Expensive per iteration
- Few iterations to converge
- Tolerates $\epsilon < \frac{\gamma}{\beta}$

RGD

$$\theta_{t+1} = \Pi_\Theta\left(\theta_t - \eta\, \mathbb{E}_{Z\sim D(\theta_t)} \nabla_\theta \ell(Z;\theta_t)\right)$$

- **Single** gradient step each round
- Cheap per iteration
- Many iterations
- Requires $\epsilon < \frac{\gamma}{(\beta+\gamma)(1+1.5\eta\beta)}$

RRM makes maximal progress per step, so it is more robust to distribution shift. RGD is computationally cheaper, but more vulnerable to the world changing between updates — it needs a stricter bound on ϵ.

Finite-Sample Extensions: RERM & REGD

Definition 3.9 (RERM & REGD)

At every iteration $t \geq 0$, collect n_t samples from $D(\theta_t)$.

RERM:

$$\theta_{t+1} = G^{n_t}(\theta_t) \stackrel{\text{def}}{=} \arg\min_{\theta} \mathbb{E}_{Z \sim D^{n_t}(\theta_t)} \ell(Z; \theta)$$

REGD:

$$\theta_{t+1} = G^{n_t}_{\text{gd}}(\theta_t) \stackrel{\text{def}}{=} \Pi_{\Theta}\left(\theta_t - \eta\, \mathbb{E}_{Z \sim D^{n_t}(\theta_t)} \nabla_{\theta} \ell(Z; \theta_t)\right)$$

With enough samples at each iteration, both algorithms converge with high probability to a **small neighborhood** around θ_{PS} but not exactly to it, since finite-sample noise prevents exact contraction once you are close.

Stable Points Approximate Optima (Theorem 4.3)

Theorem 4.3

Suppose $\ell(z;\theta)$ is L_z-Lipschitz in z, γ-strongly convex, and $D(\cdot)$ is ϵ-sensitive. Then for every performatively stable θ_{PS} and every performative optimum θ_{PO}:

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

What this tells us:

- All stable points and all optima live in a **small neighborhood** of each other, controlled by ϵ/γ.
- Low sensitivity (ϵ small) or strong convexity (γ large) $\Rightarrow$ stable points are near-optimal.
- **Practical implication:** running RRM to find θ_{PS} gives a model that approximately minimizes performative risk — even though we never directly optimized $\mathrm{PR}(\theta)$.

Strategic Classification as Performative Prediction

The setup: A two-player game (Bank Vs. Individual)

- The Bank deploys a classifier f_θ (e.g., deciding creditworthiness)
- **Individuals/Agents** who adapt their features to obtain a favorable classification (e.g., applicants manipulating their profiles)

This has a **Stackelberg structure**: the bank moves first (deploys f_θ), then agents/ individuals best-respond to maximize utility.

Distribution Map for Strategic Classification

Given base distribution D, classifier f_θ, cost c, and utility u:

1. Sample $(x, y) \sim D$
2. Compute best response: $x_{BR} \leftarrow \arg\max_{x'} \; u(x', \theta) - c(x', x)$
3. Output (x_{BR}, y)

Stackelberg Equilibria are Performative Optima

The institution's optimal strategy is the **Stackelberg equilibrium**: the classifier that minimizes loss over the distribution where agents have already best-responded.

Key Equivalence

$$f_{\theta_{SE}} \text{ is a Stackelberg equilibrium } \iff \theta_{SE} \in \arg\min_{\theta} \ \mathrm{PR}(\theta)$$

What this buys us:

- All of the theory we developed for performative prediction transfers directly to strategic classification.
- In particular, we can ask: does **retraining** converge in the strategic setting? This was previously an open question.

Retraining Overcomes Strategic Effects (Corollary 5.1)

Corollary 5.1

Let the institution's loss $\ell(z;\theta)$ be L_z- and L_θ-Lipschitz, β-jointly smooth, and γ-strongly convex. If $D(\cdot)$ is ϵ-sensitive with $\epsilon < \frac{\gamma}{\beta}$, then RRM converges at a linear rate to a performatively stable classifier θ_{PS} satisfying:

$$\theta_{PS} - \theta_{SE} \leq \frac{2L_z\epsilon(L_\theta + L_z\epsilon)}{\gamma}$$

What this says:

- The "retrain and redeploy" heuristic that practitioners already use in strategic settings is **theoretically justified** under natural conditions.
- The resulting classifier is not just stable — it is **near-optimal** in performative risk, i.e., close to the Stackelberg equilibrium.

Ethical and Technical Risks of Performative Prediction

Self-Fulfilling Prophecies: The bank's action creates the default, leading to systemic bias

Social Cost: In strategic classification, users adapting to models can incur unfair impacts to others.

The Technical Catch-22: Without strong convexity, retraining can oscillate forever or fail to find stable ground. Performative risk ($\mathrm{PR}(\theta)$) can become non-convex or concave, leading to poor stable points for both the bank and individuals.

In summary, the 'performative aspect' (high sensitivity ε) significantly biases actual outcomes, potentially leading to systemic instability or social harm.

Key Takeaways

- **Retraining as equilibration**
- **All three assumptions are necessary**
- **Stability $\approx$ optimality**
- **Broad applicability**