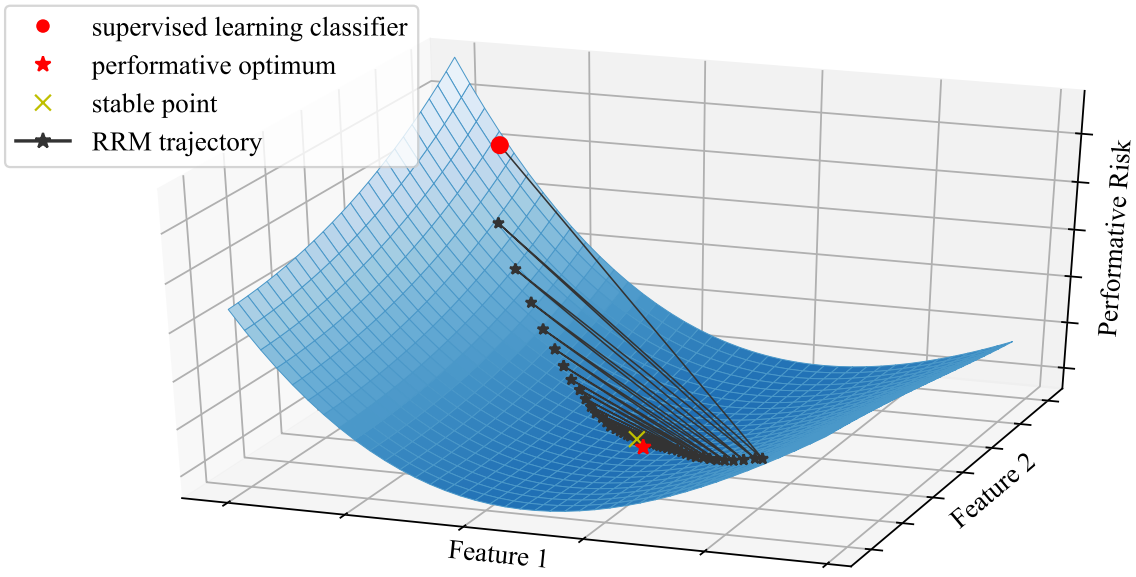
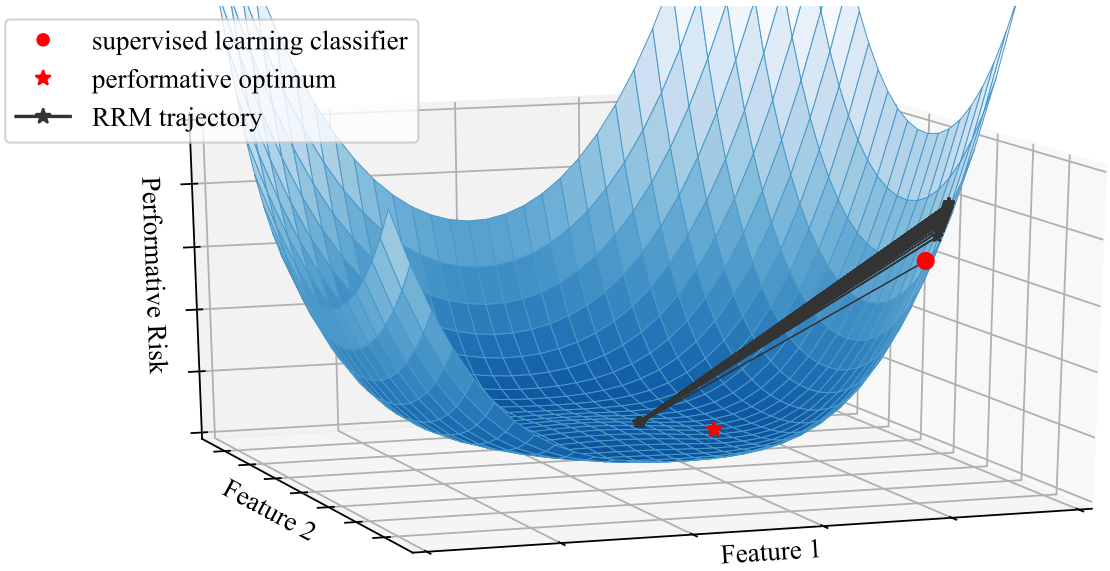
A. Visualizing the Performative Risk and Trajectory of RRM



(a) $\varepsilon = 25$



(b) $\varepsilon = 100$

Figure 4. Performative risk surface and trajectory of repeated risk minimization for two different values of sensitivity parameter ε. The initial iterate is the risk minimizer on the base dataset (●). We mark the performative optimum (⋆) and performatively stable point (×).

We provide additional experimental results in which we visualize the trajectory of repeated risk minimization on the surface of the performative risk. We adopt the general setting of Section 5. However, to properly visualize the loss, we rerun the experiments on a reduced version of the dataset with only two features (i.e $x \in \mathbb{R}^2$), both of which are adapted strategically according to the update described in Section 5.

Figure 4 plots the performative risk surface, together with the trajectory of RRM given by straight black lines. The top plot shows the trajectory for a suitably small sensitivity parameter ε. We see that RRM converges to a stable point which is close to the performative optimum. We contrast this behavior with that of RRM when ε is large in the bottom plot. Here, we observe that the iterates oscillate and that the algorithm fails to converge.

Both plots mark the risk minimizer on the initial data set ($\bullet$), before any strategic adaptation takes place. This point also corresponds to the initial iterate of RRM θ_0. We additionally mark the performative optimum ($\star$) on the risk curve. The top plot additionally marks the last iterate of RRM, which serves as a proxy for the performatively stable point ($\times$). As predicted by our theory, this stable point is in a small neighborhood around the performative optimum.

B. Applications of Performativity

To illustrate the fact that performativity is a common cause for concept drift, we review a table of concepts drift applications from Žliobaitė et al. (2016). In Table 1, we highlight those settings that naturally occur due to performativity. Below we briefly discuss the role of performativity in such applications.

Indust. \ Appl.	**Monitoring & control**	**Information management**	**Analytics & diagnostics**
Security, Police	fraud detection, insider trading detection, adversary actions detection	next crime place prediction	crime volume prediction
Finance, Banking, Telecom, Insurance, Marketing, Retail, Advertising	monitoring & management of customer segments, bankruptcy prediction	product or service recommendation, including complimentary, user intent or information need prediction	demand prediction, response rate prediction, budget planning
Production industry	controlling output quality	-	predict bottlenecks
Education (e-Learning, e-Health), Media, Entertainment	gaming the system, drop out prediction	music, VOD, movie, news, learning object personalized search & recommendations	player-centered game design, learner-centered education

Table 1. Table of concept drift applications from Žliobaitė et al. (2016).

The role of *fraud detection* systems is to predict whether an instance such as a transaction or email is legitimate or not. It is well-known that designers of such fraudulent instances adapt to the fraud detection system in place in order to breach security. Therefore, the deployment of fraud detection systems shapes the features of fraudulent instances.

Crime place prediction uses historical data to estimate the likelihood of crime at a given location. Those locations where criminal behavior is deemed likely by the system typically get more police patrols and better surveillance. These actions resulting from prediction significantly decrease the probability of crime taking place, thus changing the data used for future predictions.

In *personalized recommendations*, instances are recommended to a user based on their historical context, such as their ratings or purchases. The set of recommendations thus depends on the trained machine learning model, which in turn changes the user's future ratings or purchases. In other words, user features serving as input to a recommender inevitably depend on the previously used recommendation mechanisms.

In *online two-player games*, it is common to request an AI opponent. The level of sophistication of the AI opponent might be chosen depending on the user's success history in the given game, with the goal of making the game appropriately challenging. This choice of AI opponent changes players' future success profiles, again causing a distribution shift in the features serving as an input to the prediction system.

Gaming the system falls under the umbrella of strategic classification, which we discuss in detail in Section 5, so we avoid further discussion in this section.

C. Detailed Derivation of Example 2.2

Since we have closed form expressions for all the relevant distributional quantities, we can write out a precise expression of the performative risk in our biased coin toss example. In particular, by factoring out the expectation we have that,

$$\mathrm{PR}(\theta) = \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}(\theta)} \big(Y - f_\theta(X)\big)^2 = \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}(\theta)} \Big[\mathbb{E}\left[\big(Y - f_\theta(X)\big)^2 \quad X\right]\Big],$$

where all the distributions are taken with respect to $\mathcal{D}(\theta)$. If we now expand out the squared loss, we can write,

$$\mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}(\theta)} \Big[\mathbb{E}\left[\big(Y - f_\theta(X)\big)^2 \quad X\right]\Big] = \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}(\theta)} \left[\mathbb{E}[Y^2|X] - 2\,\mathbb{E}[Y|X] f_\theta(X) + f_\theta(X)^2\right].$$

Using our knowledge of the distribution map, we know that

$$\mathbb{E}[Y^2|X] = \mathbb{E}[Y|X] = \frac{1}{2} + \mu X + \varepsilon\theta X.$$

Plugging these into our previous expression and simplifying terms, we find that we can write the performative risk as,

$$\mathrm{PR}(\theta) = \mathbb{E}\left[\frac{1}{4} - 2\mu\theta X^2 + (1-2\varepsilon)\theta^2 X^2\right] = \frac{1}{4} - 2\mu\theta + (1-2\varepsilon)\theta^2.$$

Importantly, we see that the distribution over X is irrelevant for the performative risk, since $X^2 = 1$ with probability 1. The performative risk is then a quadratic function of θ. The minimizer of this objective, the performative optimum, is therefore

$$\theta_{\mathrm{PO}} = \frac{\mu}{1-2\varepsilon}.$$

For performative stability, we recall a classifier θ_{PS} is performatively stable if and only if

$$\theta_{\mathrm{PS}} = \mathop{\arg\min}_{\theta\in[0,1]} \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta_{\mathrm{PS}})} \big(Y - f_\theta(X)\big)^2.$$

Furthermore, we know that the unique minimizer of the squared loss is the conditional expectation function. Therefore, if we can find a classifier within our class that satisfies

$$f_\theta(x) = \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}(\theta)}[Y \mid X = x], \quad \forall x,$$

then this classifier is the (unique) performatively stable classifier. Rewriting this above expression, we have that a classifier f_θ is stable if

$$\frac{1}{2} + \varepsilon\theta x + \mu x = \frac{1}{2} + \theta x.$$

Solving the above expression for θ, we get

$$\theta_{\mathrm{PS}} = \frac{\mu}{1-\varepsilon}.$$

As a final note, another way of seeing that there is a unique performatively stable classifier is that this example satisfies the conditions of Theorem 3.8. Repeated risk minimization is hence a contraction which implies that there must be a unique stable classifier.

D. Auxiliary Lemmas

Lemma D.1 (Bubeck (2015), Proposition 1.3)**.** *Let f be convex and let Ω be a closed convex set on which f is differentiable, then*

$$x_* \in \mathop{\arg\min}_{x\in\Omega} f(x)$$

if and only if

$$\nabla f(x_*)^T (y - x_*) \geqslant 0, \quad \forall y \in \Omega.$$

Lemma D.2 (Bubeck (2015), Lemma 3.11)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be β-smooth and γ-strongly convex, then for all x, y in $\mathbb{R}^d$,*

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geqslant \frac{\gamma\beta}{\gamma + \beta}\|x - y\|_2^2 + \frac{1}{\gamma + \beta}\|\nabla f(x) - \nabla f(y)\|_2^2.$$

Lemma D.3 (Kantorovich-Rubinstein)**.** *A distribution map $\mathcal{D}(\cdot)$ is ε-sensitive if and only if for all $\theta, \theta' \in \Theta$:*

$$\sup \left\{ \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta)} g(Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} g(Z) \;\leqslant \varepsilon\|\theta - \theta'\|_2 \;:\; g : \mathbb{R}^p \to \mathbb{R},\ g \text{ 1-Lipschitz} \right\}. \tag{1}$$

Lemma D.4. *Let $f : \mathbb{R}^n \to \mathbb{R}^d$ be an L-Lipschitz function, and let $X, X' \in \mathbb{R}^n$ be random variables such that $W_1(X, X') \leqslant C$. Then*

$$\|\mathbb{E}[f(X)] - \mathbb{E}[f(X')]\|_2 \leqslant LC.$$

Proof.

$$\begin{aligned}\|\mathbb{E}[f(X)] - \mathbb{E}[f(X')]\|_2^2 &= (\mathbb{E}[f(X)] - \mathbb{E}[f(X')])^\top (\mathbb{E}[f(X)] - \mathbb{E}[f(X')]) \\ &= \|\mathbb{E}[f(X)] - \mathbb{E}[f(X')]\|_2 \frac{(\mathbb{E}[f(X)] - \mathbb{E}[f(X')])^\top}{\|\mathbb{E}[f(X)] - \mathbb{E}[f(X')]\|_2} (\mathbb{E}[f(X)] - \mathbb{E}[f(X')]).\end{aligned}$$

Now define the unit vector $v := \frac{\mathbb{E}[f(X)] - \mathbb{E}[f(X')]}{\|\mathbb{E}[f(X)] - \mathbb{E}[f(X')]\|_2}$. By linearity of expectation, we can further write

$$\|\mathbb{E}[f(X)] - \mathbb{E}[f(X')]\|_2^2 = \|\mathbb{E}[f(X)] - \mathbb{E}[f(X')]\|_2 \, (\mathbb{E}[v^\top f(X)] - \mathbb{E}[v^\top f(X')]).$$

For any unit vector v and L-Lipschitz function f, $v^\top f$ is a one-dimensional L-Lipschitz function, so we can apply Lemma D.3 to obtain

$$\|\mathbb{E}[f(X)] - \mathbb{E}[f(X')]\|_2^2 \leqslant \|\mathbb{E}[f(X)] - \mathbb{E}[f(X')]\|_2 LC.$$

Canceling out $\|\mathbb{E}[f(X)] - \mathbb{E}[f(X')]\|_2$ from both sides concludes the proof.

■

E. Proofs of Main Results

E.1. Proof of Theorem 3.5

Fix $\theta, \theta' \in \Theta$. Let $f(\varphi) = \mathbb{E}_{Z\sim\mathcal{D}(\theta)} \ell(Z; \varphi)$ and $f'(\varphi) = \mathbb{E}_{Z\sim\mathcal{D}(\theta')} \ell(Z; \varphi)$. Since f is γ-strongly convex and $G(\theta)$ is the unique minimizer of $f(x)$ we know that,

$$\begin{aligned}f(G(\theta)) - f(G(\theta')) &\geqslant (G(\theta) - G(\theta'))^\top \nabla f(G(\theta')) + \frac{\gamma}{2}\|G(\theta) - G(\theta')\|_2^2, \\ f(G(\theta')) - f(G(\theta)) &\geqslant \frac{\gamma}{2}\|G(\theta) - G(\theta')\|_2^2.\end{aligned}$$

Together, these two inequalities imply that

$$-\gamma\|G(\theta) - G(\theta')\|_2^2 \geqslant (G(\theta) - G(\theta'))^\top \nabla f(G(\theta')).$$

Next, we observe that $(G(\theta) - G(\theta'))^\top \nabla_\theta \ell(z; G(\theta'))$ is $\|G(\theta) - G(\theta')\|_2\beta$-Lipschitz in z. This follows from applying Cauchy-Schwarz and the fact that the loss is β-jointly smooth. Using the dual formulation of the earth mover's distance (Lemma D.3) and ε-sensitivity of $\mathcal{D}(\cdot)$, we can write

$$(G(\theta) - G(\theta'))^\top \nabla f(G(\theta')) - (G(\theta) - G(\theta'))^\top \nabla f'(G(\theta')) \geqslant -\varepsilon\beta\|G(\theta) - G(\theta')\|_2\|\theta - \theta'\|_2.$$

Furthermore, using the first-order optimality conditions for convex functions, we have $(G(\theta) - G(\theta'))^\top \nabla f'(G(\theta')) \geqslant 0$, and hence $(G(\theta) - G(\theta'))^\top \nabla f(G(\theta')) \geqslant -\varepsilon\beta\|G(\theta) - G(\theta')\|_2\|\theta - \theta'\|_2$. Therefore, we conclude that,

$$-\gamma\|G(\theta) - G(\theta')\|_2^2 \geqslant -\varepsilon\beta\|G(\theta) - G(\theta')\|_2\|\theta - \theta'\|_2.$$

Claim (a) then follows by rearranging.

To prove claim (b) we note that $\theta_t = G(\theta_{t-1})$ by the definition of RRM, and $G(\theta_{\mathrm{PS}}) = \theta_{\mathrm{PS}}$ by the definition of stability. Applying the result of part (a) yields

$$\|\theta_t - \theta_{\mathrm{PS}}\|_2 \leqslant \varepsilon\frac{\beta}{\gamma}\|\theta_{t-1} - \theta_{\mathrm{PS}}\|_2 \leqslant \left(\varepsilon\frac{\beta}{\gamma}\right)^t \|\theta_0 - \theta_{\mathrm{PS}}\|_2.$$

Setting this expression to be at most δ and solving for t completes the proof of claim (b). Alternatively, part (b) can be concluded by a direct application of the Banach fixed-point theorem to part (a).

E.2. Proof of Proposition 3.6

Proof of (a): Consider the linear loss defined as $\ell((x, y); \theta) = \beta y\theta$, for $\theta \in [-1, 1]$. Note that this objective is β-jointly smooth and convex, but not strongly convex. Let the distribution of Y according to $\mathcal{D}(\theta)$ be a point mass at $\varepsilon\theta$, and let the distribution of X be invariant with respect to θ. Clearly, this distribution is ε-sensitive.

Here, the decoupled performative risk has the following form $\mathrm{DPR}(\theta, \varphi) = \varepsilon\beta\theta\varphi$. The unique performatively stable point is 0. However, if we initialize RRM at any point other than 0, the procedure generates the sequence of iterates $\ldots, 1, -1, 1, -1 \ldots$, thus failing to converge. Furthermore, this behavior holds for all $\varepsilon, \beta > 0$.

Proof of (b): Consider a type of regularized hinge loss $\ell(z; \theta) = C\max(-1, y\theta) + \frac{\gamma}{2}(\theta - 1)^2$, and suppose $\Theta \supseteq [-\frac{1}{2\varepsilon}, \frac{1}{2\varepsilon}]$.

Let the distribution of Y according to $\mathcal{D}(\theta)$ be a point mass at $\varepsilon\theta$, and let the distribution of X be invariant with respect to θ. Clearly, this distribution is ε-sensitive.

Let $\theta_0 = 2$. Then, by picking C big enough, RRM prioritizes to minimize the first term exactly, and hence we get $\theta_1 = -\frac{1}{2\varepsilon}$. In the next step, again due to large C, we get $\theta_2 = 2$. Thus, RRM keeps oscillating between 2 and $-\frac{1}{2\varepsilon}$, failing to converge. This argument holds for all $\gamma, \varepsilon > 0$.

Proof of (c): Suppose that the loss function is the squared loss, $\ell(z; \theta) = (y - \theta)^2$, where $y, \theta \in \mathbb{R}$. Note that this implies $\beta = \gamma$. Let the distribution of Y according to $\mathcal{D}(\theta)$ be a point mass at $1 + \varepsilon\theta$, and let the distribution of X be invariant with respect to θ. This distribution family satisfies ε-sensitivity, because

$$W_1(\mathcal{D}(\theta), \mathcal{D}(\theta')) = \varepsilon|\theta - \theta'|.$$

By properties of the squared loss, we know

$$\arg\min_{\theta'} \mathrm{DPR}(\theta, \theta') = \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta)}[Y] = 1 + \varepsilon\theta.$$

It is thus not hard to see that RRM does not contract if $\varepsilon \geqslant \frac{\gamma}{\beta} = 1$:

$$|G(\theta) - G(\theta')| = |1 + \varepsilon\theta - 1 - \varepsilon\theta'| = \varepsilon|\theta - \theta'|,$$

which exactly matches the bound of Theorem 3.5 and proves the first statement of the proposition. The unique performatively stable point of this problem is θ such that $\theta = 1 + \varepsilon\theta$, which is $\theta_{\mathrm{PS}} = \frac{1}{1-\varepsilon}$ for $\varepsilon > 1$.

For $\varepsilon = 1$, no performatively stable point exists, thereby proving the second claim of the proposition. If $\varepsilon > 1$ on the other hand, and $\theta_0 \neq \theta_{\mathrm{PS}}$, we either have $\theta_t \to \infty$ or $\theta_t \to -\infty$, because

$$\theta_t = 1 + \varepsilon\theta_{t-1} = \sum_{k=0}^{t-1}\varepsilon^k + \theta_0\varepsilon^t = \frac{\varepsilon^t - 1}{\varepsilon - 1} + \theta_0\varepsilon^t,$$

thus concluding the proof.

E.3. Proof of Theorem 3.8

Since projecting onto a convex set can only bring two iterates closer together, in this proof we ignore the projection operator Π_Θ and treat G_{gd} as performing merely the gradient step.

We begin by expanding out $\|G_{\mathrm{gd}}(\theta) - G_{\mathrm{gd}}(\theta')\|_2^2$,

$$\begin{aligned}
\|G_{\mathrm{gd}}(\theta) - G_{\mathrm{gd}}(\theta')\|_2^2 &= \Big\| \theta - \eta \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta)} \nabla_\theta \ell(Z;\theta) - \theta' + \eta \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta') \Big\|_2^2 \\
&= \|\theta - \theta'\|_2^2 - 2\eta(\theta - \theta')^\top \left(\mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta)} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta') \right) \\
&\quad + \eta^2 \Big\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta)} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta') \Big\|_2^2 \\
&\stackrel{\mathrm{def}}{=} T_1 - 2\eta T_2 + \eta^2 T_3.
\end{aligned}$$

Next, we analyze each term individually,

$$\begin{aligned}
T_1 &\stackrel{\mathrm{def}}{=} \|\theta - \theta'\|_2^2, \\
T_2 &\stackrel{\mathrm{def}}{=} (\theta - \theta')^\top \left(\mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta)} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta') \right), \\
T_3 &\stackrel{\mathrm{def}}{=} \| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta)} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta') \|_2^2.
\end{aligned}$$

We start by lower bounding T_2:

$$\begin{aligned}
T_2 &= (\theta - \theta')^\top \left(\mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta)} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta) + \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta') \right) \\
&= (\theta - \theta')^\top \left(\mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta)} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta) \right) + (\theta - \theta')^\top \left(\mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta') \right) \\
&\geqslant -\|\theta - \theta'\|_2 \, \| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta)} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta) \|_2 + (\theta - \theta')^\top \left(\mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta') \right),
\end{aligned}$$

where in the last step we apply the Cauchy-Schwarz inequality. By smoothness, $\nabla_\theta \ell(Z;\theta)$ is β-Lipschitz in Z. Together with the fact that Z is ε-sensitive, we can lower bound the first term in the above expression by applying Lemma D.4, which results in $-\beta\varepsilon\|\theta - \theta'\|_2^2$.

We apply Lemma D.2 to lower bound the second term by

$$\begin{aligned}
& (\theta - \theta')^\top \left(\mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta') \right) \\
\geqslant\ & \frac{\beta\gamma}{\beta+\gamma}\|\theta - \theta'\|_2^2 + \frac{1}{\beta+\gamma} \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \left[\|\nabla_\theta \ell(Z;\theta) - \nabla_\theta \ell(Z;\theta')\|_2^2 \right] \\
\geqslant\ & \frac{\beta\gamma}{\beta+\gamma}\|\theta - \theta'\|_2^2 + \frac{1}{\beta+\gamma} \Big\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \left[\nabla_\theta \ell(Z;\theta) - \nabla_\theta \ell(Z;\theta') \right] \Big\|_2^2,
\end{aligned}$$

where we have applied Jensen's inequality in the last line. Putting everything together, we get

$$T_2 \geqslant \left(\frac{\beta\gamma}{\beta+\gamma} - \beta\varepsilon \right) \|\theta - \theta'\|_2^2 + \frac{1}{\beta+\gamma} \Big\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \left[\nabla_\theta \ell(Z;\theta) - \nabla_\theta \ell(Z;\theta') \right] \Big\|_2^2.$$

Now we upper bound T_3. We begin by expanding out the square just as before,

$$\begin{aligned}
T_3 &= \Big\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta)} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta) + \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta') \Big\|_2^2 \\
&= \Big\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta)} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta) \Big\|_2^2 + \Big\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta') \Big\|_2^2 \\
&\quad + 2 \left(\mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta)} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta) \right)^\top \left(\mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta') \right).
\end{aligned} \tag{2}$$

We again bound each term individually. By the smoothness of the loss and Lemma D.4,

$$\left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta)} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta) \right\|_2^2 \leqslant \beta^2\varepsilon^2 \|\theta-\theta'\|_2^2.$$

Moving on to the last term in (2):

$$\begin{aligned}
&2\left(\mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta)} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta) \right)^\top \left(\mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta') \right) \\
\overset{\text{def}}{=}\ & 2 \left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta') \right\|_2 \left(\mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta)} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta) \right)^\top v \\
=\ & 2 \left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta') \right\|_2 \left(\mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta)} v^\top \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} v^\top \nabla_\theta \ell(Z;\theta) \right),
\end{aligned}$$

where we define the unit vector $v \overset{\text{def}}{=} \frac{\mathbb{E}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta) - \mathbb{E}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta')}{\|\mathbb{E}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta) - \mathbb{E}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta')\|_2}$. By smoothness of the loss, we can conclude that $v^\top \nabla_\theta \ell(Z,\theta)$ is β-Lipschitz, so by ε-sensitivity we get

$$\begin{aligned}
&2\left(\mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta)} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta) \right)^\top \left(\mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta') \right) \\
\leqslant\ & 2 \left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta') \right\|_2 \beta\varepsilon\|\theta-\theta'\|_2 \\
\leqslant\ & 2\beta^2\varepsilon\|\theta-\theta'\|_2^2,
\end{aligned}$$

where in the last step we again apply smoothness. Hence,

$$T_3 \leqslant (\varepsilon^2\beta^2 + 2\beta^2\varepsilon)\|\theta-\theta'\|_2^2 + \left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta') \right\|_2^2.$$

Having bounded all the terms, we now conclude that

$$\begin{aligned}
\|G_{\text{gd}}(\theta) - G_{\text{gd}}(\theta')\|_2^2 \leqslant & \left(1 + \eta^2\varepsilon^2\beta^2 + 2\eta^2\beta^2\varepsilon - 2\eta\frac{\beta\gamma}{\beta+\gamma} + 2\eta\beta\varepsilon\right)\|\theta-\theta'\|_2^2 \\
& - \left(\frac{2\eta}{\beta+\gamma} - \eta^2\right) \left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta')} \nabla_\theta \ell(Z;\theta') \right\|_2^2.
\end{aligned}$$

If we take the step size η to be small enough, namely $\eta \leqslant \frac{2}{\beta+\gamma}$, we get

$$\|G_{\text{gd}}(\theta) - G_{\text{gd}}(\theta')\|_2^2 \leqslant \left(1 + \eta^2\varepsilon^2\beta^2 + 2\eta^2\beta^2\varepsilon - 2\eta\frac{\beta\gamma}{\beta+\gamma} + 2\eta\beta\varepsilon\right)\|\theta-\theta'\|_2^2.$$

To ensure a contraction, we need $2\eta\frac{\beta\gamma}{\beta+\gamma} - \eta^2\varepsilon^2\beta^2 - 2\eta^2\beta^2\varepsilon - 2\eta\beta\varepsilon > 0$. Canceling out $\eta\beta$, and assuming $\varepsilon \leqslant 1$, it suffices to have $\frac{2\gamma}{\beta+\gamma} - 3\eta\varepsilon\beta - 2\varepsilon > 0$. Therefore, if $\varepsilon < \frac{\gamma}{(\beta+\gamma)(1+1.5\eta\beta)} \leqslant 1$, the map G_{gd} is contractive. In particular, we have

$$\begin{aligned}
\|G_{\text{gd}}(\theta) - G_{\text{gd}}(\theta')\|_2 &\leqslant \sqrt{\left(1 - \eta\left(2\frac{\beta\gamma}{\beta+\gamma} - \varepsilon(3\eta\beta^2 + 2\beta)\right)\right)}\|\theta-\theta'\|_2 \\
&\leqslant \left(1 - \eta\left(\frac{\beta\gamma}{\beta+\gamma} - \varepsilon(1.5\eta\beta^2 + \beta)\right)\right)\|\theta-\theta'\|_2,
\end{aligned}$$

where we use the fact that $\sqrt{1-x} \leqslant 1 - \frac{x}{2}$ for $x \in [0,1]$. This completes the proof of part (a).

Since we have shown G_{gd} is contractive, by the Banach fixed-point theorem we know that there exists a unique fixed point of G_{gd}. That is, there exists θ_{PS} such that $\mathbb{E}_{Z\sim\mathcal{D}(\theta_{\text{PS}})} \nabla_\theta \ell(Z;\theta_{\text{PS}}) = 0$. By convexity of the loss function, this means that θ_{PS}

is the optimum of $\mathbb{E}_{Z\sim\mathcal{D}(\theta_{\mathrm{PS}})}\,\ell(Z;\theta)$ over θ, which in turn implies that θ_{PS} is performatively stable. Recursively applying the result of part (a) we get the rate of convergence of RRM to θ_{PS}:

$$\begin{aligned}\|\theta_t-\theta_{\mathrm{PS}}\|_2 &\leqslant \left(1-\eta\left(\frac{\beta\gamma}{\beta+\gamma}-\varepsilon(1.5\eta\beta^2+\beta)\right)\right)^t\|\theta_0-\theta_{\mathrm{PS}}\|_2\\ &\leqslant \exp\left(-t\eta\left(\frac{\beta\gamma}{\beta+\gamma}-\varepsilon(1.5\eta\beta^2+\beta)\right)\right)\|\theta_0-\theta_{\mathrm{PS}}\|_2,\end{aligned}$$

where in the last step we use the fact that $1-x\leqslant e^{-x}$. Setting this expression to be at most δ and solving for t completes the proof.

E.4. Proof of Theorem 3.10

Proof of (a): We introduce the main proof idea and then present the full argument. The proof proceeds by case analysis. First, we show that if $\|\theta_t-\theta_{\mathrm{PS}}\|_2>\delta$, performing ERM ensures that with high probability $\|\theta_{t+1}-\theta_{\mathrm{PS}}\|_2\leqslant 2\varepsilon\frac{\beta}{\gamma}\|\theta_t-\theta_{\mathrm{PS}}\|_2$. Using our assumption that $\varepsilon<\frac{\gamma}{2\beta}$, this implies that the iterate θ_{t+1} contracts toward θ_{PS}.

On the other hand, if $\|\theta_t-\theta_{\mathrm{PS}}\|_2\leqslant\delta$, we show that while ERM might not contract, it cannot push θ_{t+1} too far from θ_{PS} either. In particular, θ_{t+1} must be in a $\frac{\varepsilon\beta}{2\gamma}\delta$-ball around θ_{PS}. The proof then concludes by arguing that θ_t for $t\geqslant\frac{\log(\|\theta_0-\theta_{\mathrm{PS}}\|_2/\delta)}{\log(\gamma/2\varepsilon\beta)}$ must enter a ball of radius δ around θ_{PS}. Once this event occurs, no future iterate can exit the $\frac{\varepsilon\beta}{2\gamma}\delta$-ball around θ_{PS}.

Case 1: $\|\theta_t-\theta_{\mathrm{PS}}\|_2>\delta$. If the current iterate is outside the ball, we show that with high probability the next iterate contracts towards a performatively stable point. In particular,

$$\|\theta_{t+1}-\theta_{\mathrm{PS}}\|_2\leqslant\frac{2\varepsilon\beta}{\gamma}\|\theta_t-\theta_{\mathrm{PS}}\|_2.$$

To prove this claim, we begin by showing that

$$W_1(\mathcal{D}^{n_t}(\theta_t),\mathcal{D}(\theta_{\mathrm{PS}}))\leqslant 2\varepsilon\|\theta_t-\theta_{\mathrm{PS}}\|_2,\ \text{with probability } 1-\frac{6p}{\pi^2t^2}. \tag{3}$$

Since the W_1-distance is a metric on the space of distributions, we can apply the triangle inequality to get

$$W_1(\mathcal{D}^{n_t}(\theta_t),\mathcal{D}(\theta_{\mathrm{PS}}))\leqslant W_1(\mathcal{D}^{n_t}(\theta_t),\mathcal{D}(\theta_t))+W_1(\mathcal{D}(\theta_t),\mathcal{D}(\theta_{\mathrm{PS}})).$$

The second term is bounded deterministically by $\varepsilon\|\theta_t-\theta_{\mathrm{PS}}\|_2$ due to ε-sensitivity. By Theorem 2 of Fournier & Guillin (2015), for $n_t\geqslant\frac{1}{c_2(\varepsilon\delta)^m}\log\left(\frac{t^2\pi^2c_1}{6p}\right)$, the probability that the first term is greater than $\varepsilon\delta$ is less that $\frac{6p}{t^2\pi^2}$. Here, the positive constants c_1,c_2 depend on $\alpha,\mu,\xi_{\alpha,\mu}$ and m. Therefore,

$$W_1(\mathcal{D}^{n_t}(\theta_t),\mathcal{D}(\theta_{\mathrm{PS}}))\leqslant\varepsilon\delta+\varepsilon\|\theta_t-\theta_{\mathrm{PS}}\|_2\leqslant 2\varepsilon\|\theta_t-\theta_{\mathrm{PS}}\|_2,\ \text{with probability } 1-\frac{6p}{\pi^2t^2}.$$

Using this, we can now prove that the iterates contract. Following the first steps of the proof of Theorem 3.5, we have that

$$\begin{aligned}&\left(G^{n_t}(\theta_t)-G(\theta_{\mathrm{PS}})\right)^\top\left(\mathop{\mathbb{E}}_{Z\sim\mathcal{D}^{n_t}(\theta_t)}\nabla_\theta\ell(Z;G^{n_t}(\theta_t))-\mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta_{\mathrm{PS}})}\nabla_\theta\ell(Z;G^{n_t}(\theta_t))\right)\\ &+\left(G^{n_t}(\theta_t)-G(\theta_{\mathrm{PS}})\right)^\top\left(\mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta_{\mathrm{PS}})}\nabla_\theta\ell(Z;G^{n_t}(\theta_t))-\mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta_{\mathrm{PS}})}\nabla_\theta\ell(Z;G(\theta_{\mathrm{PS}}))\right)\leqslant 0.\end{aligned} \tag{4}$$

Like in the proof of Theorem 3.5, the term $\left(G^{n_t}(\theta_t)-G(\theta_{\mathrm{PS}})\right)^\top\mathbb{E}_{Z\sim\mathcal{D}^{n_t}(\theta_t)}\nabla_\theta\ell(Z;G^{n_t}(\theta_t))$ is $\|G^{n_t}(\theta_t)-G(\theta_{\mathrm{PS}})\|_2\cdot\beta$ Lipschitz in Z. Using equation (3), with probability $1-\frac{6p}{\pi^2t^2}$ we can bound the first term by

$$\begin{aligned}&\left(G^{n_t}(\theta_t)-G(\theta_{\mathrm{PS}})\right)^\top\left(\mathop{\mathbb{E}}_{Z\sim\mathcal{D}^{n_t}(\theta_t)}\nabla_\theta\ell(Z;G^{n_t}(\theta_t))-\mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta_{\mathrm{PS}})}\nabla_\theta\ell(Z;G^{n_t}(\theta_t))\right)\\ \geqslant\ &-2\varepsilon\beta\left\|G^{n_t}(\theta_t)-G(\theta_{\mathrm{PS}})\right\|_2\|\theta_t-\theta_{\mathrm{PS}}\|_2.\end{aligned}$$

And by strong convexity,

$$\begin{aligned}&(G^{n_t}(\theta_t) - G(\theta_{\mathrm{PS}}))^\top \Big(\mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta_{\mathrm{PS}})} \nabla_\theta \ell(Z; G^{n_t}(\theta_t)) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta_{\mathrm{PS}})} \nabla_\theta \ell(Z; G(\theta_{\mathrm{PS}}))\Big)\\ \geqslant\ & \gamma \|G^{n_t}(\theta_t) - G(\theta_{\mathrm{PS}})\|_2^2.\end{aligned}$$

Plugging back into equation (4), we conclude that with high probability

$$\|\theta_{t+1} - \theta_{\mathrm{PS}}\|_2 \leqslant \frac{2\varepsilon\beta}{\gamma} \|\theta_t - \theta_{\mathrm{PS}}\|_2.$$

Applying a union bound, we conclude that the iterates contract at every iteration where $\|\theta_t - \theta_{\mathrm{PS}}\|_2 > \delta$ with probability at least $1 - \sum_{t=1}^{\infty} \frac{6p}{\pi^2 t^2} = 1 - p$. Therefore, for $t \geqslant \left(1 - \frac{2\varepsilon\beta}{\gamma}\right)^{-1} \log\left(\frac{\|\theta_0 - \theta_{\mathrm{PS}}\|_2}{\delta}\right)$ steps we have

$$\|\theta_t - \theta_{\mathrm{PS}}\|_2 \leqslant \left(\frac{2\varepsilon\beta}{\gamma}\right)^t \|\theta_0 - \theta_{\mathrm{PS}}\|_2 \leqslant \left(\frac{2\varepsilon\beta}{\gamma}\right)^t \|\theta_0 - \theta_{\mathrm{PS}}\|_2 \leqslant \exp\left(-t\left(1 - \frac{2\varepsilon\beta}{\gamma}\right)\right) \|\theta_0 - \theta_{\mathrm{PS}}\|_2 \leqslant \delta,$$

where we use $1 - x \leqslant e^{-x}$. This implies that θ_t eventually contracts to a ball of radius δ around θ_{PS}.

Case 2: $\|\theta_t - \theta_{\mathrm{PS}}\|_2 \leqslant \delta$. We show that the RERM iterates can leave a ball of radius δ around θ_{PS} only with negligible probability. We begin by applying the triangle inequality just as we did in the previous case,

$$W_1(\mathcal{D}^{n_t}(\theta_t), \mathcal{D}(\theta_{\mathrm{PS}})) \leqslant W_1(\mathcal{D}^{n_t}(\theta_t), \mathcal{D}(\theta_t)) + W_1(\mathcal{D}(\theta_t), \mathcal{D}(\theta_{\mathrm{PS}})) \leqslant W_1(\mathcal{D}^{n_t}(\theta_t), \mathcal{D}(\theta_t)) + \varepsilon\delta.$$

For our choice of n_t, with probability at least $1 - \frac{6p}{\pi^2 t^2}$ this quantity is upper bounded by

$$W_1(\mathcal{D}^{n_t}(\theta_t), \mathcal{D}(\theta_{\mathrm{PS}})) \leqslant 2\varepsilon\delta.$$

With this information, we can now apply the exact same steps as in the previous case, but now using the fact that $W_1(\mathcal{D}^{n_t}(\theta_t), \mathcal{D}(\theta_{\mathrm{PS}})) \leqslant 2\varepsilon\delta$ instead of $W_1(\mathcal{D}^{n_t}(\theta_t), \mathcal{D}(\theta_{\mathrm{PS}})) \leqslant 2\varepsilon\|\theta_t - \theta_{\mathrm{PS}}\|_2$, to conclude that with probability at least $1 - \frac{6p}{\pi^2 t^2}$

$$\|\theta_{t+1} - \theta_{\mathrm{PS}}\|_2 \leqslant 2\varepsilon\frac{\beta}{\gamma}\delta \leqslant \delta.$$

As before, a union bound argument proves that the entire analysis holds with probability $1 - p$.

Proof of (b): The only difference between part (b) in relation to part (a) is the fact that one needs to invoke the steps of Theorem 3.8 rather than Theorem 3.5.

E.5. Proof of Proposition 4.2

We make a slight modification to Example 2.2 to prove the proposition. As in the example, $\mathcal{D}(\theta)$ is given as follows: X is a single feature supported on $\{\pm 1\}$ and $Y \mid X \sim \text{Bernoulli}(\frac{1}{2} + \mu X + \varepsilon\theta X)$, where $\Theta = [0, 1]$. We let $\varepsilon \geqslant \frac{1}{2}$, and constrain μ to satisfy $|\mu + \varepsilon| \leqslant \frac{1}{2}$. We assume that outcomes are predicted according to the model $f_\theta(x) = \theta x + \frac{1}{2}$ and that performance is measured via the squared loss, $\ell(z; \theta) = (y - f_\theta(x))^2$. This loss has condition number $\frac{\beta}{\gamma} = 1$.

A direct calculation demonstrates that the performative risk is a quadratic in θ:

$$\mathrm{PR}(\theta) = \frac{1}{4} - 2\theta\mu + (1 - 2\varepsilon)\theta^2.$$

Therefore, if $\varepsilon \in \left[\frac{1}{2}, 1\right)$, the performative risk is a concave function of θ, even though $\varepsilon < \frac{\gamma}{\beta}$.

E.6. Proof of Theorem 4.3

By definition of performative optimality and performative stability we have that:

$$\mathrm{DPR}(\theta_{\mathrm{PO}}, \theta_{\mathrm{PO}}) \leqslant \mathrm{DPR}(\theta_{\mathrm{PS}}, \theta_{\mathrm{PS}}) \leqslant \mathrm{DPR}(\theta_{\mathrm{PS}}, \theta_{\mathrm{PO}}).$$

We claim that $\mathrm{DPR}(\theta_{\mathrm{PS}}, \theta_{\mathrm{PO}}) - \mathrm{DPR}(\theta_{\mathrm{PS}}, \theta_{\mathrm{PS}}) \geqslant \frac{\gamma}{2}\|\theta_{\mathrm{PO}} - \theta_{\mathrm{PS}}\|_2^2$. By definition of DPR, we can write

$$\mathrm{DPR}(\theta_{\mathrm{PS}}, \theta_{\mathrm{PO}}) - \mathrm{DPR}(\theta_{\mathrm{PS}}, \theta_{\mathrm{PS}}) = \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta_{\mathrm{PS}})}\left[\ell(Z;\theta_{\mathrm{PO}}) - \ell(Z;\theta_{\mathrm{PS}})\right].$$

Since $\ell(z;\theta_{\mathrm{PO}}) \geqslant \ell(z;\theta_{\mathrm{PS}}) + \nabla_\theta \ell(z;\theta_{\mathrm{PS}})^\top(\theta_{\mathrm{PO}} - \theta_{\mathrm{PS}}) + \frac{\gamma}{2}\|\theta_{\mathrm{PO}} - \theta_{\mathrm{PS}}\|_2^2$ for all z, we have that

$$\mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta_{\mathrm{PS}})}\left[\ell(Z;\theta_{\mathrm{PO}}) - \ell(Z;\theta_{\mathrm{PS}})\right] \geqslant \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta_{\mathrm{PS}})}\left[\nabla_\theta \ell(Z;\theta_{\mathrm{PS}})^\top(\theta_{\mathrm{PO}} - \theta_{\mathrm{PS}})\right] + \frac{\gamma}{2}\|\theta_{\mathrm{PO}} - \theta_{\mathrm{PS}}\|_2^2. \tag{5}$$

Now, by Lemma D.1, $\mathbb{E}_{Z\sim\mathcal{D}(\theta_{\mathrm{PS}})}\left[\nabla_\theta \ell(Z;\theta_{\mathrm{PS}})^\top(\theta_{\mathrm{PO}} - \theta_{\mathrm{PS}})\right] \geqslant 0$, so we get that equation (5) implies

$$\mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta_{\mathrm{PS}})}\left[\ell(Z;\theta_{\mathrm{PO}}) - \ell(Z;\theta_{\mathrm{PS}})\right] \geqslant \frac{\gamma}{2}\|\theta_{\mathrm{PO}} - \theta_{\mathrm{PS}}\|_2^2.$$

Since the population distributions are ε-sensitive and the loss is L_z-Lipschitz in z, we have that $\mathrm{DPR}(\theta_{\mathrm{PS}}, \theta_{\mathrm{PO}}) - \mathrm{DPR}(\theta_{\mathrm{PO}}, \theta_{\mathrm{PO}}) \leqslant L_z\varepsilon\|\theta_{\mathrm{PO}} - \theta_{\mathrm{PS}}\|_2$. If $\varepsilon < \frac{\gamma\|\theta_{\mathrm{PO}} - \theta_{\mathrm{PS}}\|_2}{2L_z}$ then we have that $L_z\varepsilon\|\theta_{\mathrm{PO}} - \theta_{\mathrm{PS}}\|_2 < \frac{\gamma}{2}\|\theta_{\mathrm{PO}} - \theta_{\mathrm{PS}}\|_2^2$ which is a contradiction since it must hold that

$$\mathrm{DPR}(\theta_{\mathrm{PS}}, \theta_{\mathrm{PO}}) - \mathrm{DPR}(\theta_{\mathrm{PO}}, \theta_{\mathrm{PO}}) \geqslant \mathrm{DPR}(\theta_{\mathrm{PS}}, \theta_{\mathrm{PO}}) - \mathrm{DPR}(\theta_{\mathrm{PS}}, \theta_{\mathrm{PS}}).$$

E.7. Proof of Corollary 5.1

By Theorem 3.8 we know that repeated risk minimization converges at a linear rate to a performatively stable point θ_{PS}. Furthermore, by Theorem 4.3, this performatively stable point is close in domain to the institution's Stackelberg equilibrium classifier θ_{SE},

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

We can then use the fact that the loss is Lipschitz to show that this performatively stable classifier is close in objective value to the Stackelberg equilibrium:

$$\begin{aligned}
\mathrm{PR}(\theta_{\mathrm{PS}}) - \mathrm{PR}(\theta_{\mathrm{SE}}) &\leqslant \mathrm{PR}(\theta_{\mathrm{PS}}) - \mathrm{DPR}(\theta_{\mathrm{PS}}, \theta_{\mathrm{SE}}) + \mathrm{DPR}(\theta_{\mathrm{PS}}, \theta_{\mathrm{SE}}) - \mathrm{PR}(\theta_{\mathrm{SE}}) \\
&\leqslant L_\theta\|\theta_{\mathrm{SE}} - \theta_{\mathrm{PS}}\|_2 + L_z\varepsilon\|\theta_{\mathrm{SE}} - \theta_{\mathrm{PS}}\|_2 \\
&\leqslant \frac{2L_z\varepsilon(L_\theta + L_z\varepsilon)}{\gamma}
\end{aligned}$$

Here, we have used the Kantorovich-Rubinstein Lemma (D.3) to bound the second term.

F. Approximately Minimizing Performative Risk via Regularization

Recall that in Proposition 3.6 we have shown that RRM might not converge at all if the objective is smooth and convex, but not strongly convex. In this section, we show how adding a small amount of quadratic regularization to the objective guarantees that RRM will converge to a stable point which approximately minimizes the performative risk on the original loss.

To do so, we additionally require that the space of model parameters Θ be bounded with diameter $D = \sup_{\theta,\theta'\in\Theta}\|\theta - \theta'\|_2$. We can assume without loss of generality that $D = 1$.

Proposition F.1. *Suppose that the loss $\ell(z;\theta)$ is L_z-Lipschitz iz z and L_θ-Lipschitz in θ, β-jointly smooth (A1) and convex (but not necessarily strongly convex). Furthermore, suppose that distribution map $\mathcal{D}(\cdot)$ is ε-sensitive with $\varepsilon < 1$, and that the set Θ is bounded with diameter 1. Then, there exists a choice of α, such that running RRM with loss $\ell^{reg}(z;\theta) \stackrel{\mathrm{def}}{=} \ell(z;\theta) + \frac{\alpha}{2}\|\theta - \theta_0\|_2^2$ converges to a performatively stable point $\theta_{\mathrm{PS}}^{reg}$ which satisfies the following*

$$\mathrm{PR}(\theta_{\mathrm{PS}}^{reg}) \leqslant \min_\theta \mathrm{PR}(\theta) + O\left(\frac{\sqrt{\varepsilon}}{1-\varepsilon}\right).$$

We note that in the case where $\varepsilon = 0$, the limit point $\theta_{\mathrm{PS}}^{\mathrm{reg}}$ of regularized repeated risk minimization is also performatively optimal.

Proof. First, we observe that the regularized loss function $\ell^{\text{reg}}(z;\theta)$ is α-strongly convex and $\alpha+\beta$-jointly smooth. Since $\varepsilon < 1$, we can then choose an α such that $\varepsilon < \frac{\alpha}{\alpha+\beta}$. In particular, we choose $\alpha = \sqrt{\varepsilon}\beta/(1-\varepsilon)$.

From our choice of α, we have that ε is smaller than the inverse condition number. Hence, by Theorem 3.5 repeated risk minimization converges at a linear rate to a performatively stable solution θ^{reg}_{PS} of the regularized objective.

To finish the proof, we show that the objective value at the θ^{reg}_{PS} is close to the objective value at the performative optima of the original objective θ_{PO}. We do so by bounding their difference using the triangle inequality:

$$\begin{aligned}\mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta^{\text{reg}}_{\text{PS}})}\ell^{\text{reg}}(Z;\theta^{\text{reg}}_{\text{PS}}) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta_{\text{PO}})}\ell(Z;\theta_{\text{PO}}) &= \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta^{\text{reg}}_{\text{PS}})}\ell^{\text{reg}}(Z;\theta^{\text{reg}}_{\text{PS}}) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta^{\text{reg}}_{\text{PO}})}\ell^{\text{reg}}(Z;\theta^{\text{reg}}_{\text{PO}}) \\ &\quad + \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta^{\text{reg}}_{\text{PO}})}\ell^{\text{reg}}(Z;\theta^{\text{reg}}_{\text{PO}}) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta_{\text{PO}})}\ell(Z;\theta_{\text{PO}})\end{aligned}$$

We can bound the first difference via Lipschitzness:

$$\begin{aligned}\mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta^{\text{reg}}_{\text{PS}})}\ell^{\text{reg}}(Z;\theta^{\text{reg}}_{\text{PS}}) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta^{\text{reg}}_{\text{PO}})}\ell^{\text{reg}}(Z;\theta^{\text{reg}}_{\text{PO}}) &= \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta^{\text{reg}}_{\text{PS}})}\ell^{\text{reg}}(Z;\theta^{\text{reg}}_{\text{PS}}) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta^{\text{reg}}_{\text{PS}})}\ell^{\text{reg}}(Z;\theta^{\text{reg}}_{\text{PO}}) \\ &\quad + \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta^{\text{reg}}_{\text{PS}})}\ell^{\text{reg}}(Z;\theta^{\text{reg}}_{\text{PO}}) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta^{\text{reg}}_{\text{PO}})}\ell^{\text{reg}}(Z;\theta^{\text{reg}}_{\text{PO}}) \\ &\leqslant (L_\theta + \alpha \sup_{\theta,\theta'\in\Theta}\|\theta-\theta'\|_2)\|\theta^{\text{reg}}_{\text{PS}} - \theta^{\text{reg}}_{\text{PO}}\|_2 \\ &\quad + \varepsilon L_z\|\theta^{\text{reg}}_{\text{PS}} - \theta^{\text{reg}}_{\text{PO}}\|_2 \\ &= (L_\theta + \alpha + \varepsilon L_z)\|\theta^{\text{reg}}_{\text{PS}} - \theta^{\text{reg}}_{\text{PO}}\|_2 \\ &\leqslant \frac{2(L_\theta+\alpha+\varepsilon L_z)L_z\varepsilon}{\alpha}.\end{aligned}$$

In the last two lines, we have applied the fact that $D = \sup_{\theta,\theta'\in\Theta}\|\theta-\theta'\|_2 = 1$ as well as Theorem 4.3. For the second difference, by definition of performative optimality we have that,

$$\mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta^{\text{reg}}_{\text{PO}})}\ell^{\text{reg}}(Z;\theta^{\text{reg}}_{\text{PO}}) \leqslant \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta_{\text{PO}})}\ell^{\text{reg}}(Z;\theta_{\text{PO}}) \leqslant \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta_{\text{PO}})}\ell(Z;\theta_{\text{PO}}) + \frac{\alpha}{2}.$$

Where we have again used the fact that $D = 1$ for the last inequality. Combining these two together, we can bound the total difference:

$$\mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta^{\text{reg}}_{\text{PS}})}\ell^{\text{reg}}(Z;\theta^{\text{reg}}_{\text{PS}}) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta_{\text{PO}})}\ell(Z;\theta_{\text{PO}}) \leqslant \frac{2(L_\theta+\alpha+\varepsilon L_z)L_z\varepsilon}{\alpha} + \frac{\alpha}{2}.$$

Plugging in $\alpha = \frac{\sqrt{\varepsilon}\beta}{1-\varepsilon}$ completes the proof. ■

G. Experimental Details

Base distribution. The base distribution consists of the Kaggle data set (Kaggle, 2012). We subsample $n = 18,357$ points from the original training set such that both classes are approximately balanced (45% of points have y equal to 1). There are a total of 10 features, 3 of which we treat as strategic features: utilization of credit lines, number of open credit lines, and number of real estate loans. We scale features in the base distribution so that they have zero mean and unit variance.

Verifying ε-sensitivity. We verify that the map $\mathcal{D}(\cdot)$, as described in Section 5, is ε-sensitive. To do so, we analyze $W_1(\mathcal{D}(\theta), \mathcal{D}(\theta'))$, for arbitrary $\theta, \theta' \in \Theta$. Fix a sample point $x \in \mathbb{R}^{m-1}$ from the base dataset. Because the base distribution $\mathcal{D}$ is supported on n points, we can upper bound the optimal transport distance between any pair of distributions $\mathcal{D}(\theta)$ and $\mathcal{D}(\theta')$ by the Euclidean distance between the shifted versions of x in $\mathcal{D}(\theta)$ and $\mathcal{D}(\theta')$.

In our construction, the point x is shifted to $x - \varepsilon\theta$ and to $x - \varepsilon\theta'$ in $\mathcal{D}(\theta)$ and $\mathcal{D}(\theta')$ respectively. The distance between these two shifted points is $\|x - \varepsilon\theta - x + \varepsilon\theta'\|_2 = \varepsilon\|\theta - \theta'\|_2$. Since the same relationship holds for all other samples x in the base dataset, the optimal transport from $\mathcal{D}(\theta)$ to $\mathcal{D}(\theta')$ is at most $\varepsilon\|\theta - \theta'\|_2$.

Verifying joint smoothness of the objective. For the experiments described in Figure 1, we run repeated risk minimization and repeated gradient descent on the logistic loss with ℓ_2 regularization:

$$\frac{1}{n}\sum_{i=1}^{n} -y_i\theta^\top x_i + \log\left(1+\exp(\theta^\top x_i)\right) + \frac{\gamma}{2}\|\theta\|_2^2 \tag{6}$$

For both the repeated risk minimization and repeated gradient descent we set $\gamma = 1000/n$, where n is the size of the base dataset.

For a particular feature-outcome pair (x_i, y_i), the logistic loss is $\frac{1}{4}\|x_i\|_2^2$ smooth (Shalev-Shwartz & Ben-David, 2014). Therefore, the entire objective is $\frac{1}{4n}\sum_{i=1}^{n}\|x_i\|_2^2 + \gamma$ smooth. Due to the strategic updates, $x_{BR} = x - \varepsilon\theta$, the norm of individual features change depending on the choice of model parameters.

Theoretically, we can upper bound the smoothness of the objective by finding the implicit constraints on Θ, which can be revealed by looking at the dual of the objective function for every fixed value of ε. However, for simplicity, we simply calculate the worst-case smoothness of the objective, given the trajectory of iterates $\{\theta_t\}$, for every fixed ε.

Furthermore, we can verify the logistic loss is jointly smooth. For a fixed example $z = (x, y)$, the gradient of the regularized logistic loss with respect to θ is,

$$\nabla_\theta \ell(z;\theta) = yx + \frac{\exp(\theta^\top x)}{1+\exp(\theta^\top x)}x + \gamma\theta,$$

which is 2-Lipschitz in z due to $y \in \{0, 1\}$. Hence, the overall objective is β-jointly smooth with parameter

$$\beta = \max\left\{2, \frac{1}{4n}\sum_{i=1}^{n}\|x_i\|_2^2 + \gamma\right. \quad .$$

For RRM, ε is less than $\frac{\gamma}{\beta}$ only in the case that $\varepsilon = 0.01$. For RGD, ε is never smaller than the theoretical cutoff of $\frac{\gamma}{(\beta+\gamma)(1+1.5\eta\beta)}$.

Optimization details. The definition of RRM requires exact minimization of the objective at every iteration. We approximate this requirement by minimizing the objective described in expression (6) to small tolerance, 10^{-8}, using gradient descent. We choose the step size at every iteration using backtracking line search.

In the case of repeated gradient descent, we run the procedure as described in Definition 3.7 with a fixed step size of $\eta = \frac{2}{\beta+\gamma}$.