For all of the proofs in the Appendix, we assume WLOG that the utility of a positive outcomes γ is equal to 1.

A. Additional discussion of assumptions

A.1. Cost function

Let us discuss some context and implications of Assumption 1 defining a valid cost function. Unlike prior work (Milli et al., 2019; Miller et al., 2020; Braverman and Garg, 2020), we model a nonzero cost for *all* modifications to features, regardless of whether these modifications are in the right direction. In the spirit of generalizing beyond standard microfoundations, this accounts for how agents may erroneously expend effort on changing their features in an incorrect direction, as empirically demonstrated in Example 3. We further note that the definition of a valid cost function does not require symmetry in the arguments, which differentiates it from a metric.

A.2. Measurability requirements for alternative microfoundations

We now describe the measurability requirements that we need in order to define and work with maps $M \in \mathcal{M}$. If we ignore measurability requirements for a moment, then notice that each map $M \in \mathcal{M}$ can be associated with a distribution $\mathcal{D}_{TXY} \in \Delta(\mathcal{T} \times X \times Y)$ given by $(M(x,y),x,y)$. Since it is easier to define measurability requirements on $\mathcal{D}_{TXY}$, we specify requirements on $\mathcal{D}_{TXY}$, which gives an implicit specification of requirements on M. First, we define the probability space. Consider the sample space $\mathcal{T} \times X \times Y$. We can define a sigma algebra $\mathcal{F}$ over Ω by viewing $\mathcal{T}$ as the set of functions $X \times \Theta \to X$, and using that $X \subseteq \mathbb{R}^d, \Theta \subseteq \mathbb{R}^d$. The probability measure can then be given by $\mathcal{D}_{TXY}$.

Since image(M) = supp$(\mathcal{D}_{TXY})$ contains a very small fraction of the sample space $\mathcal{T} \times X \times Y$, we can work with a much smaller probability space in this context. This probability space is defined as follows: the sample space is supp$(\mathcal{D}_{TXY}) \in \mathcal{F}$ (i.e. a subset of $\mathcal{T} \times X \times Y$ in the sigma-algebra), and the sigma-algebra is intersections of every set in $\mathcal{F}$ with supp$(\mathcal{D}_{TXY})$. The probability measure given by $\mathcal{D}_{TXY}$ can be defined over this smaller probability space.

The distribution map $\mathcal{D}$ can thus be viewed as random variables over this probability space. In particular, $\mathcal{D}(\theta)$ is the distribution of the random variable $(\mathcal{R}_t(x,\theta),y)$. In order for this random variable to be well-defined, we place the following measurability assumption.

Assumption 2 (Measurability requirement on $\mathcal{R}$)**.** We require that for each $\theta \in \Theta$ the function $F_\theta : \text{supp}(\mathcal{D}_{TXY}) \to X \times Y$ given by $F_\theta(t,x,y) = (\mathcal{R}_t(x,\theta),y)$ is measurable.

A.3. Assumption on gaming behavior

In Proposition 3, we make the assumption that agent cannot have differing types solely on the basis of their true label. In other words, the map M cannot take into account the true label.

Assumption 3. For a map $M \in \mathcal{M}$, we require that $M(x,0) = M(x,1)$ for all $x \in X$.

Assumption 3 means that agents with features x who have true label 0 versus true label 1 have identical distributions over response types in aggregate. We need this assumption to reason about performatively optimal points, because a decision maker has no access to the true labels beyond agents' reported features when anticipating strategic behavior.

A.4. Compactness of X

The compactness assumption guarantees that the behavior of agents who follow standard microfoundations is well-defined.

Fact 1. *Suppose that c is a valid cost function, and X is compact. Then $\sup_{x' \in X}(f_\theta(x') - c(x,x'))$ is attained on some $x^* \in X$ and the behavior of rational agents with perfect information is well-defined.*

B. Proofs for Section 2

B.1. Proof of Proposition 1

In order to prove Proposition 1, we show that the gaming behavior of rational agents with perfect information can be characterized in the following way: Any rational agent with perfect information either will not change their features at all or

will change their features exactly up to the decision boundary. We use the notation:

$$R_{t_{\text{SM}}}(x,\theta) := \arg\max_{x' \in X}(f_\theta(x') - c(x,x')) \tag{5}$$

to denote how an agent with features x who follows standard microfoundations will change their features in response to f_θ.

Lemma 9. *Suppose that c is a valid cost function. Then for any x the response (5) is either $R_{t_{SM}}(x,\theta) = x$ or $R_{t_{SM}}(x,\theta)$ is on the decision boundary of f_θ.*

Proof of Lemma 9. By Fact 1, we know that the quantity $\arg\max_{x' \in X}(f_\theta(x') - c(x,x'))$ is well defined. It suffices to show that if $R_{t_{\text{SM}}}(x,\theta) \neq x$, then $R_{t_{\text{SM}}}(x,\theta)$ is on the decision boundary of f_θ. If $R_{t_{\text{SM}}}(x,\theta) \neq x$, then we know that $c(x, R_{t_{\text{SM}}}(x,\theta)) > 0$. This means that $f_\theta(x) = 0$ and $R_{t_{\text{SM}}}(x,\theta) \in \arg\max_{x' \in X}(f_\theta(x') - c(x,x')) = \operatorname{argmin}_{x' \in X_{\text{pos}}} c(x,x')$, where $X_{\text{pos}} := \{x \in X \mid f_\theta(x) = 1\}$. Assume for sake of contradiction that $R_{t_{\text{SM}}}(x,\theta)$ is not on the decision boundary. Then since $x \notin X_{\text{pos}}$ and $R_{t_{\text{SM}}}(x,\theta) \in X_{\text{pos}}$, there must exist x' on the line segment between x and $R_{t_{\text{SM}}}(x,\theta)$ such that x' is on boundary of X_{pos}, and thus the decision boundary of f_θ. Moreover, by Assumption 1, we know that $c(x,x') < c(x, R_{t_{\text{SM}}}(x,\theta))$. Since X_{pos} is closed, we see that $f_\theta(x') = 1$. Thus, $[f_\theta(x') - c(x,x')] < [f_\theta(R_{t_{\text{SM}}}(x,\theta)) - c(x,x')]$ which is a contradiction. □

Now, we use Lemma 9 to prove Proposition 1.

Proof of Proposition 1. It suffices to show that $\mathcal{D}(\theta)$ is either equal to $\mathcal{D}_{XY}$ or is a discontinuous distribution. Let $Q(\theta) \subseteq X$ be the set of agents who change their features at f_θ, i.e.

$$Q(\theta) := \left\{x \in X \mid R_{t_{\text{SM}}}(x,\theta) \neq x\right\}.$$

If $\mathbb{P}_{(x,y)\in\mathcal{D}_{XY}}[x \in Q(\theta)] = 0$, then $\mathcal{D}(\theta) = \mathcal{D}_{XY}$. Otherwise, suppose that $\mathbb{P}_{(x,y)\in\mathcal{D}_{XY}}[x \in Q(\theta)] > 0$. By Lemma 9, all of the agents in $Q(\theta)$ will game to somewhere on the decision boundary: that is, $R_{t_{\text{SM}}}(x,\theta)$ will be on the decision boundary for all $x \in Q(\theta)$. Thus, in $\mathcal{D}(\theta)$, there will be at least a $\mathbb{P}_{(x,y)\in\mathcal{D}_{XY}}[x \in Q(\theta)]$ probability mass of agents at the decision boundary, which is measure 0. This means that $\mathcal{D}(\theta)$ is not a continuous distribution. □

B.2. Proof of Proposition 2

For convenience, we break down Proposition 2 into a series of propositions, roughly corresponding to part (a), part (b), and part (c), which we prove one-by-one.

First, let's consider the case where $p = 0$. By the assumptions in Setup 1, we know that there exists a unique $\theta \in \Theta$ such that $p(\theta) = 0.5$. We call this value θ_{SL} (and it is in the interior of Θ). We claim that this is the unique locally stable point when $p = 0$.

Lemma 10. *Consider Setup 1, where a $p = 1$ fraction of agents are non-strategic. Then, θ_{SL} (defined above) is the unique locally stable point.*

Proof. Since $p = 1$, the distribution map is given by $\mathcal{D}(\theta) = \mathcal{D}_{XY}$. A locally stable point θ must be a local minimum or a stationary point of the following optimization problem:

$$\min_{\theta\in\Theta} \mathbb{E}_{(x,y)\in\mathcal{D}_{XY}}[\mathbb{1}\{f_\theta(x) = y\}] = \min_{\theta\in\Theta}\left(\mathbb{E}_{(x,y)\in\mathcal{D}_{XY}}[\mathbb{1}\{x \geq \theta\}(1 - p(x))] + \mathbb{E}_{(x,y)\in\mathcal{D}_{XY}}[\mathbb{1}\{x < \theta\}p(x)]\right).$$

Notice that the unique such θ is θ_{SL}. □

We introduce some basic properties and notation for agents who behave according to standard microfoundations. By Lemma 9, we know if an agent games when the classifier f_θ is deployed, then they will game up to boundary, which in this case, is θ. We adopt similar notation to the proof of Proposition 1, and we denote the set of who game by:

$$Q(\theta) := \left\{x \in X \mid R_{t_{\text{SM}}}(x,\theta) \neq x\right\} = \{x \in X \mid c(x,\theta) \leq 1, x < \theta\}.$$

(Technically, the agents $x \in Q(\theta)$ for whom $c(x,\theta) = 1$ are indifferent between not gaming and gaming to θ, but this is a measure 0 set by the assumption that $\mathcal{D}_{XY}$ is continuous, and the assumption that c is valid (Assumption 1)). For

$\theta \neq \min(\Theta)$, we see that for $\mathcal{D}(\theta)$, there will be a point mass at θ (from agents in $Q(\theta)$), the region $Q(\theta)$ will have zero probability density, and the rest of the distribution will remain identical to $\mathcal{D}_{XY}$.

We first characterize the set of stable points at $p = 0$. This follows a very similar argument to Lemma 3.2 in (Milli et al., 2019), but since our assumptions as well as our requirements for stability are slightly weaker, the characterization result looks slightly different. (In particular, points above the Stackelberg equilibrium can be locally stable points.)

Lemma 11. *Consider Setup 1, where a $p = 0$ fraction of agents are non-strategic. Then, there exists a locally stable point, and moreover, the set of locally stable points forms an interval $[\theta_{min}, \max(\Theta)]$, where θ_{min} is the unique value such that:*

$$\frac{\mathbb{E}_{(x,y)\in\mathcal{D}_{XY}}[p(x)I_{x\in Q(\theta_{min})}]}{\mathbb{E}_{(x,y)\in\mathcal{D}_{XY}}[I_{x\in Q(\theta_{min})}]} = 0.5.$$

(Moreover, it holds that $\theta_{min} > \theta_{SL}$, and $c(\theta_{SL}, \theta_{min}) < 1$.)

Proof. First, we show that $\theta^* = \min(\Theta)$ cannot be a stable point. Notice that $\mathcal{D}(\theta^*) = \mathcal{D}_{XY}$. Thus, θ^* is a local minimum or stationary point of $\min_{\theta\in\Theta} \mathbb{E}_{(x,y)\sim\mathcal{D}_{XY}}\, \mathbb{1}\{y \neq f_\theta(x)\}$. However, this is not possible because $p(\min(\Theta)) < 0.5$ by the assumptions in Setup 1.

Now, we consider $\theta \neq \min(\Theta)$. In this case, as discussed above, $\mathcal{D}(\theta)$ has a point mass at θ. Roughly speaking, the only property that needs to be satisfied for θ in the interior of Θ to be a local minimum of $\mathbb{E}_{(x,y)\in\mathcal{D}(\theta)}[\mathbb{1}\{f_{\theta'}(x) = y\}]$ is that it needs to be suboptimal for the decision maker to move just above the point mass (the decision maker never benefits from moving to $\theta - \epsilon$ because there is a region of zero probability density underneath θ). The loss induced from the point mass at θ is $\mathbb{E}_{(x,y)\in\mathcal{D}_{XY}}[(1 - p(x))\mathbb{1}\{x \in Q(\theta)\}]$, while if the decision-maker moves just above θ is $\mathbb{E}_{(x,y)\in\mathcal{D}_{XY}}[p(x)\mathbb{1}\{x \in Q(\theta)\}]$. The condition thus becomes $\mathbb{E}_{(x,y)\in\mathcal{D}_{XY}}[p(x)\mathbb{1}\{x \in Q(\theta)\}] \geq \mathbb{E}_{(x,y)\in\mathcal{D}_{XY}}[(1 - p(x))\mathbb{1}\{x \in Q(\theta)\}]$, which can be written as

$$\Gamma(\theta) := \frac{\mathbb{E}_{(x,y)\in\mathcal{D}_{XY}}[p(x)\mathbb{1}\{x \in Q(\theta)\}]}{\mathbb{E}_{(x,y)\in\mathcal{D}_{XY}}[\mathbb{1}\{x \in Q(\theta)\}]} \geq 0.5. \tag{6}$$

It suffices to show that the set of points where (6) is satisfied is an interval of the form $[\theta_{\min}, \max(\Theta)]$.

First, we show that the set of stable points forms an interval. It suffices to show that $\Gamma(\theta)$ is continuous and strictly increasing in θ. By the assumption on c (Assumption 1), we see that the endpoints of the interval $Q(\theta)$ are strictly increasing in θ. This, coupled with the fact that ℓ is strictly increasing in x (assumed in Setup 1), implies that $\Gamma(\theta)$ is continuous and strictly increasing as desired.

Furthermore, when $p(\theta) \leq 0.5$, the condition in (6) is never satisfied, and thus all stable points θ satisfy $\theta > \theta_{\mathrm{SL}}$, and hence $\theta_{\min} > \theta_{\mathrm{SL}}$.

Lastly, we show that this interval is not nonempty, and that $c(\theta_{\mathrm{SL}}, \theta_{\min}) \leq 1$. Consider θ such that $c(\theta_{\mathrm{SL}}, \theta) = 1$ (which we know exists by Setup 1), we see that $Q_\theta = [\theta_{\mathrm{SL}}, \theta]$. Using the conditions on ℓ, this means that condition (6) is satisfied and there is actually a strict equality. Using that c is valid, this means that $c(\theta_{\mathrm{SL}}, \theta_{\min}) < 1$. □

We now prove that no locally stable points exist for $0 < p < 1$.

Lemma 12. *Consider Setup 1, where a $0 < p < 1$ fraction of agents are non-strategic. Then, there are no locally stable points.*

Proof. When $0 < p < 1$, we show that there are no locally stable points. Assume for sake of contradiction that θ^* is a locally stable point. Recall that for θ^* to be locally stable, θ^* must either be a stationary point or a local minimum of $\min_{\theta\in\Theta} \mathbb{E}_{(x,y)\sim\mathcal{D}(\theta_{\mathrm{PS}})}\, \mathbb{1}\{y \neq f_\theta(x)\}$. We divide into three cases: (1) $\theta^* = \min(\Theta)$, (2) $\theta^* > \min(\Theta) \wedge p(\theta^*) \leq 0.5$, (3) $\theta^* > \min(\Theta) \wedge p(\theta^*) > 0.5$, and show that each results in a contradiction.

For the case (1), where $\theta^* := \min(\Theta)$, we see that $\mathcal{D}(\theta^*) = \mathcal{D}_{XY}$. Thus, θ^* is a local minimum or stationary point of $\min_{\theta\in\Theta} \mathbb{E}_{(x,y)\sim\mathcal{D}_{XY}}\, \mathbb{1}\{y \neq f_\theta(x)\}$. However, this is not possible because $p(\min(\Theta)) < 0.5$ by the assumptions in Setup 1. For the remaining two cases, we know that $\mathcal{D}(\theta^*)$ has a point mass at θ^*. This means that $\mathbb{E}_{(x,y)\in\mathcal{D}(\theta^*)}\mathbb{1}\{y \neq f_{\theta'}(x)\}$ is not differentiable at $\theta' = \theta^*$, and so θ^* must be a local minimum.

For case (2), notice that $Q(\theta^*)$ consists a nonzero density of agents for whom $p(x) < 0.5$, and for all agents $x \in Q(\theta^*)$, it holds that $p(x) \leq 0.5$. The decision maker thus wishes to move just to the other side of the point mass. (This is possible because

$\theta^* < \max(\Theta)$ based on the fact that $p(\theta^*) < 0.5$ and the assumptions in Setup 1.) In particular, $\lim_{\epsilon \to 0} \mathbb{E}_{(x,y)\in\mathcal{D}(\theta)} \mathbb{1}\{y \neq f_{\theta^*+\epsilon}(x)\} < \lim_{\epsilon \to 0} \mathbb{E}_{(x,y)\in\mathcal{D}(\theta)} \mathbb{1}\{y \neq f_{\theta^*}(x)\}$.

For case (3), notice that there exists $\epsilon > 0$ such that $p(\theta) > 0.5$ and $\theta \in Q(\theta^*)$ for all $\theta \in (\theta^* - \epsilon, \theta^*)$. The presence of non-strategic agents means that the decision-maker wishes to move to $\theta^* - \epsilon$ to achieve better performance on non-strategic agents. Since there are no strategic agents within $(\theta^* - \epsilon, \theta^*)$, this can be done without affecting the classification of strategic agents. In particular, $\mathbb{E}_{(x,y)\in\mathcal{D}(\theta)} \mathbb{1}\{y \neq f_{\theta-\epsilon}(x)\} < \mathbb{E}_{(x,y)\in\mathcal{D}(\theta)} \mathbb{1}\{y \neq f_{\theta}(x)\}$. □

Now, we prove that repeated risk minimization oscillates when $0 < p < 1$.

Lemma 13. *Consider Setup 1, where a $0 < p < 1$ fraction of agents are non-strategic. Repeated risk minimization will oscillate according to the following behavior. Let θ^1_{PS} denotes the (unique) locally performatively stable point at $p = 1$ and let θ^0_{PS} denotes the minimum locally performatively stable point at $p = 0$. RRM will oscillate between θ^1_{PS} and a threshold $f(p) > \theta^1_{\mathrm{PS}}$, where $f(p)$ is decreasing in p, approaching θ^1_{PS} as $p \to 1$ and approaching θ^0_{PS} as $p \to 0$.*

Proof. Using Lemma 10, we see that there is a unique performatively stable point for $p = 1$, given by $\theta^1_{\mathrm{PS}} := \theta_{\mathrm{SL}}$. Using Lemma 11, we see that the smallest locally stable point is given by $\theta^0_{\mathrm{PS}} := \theta_{\min}$.

In the case of $p \in (0, 1)$, the distribution map $\mathcal{D}(\theta)$ takes the form of a mixture with p weight on $\mathcal{D}_{XY}$ and with $1 - p$ weight on the distribution map of agents who behave according to standard microfoundations (which has a point mass at θ, zero density within Q_θ, and the same as the original distribution elsewhere). The main step in our proof is an analysis of the global optima of

$$B(\theta) = \operatorname{argmin}_{\theta' \in \Theta} \mathbb{E}_{(x,y)\sim\mathcal{D}(\theta)} \mathbb{1}\{y \neq f_{\theta'}(x)\}.$$

for each $\theta \in \Theta$. For convenience, we let

$$\mathrm{DPR}(\theta, \theta') := \mathbb{E}_{(x,y)\sim\mathcal{D}(\theta)} \mathbb{1}\{y \neq f_{\theta'}(x)\}.$$

We split into three cases: (a) $\theta \geq \theta_{\min}$, (b) $\theta < \theta_{\mathrm{SL}}$, and (c) $\theta_{\mathrm{SL}} \leq \theta \leq \theta_{\min}$.

Case (a): $\theta \geq \theta_{\min}$. We claim that $B(\theta) = \{\theta_{\mathrm{SL}}\}$. In this case, the proof of Lemma 11 tells us that moving just above the point mass will incur no better risk than at θ. Moreover, since $p(x) > 0.5$ for all $x \geq \theta \geq \theta_{\min}$, we see that $\mathrm{DPR}(\theta, \theta') > \mathrm{DPR}(\theta, \theta)$ for all $\theta' > \theta$. Because of the presence of non-strategic agents, a p fraction of agents will be present in Q_θ, and these agents do not change their features. Moreover, for $\theta' < \min(Q_\theta)$, $\mathcal{D}(\theta)$ looks like the base distribution. Since $p(x) > 0.5$ for $\theta_{\mathrm{SL}} < x \leq \theta_{\min}$, we see that $\mathrm{DPR}(\theta, \theta') < \mathrm{DPR}(\theta, \theta)$ for all $\theta_{\mathrm{SL}} \leq \theta' \leq \theta$. Moreover, this argument actually shows that $\theta_{\mathrm{SL}} = \operatorname{argmin}_{\theta_{\mathrm{SL}} \leq \theta' \leq \theta} \mathrm{DPR}(\theta, \theta')$. Lastly, since $p(x) < 0.5$ for $x < \theta_{\mathrm{SL}}$, we see that $\mathrm{DPR}(\theta, \theta') > \mathrm{DPR}(\theta, \theta_{\mathrm{SL}})$ for all $\theta' < \theta_{\mathrm{SL}}$.

Case (b): $\theta < \theta_{\mathrm{SL}}$. If $\theta < \theta_{\mathrm{SL}}$, then we claim that $B(\theta) = \{\theta_{\mathrm{SL}}\}$. In this case, all agents x below θ_{SL} in $\mathcal{D}(\theta)$ have $p(x) < 0.5$. Thus, $\theta_{\mathrm{SL}} = \operatorname{argmin}_{\theta \leq \theta' \leq \theta_{\mathrm{SL}}} \mathrm{DPR}(\theta, \theta')$. Moreover, above θ_{SL}, $\mathcal{D}(\theta)$ looks like the base distribution. This means that $\mathrm{DPR}(\theta, \theta') > \mathrm{DPR}(\theta, \theta_{\mathrm{SL}})$ for all $\theta' > \theta_{\mathrm{SL}}$, as desired.

Case (c): $\theta_{\mathrm{SL}} \leq \theta \leq \theta_{\min}$. Using the same argument as Case (a), we see that $\mathrm{DPR}(\theta, \theta') > \mathrm{DPR}(\theta, \theta_{\mathrm{SL}})$ for all $\theta' < \theta_{\mathrm{SL}}$. Moreover, we also see that the risk obtained by the threshold right above the point mass beats any higher threshold: that is, $\lim_{\epsilon \to 0, \epsilon \geq 0} \mathrm{DPR}(\theta, \theta + \epsilon) < \mathrm{DPR}(\theta, \theta')$ for any $\theta' > \theta$. This is because all agents $x > \theta$ have $p(x) > 0.5$.

Thus, all we need to do is to compare the threshold right above the point mass with the threshold θ_{SL}. Notice that these two classifiers behave the same on strategic agents with true features $x \notin Q_\theta$ (this is because $\theta_{\mathrm{SL}} \in Q_\theta$, because by Lemma 11, we know that $c(\theta_{\mathrm{SL}}, \theta) < c(\theta_{\mathrm{SL}}, \theta_{\min}) < 1$.). Moreover, they also behave the same on non-strategic agents not in $\theta_{\mathrm{SL}} \leq x \leq \theta$. Thus, we only need to focus on strategic agents with true features in Q_θ and non-strategic agents with

$\theta_{\text{SL}} \leq x \leq \theta$. Thus, we use the expression in the proof of Lemma 11 to see that:

$$\begin{aligned}
&\lim_{\epsilon \to 0, \epsilon > 0} \text{DPR}(\theta, \theta + \epsilon) - \text{DPR}(\theta, \theta_{\text{SL}}) \\
&= p \cdot \mathbb{E}_{(x,y) \in \mathcal{D}_{XY}}[\mathbb{1}\{x \in [\theta_{\text{SL}}, \theta]\} p(x)] + (1-p)\mathbb{E}_{(x,y) \in \mathcal{D}_{XY}}[\mathbb{1}\{x \in Q(\theta)\} p(x)] \\
&- p\mathbb{E}_{(x,y) \in \mathcal{D}_{XY}}[\mathbb{1}\{x \in [\theta_{\text{SL}}, \theta]\}(1 - p(x))] - (1-p)\mathbb{E}_{(x,y) \in \mathcal{D}_{XY}}[\mathbb{1}\{x \in Q(\theta)\}(1 - p(x))] \\
&= 2p\mathbb{E}_{(x,y) \in \mathcal{D}_{XY}}[\mathbb{1}\{x \in [\theta_{\text{SL}}, \theta]\} p(x)] + 2(1-p)\mathbb{E}_{(x,y) \in \mathcal{D}_{XY}}[\mathbb{1}\{x \in Q(\theta)\} p(x)] \\
&- p\mathbb{E}_{(x,y) \in \mathcal{D}_{XY}}[\mathbb{1}\{x \in [\theta_{\text{SL}}, \theta]\}] - (1-p)\mathbb{E}_{(x,y) \in \mathcal{D}_{XY}}[\mathbb{1}\{x \in Q(\theta)\}].
\end{aligned}$$

The relevant quantity is:

$$Z(p) := p\left(\mathbb{E}_{(x,y) \in \mathcal{D}_{XY}}[I_{x \in [\theta_{\text{SL}}, \theta]}(2p(x) - 1)]\right) + (1-p)\left(\mathbb{E}_{(x,y) \in \mathcal{D}_{XY}}[I_{x \in Q(\theta)}(2p(x) - 1)]\right)$$

Let's denote by θ^+ the parameter weights "right above the point mass" (that is, the parameter weights given by approaching θ from above θ, without ever reaching θ). We see that $B(\theta) = \{\theta_{SL}\}$ if and only if $Z(p) > 0$, $B(\theta) = \{\theta^+\}$ if and only if $Z(p) < 0$, and $B(\theta) = \{\theta^+, \theta_{\text{SL}}\}$ if and only if $Z(p) = 0$.

Now, we show that $Z(p)$ is increasing in θ. Let p_{base} be the pdf of $\mathcal{D}_{XY}$. The derivative of the first term is: $p(2p(\theta)0 - 1)p_{\text{base}}(\theta) > 0$, and the derivative of the second term is: $(1-p)(2p(\theta) - 1)p_{\text{base}}(\theta) - (1-p)(2p(\min(Q(\theta)) - 1)p_{\text{base}}(\min(Q(\theta))) > 0$, as desired.

Moreover, at $\theta = \theta_{\text{SL}}$, we see that $Z(p) < 0$; at $\theta = \theta_{\min}$, on the other hand, $Z(p) > 0$.

Thus, repeated retraining will oscillate between θ_{SL} and $f(p)$, where $f(p)$ is the value such that $Z(f(p)) = 0$. To see that $f(p)$ is decreasing in p, notice that $Z(p)$ is increasing in p for all $\theta_{\text{SL}} \leq \theta \leq \theta_{\min}$. As $p \to 0$, it is easy to see that $f(p) \to \theta_{\min}$. As $p \to 1$, it is easy to see that $f(p) \to \theta_{\text{SL}}$. □

Using the above results, we can conclude Proposition 2.

Proof of Proposition 2. When $p = 1$, we can apply Lemma 10. When $p = 0$, we can apply Lemma 11 to see that a locally stable point exists. When $0 < p < 1$, we can apply Lemma 12 to see that no locally stable point exists. For the behavior of repeated risk minimization, we can apply Lemma 13. □

B.3. Formal Statement and Proof of Proposition 3

We give a formal statement of Proposition 3 using the technology of alternative microfoundations. Let c be a valid cost function. First, we formalize expenditure monotonicity (Property 1) in the language of alternative microfoundations.

Property 4. *Let Θ be a function class of threshold functions, and let c be a cost function. A mapping $M \in \mathcal{M}$ satisfies expenditure monotonicity if $c(\mathcal{R}_t(x, \theta), x) \leq \gamma$ for every $\theta \in \Theta$ and every $t \in \text{Image}(M)$, and if $f_\theta(\mathcal{R}_t(x; \theta)) = 1$, then $f_{\theta'}(\mathcal{R}_t(x; \theta')) = 1$ for all $\theta' \leq \theta$.*

Let $\mathcal{M}^*$ be the set of maps M such that every $t \in \cup_{(x,y) \in X} \text{supp}(M(x))$ satisfies expenditure monotonicity (Property 4) and such that Assumption 3 is satisfied. Let $\mathscr{D}$ be the set of distribution maps $\mathcal{D}(\cdot; M)$ for $M \in \mathcal{M}^*$.

Proposition 14. *Consider Setup 1. Let $\mathscr{D}$ be the class of distribution maps defined above. Then:*

$$\begin{aligned}
\theta_{\text{PO}}(\mathcal{D}_{SM}) &\geq \theta_{\text{PO}}(\mathcal{D}) \\
\mathsf{Burden}(\theta_{\text{PO}}(\mathcal{D}_{SM})) &\geq \mathsf{Burden}(\theta_{\text{PO}}(\mathcal{D})).
\end{aligned}$$

where $\mathcal{D}_{SM}$ denotes the distribution map given by standard microfoundations, and $\theta_{\text{PO}}(\mathcal{D})$ denotes the minimal performatively optimal point associated with the distribution map $\mathcal{D}$.

Proof. For ease of notation, let θ_{SL} be the unique value such that $p(\theta_{\text{SL}}) = 0.5$. It is easy to see that $\theta' = \theta_{\text{PO}}(\mathcal{D}_{\text{SM}})$ is the unique point such that $c(\theta_{\text{SL}}, \theta') = 1$ and $\theta' > \theta_{\text{SL}}$.

Since $\mathsf{Burden}(\cdot)$ is monotonic in its argument, all we need to do is to show $\theta_{\text{PO}}(\mathcal{D}_{\text{SM}}) \geq \theta_{\text{PO}}(\mathcal{D})$. It suffices to show that for $\theta > \theta_{\text{PO}}(\mathcal{D}_{\text{SM}})$ and for any $\mathcal{D} \in \mathcal{D}$, it holds that $\text{PR}(\theta) \leq \text{PR}(\theta_{\text{PO}}(\mathcal{D}_{\text{SM}}))$, where the performative risk is with respect to $\mathcal{D}$.

First, let's consider the set of agents $S_1 := \{(t,x,y) \mid x < \theta_{\text{SL}}\}$. For $(t,x,y) \in S_1$, notice that $c(x,\theta) > c(x,\theta_{\text{PO}}(\mathcal{D}_{\text{SM}})) > c(\theta_{\text{SL}},\theta_{\text{PO}}(\mathcal{D}_{\text{SM}})) = c(\theta_{\text{SL}},\theta_{\text{PO}}(\mathcal{D}_{\text{RDPI}})) \geq 1$. By the expenditure constraint, this means that these agents will not game on f_θ or $f_{\theta_{\text{PO}}(\mathcal{D}_{\text{SM}})}$: i.e, $\mathcal{R}_t(x,\theta) = x$ and $\mathcal{R}_t(x,\theta_{\text{PO}}(\mathcal{D}_{\text{SM}})) = x$ for $(t,x,y) \in S_1$. Thus, $f_{\theta_{\text{PO}}(\mathcal{D}_{\text{SM}})}(\mathcal{R}_t(x,\theta_{\text{PO}}(\mathcal{D}_{\text{SM}}))) = f_\theta(\mathcal{R}_t(x,\theta)) = 0$. The performative risk with respect to $\mathcal{D}$ is thus equivalent on S_1 for f_θ and $f_{\theta_{\text{PO}}(\mathcal{D}_{\text{SM}})}$.

Now, let's consider the remaining set of agents $S_2 := \{(t,x,y) \mid x \geq \theta_{\text{SL}}\}$. Let $S_2' \subseteq S_2$ be the set

$$S_2' = \{(t,x,y) \in S_2 \mid f_\theta(\mathcal{R}_t(x,\theta)) = 1\}.$$

and

$$S_2'' = \left\{(t,x,y) \in S_2 \mid f_{\theta_{\text{PO}}(\mathcal{D}_{\text{SM}})}(\mathcal{R}_t(x,\theta_{\text{PO}}(\mathcal{D}_{\text{SM}}))) = 1\right\}.$$

We claim that $S_2' \subseteq S_2''$. This is because of the second condition in expenditure monotonicity that if x was labeled positively by f_θ, then x is also labeled positively by $f_{\theta_{\text{PO}}(\mathcal{D}_{\text{SM}})}$: in particular, we can thus conclude that $f_\theta(\mathcal{R}_t(x,\theta) = 1$ implies that $f_{\theta_{\text{PO}}(\mathcal{D}_{\text{SM}})}(\mathcal{R}_t(x,\theta_{\text{PO}}(\mathcal{D}_{\text{SM}}))) = 1$.

Now, we claim that the performative risk with respect to $\mathcal{D}$ on S_2 is no better for θ than for $\theta_{\text{PO}}(\mathcal{D}_{\text{SM}})$. This follows from the fact that $S_2' \subseteq S_2''$, and the fact that $p(x) \geq 0.5$ for $(t,x,y) \in S_2$ coupled with Assumption 3. This completes the proof. □

C. Proofs for Section 3

C.1. Microfounding any Distribution Map

With such a flexible model, *any* distribution map can be microfounded, albeit with complex response types, as long as feature manipulations do not change the fraction of positively labeled agents in the population.

Proposition 15. *Let $\mathcal{D}_{XY}$ be a non-atomic distribution. Let $\mathcal{D}(\theta)$ be any distribution map that preserves the marginal distribution over Y of $\mathcal{D}_{XY}$. Then, there exists a mapping $M \in \mathcal{M}$ such that $\mathcal{D}(\cdot;M)$ is equal to $\mathcal{D}(\cdot)$.*

This result primarily serves as an existence property that implies that our general framework for microfoundations can capture any aggregate distribution, including continuous distributions that are observed empirically (e.g. Examples 1–2).

We prove Proposition 15. The intuition is that there is a response type for every possible agent response, and it remains to show that the appropriate choice of agent response types can "shift the mass" from $\mathcal{D}_{XY}$ to $\mathcal{D}(\theta)$. In fact, M only needs to map the population to two different response types. Now, we formally prove this result.

Proof of Proposition 15. We prove Proposition 15 by construction and show that there is an M that can microfound any distribution map. We construct M as follows. We construct response types t_0 and t_1, and define $M(x,0) = t_0$ for all $x \in X$ and $M(x,1) = t_1$ for all $x \in X$. In other words, we associate agents with true label 0 with the type t_0 and agents with true label 1 with the type t_1.

In order to construct t_0 and t_1, we define the following probability measures over the measure space $X \subseteq \mathbb{R}^D$ equipped with the Borel sigma-algebra. We consider $\mu^0(\theta)$ to be the probability measure given by the distribution over x when $(x,y) \in \mathcal{D}(\theta)$ and $y = 0$. We define $\mu^1(\theta)$ similarly. We let μ^0_{XY} be the probability measure given by the distribution x where $(x,y) \in \mathcal{D}_{XY}$ and $y = 0$, and we define μ^1_{XY} analogously.

First, we claim that it suffices to prove that for each $\theta \in \Theta$ there is a measurable map $f_{0,\theta} : X \to X$ that maps the probability measure μ^0_{XY} to $\mu^0(\theta)$, and a measurable map $f_{1,\theta} : X \to X$ that maps μ^1_{XY} to $\mu^1(\theta)$. In this case, we can define t_0 to be given by $\mathcal{R}_{t_0}(x,\theta) = f_{0,\theta}(x)$ and t_1 to be given by $\mathcal{R}_{t_1}(x,\theta) = f_{1,\theta}(x)$. Let's now consider the distribution given by $(\mathcal{R}_t(x,\theta),y)$ where $(t,x,y) \sim \mathcal{D}_{TXY}$. The condition distribution over $y = 0$ is given by $\mu^0(\theta)$ and the conditional distribution over $y = 1$ is given by $\mu^1(\theta)$, which means that the distribution over all is given by $\mathcal{D}(\theta)$, as desired. Moreover, the measurability requirements on $f_{0,\theta}$ and $f_{1,\theta}$ guarantee that Assumption 2 is satisfied.

Thus, it suffices to construct $f_{0,\theta}$ and $f_{1,\theta}$ for $\theta \in \Theta$ that satisfy the above conditions. To do this, we make use of Proposition 3 in (Gatzouras, 2002), which says that there exists a Borel mapping from any tight non-atomic measure to any other probability measure. Since the probability measure associated to $\mathcal{D}_{XY}$ is non-atomic, we see that μ^0_{XY} and μ^1_{XY} are non-atomic as desired, and so a Borel mapping from μ^0_{XY} to $\mu^0(\theta)$ exists and a Borel mapping from μ^1_{XY} to $\mu^1(\theta)$ exists. □

C.2. Connection between Aggregate Smoothness and Continuity for 1-d Example

We formalize the connection between aggregate smoothness and the continuity of the distribution map. In particular, the existence of the partial derivative of $\mathsf{DPR}_M(\theta, \theta')$ with respect to θ' guarantees that *each distribution $\mathcal{D}(\theta; M)$ is sufficiently continuous (and cannot have a point mass at the decision boundary)*, and assuming continuity of the derivative we guarantee that $\mathcal{D}(\theta; M)$ *changes continuously in* θ. This connection between aggregate smoothness and continuity of the distribution map can be made explicit in the case of 1-dimensional features:

Proposition 16. *Suppose that $X \subseteq \mathbb{R}$, and let $\Theta \subseteq \mathbb{R}$ be a function class of threshold functions. Then, if the distribution map $\mathcal{D}(\cdot; M)$ has the following properties, the mapping M satisfies aggregate smoothness w.r.t. Θ:*

1. *For each θ, the probability density $p_\theta(x, y)$ of $\mathcal{D}(\theta; M)$ exists everywhere and is continuous in x.*
2. *For each x, y, the probability density $p_\theta(x, y)$ is continuous in θ.*

Proof of Proposition 16. To prove Proposition 16 we show that $\mathsf{dPR}_\theta(\theta')$ is continuous in θ and θ'. We see that

$$\mathsf{DPR}(\theta, \theta') = \int_{x' \geq \theta'} p_\theta((x', 0))dx' + \int_{x' < \theta'} p_\theta((x', 1))dx'.$$

Let's take a derivative with respect to θ' to obtain:

$$\mathsf{dPR}_\theta(\theta') = -p_\theta((\theta', 0)) + p_\theta((\theta', 1)).$$

The first continuity requirement tells us that this is continuous in θ', and the second continuity requirement tells us that this continuous in θ. □

C.3. Proof of Theorem 5

We first recall the definition of the decoupled performative risk (Perdomo et al., 2020):

$$\mathsf{DPR}(\theta, \theta') := \mathbb{E}_{(x,y)\in\mathcal{D}(\theta)}\left[\mathbb{1}\left(f_{\theta'}(x) \neq y\right)\right].$$

The gradient of the decoupled performative risk plays an important role in our analysis of locally stable points. In order to take derivatives at the boundary, we consider an open set $\Theta' \supset \Theta$ that is also bounded and convex, and assume there are classifiers associated with each $\theta \in \Theta'$, although the decision maker only considers classifier weights in Θ. We use the notation:

$$\mathsf{dPR}_\theta(\theta') := \nabla_{\theta'}\mathsf{DPR}(\theta, \theta') = \nabla_{\theta'}\mathbb{E}_{(x,y)\sim\mathcal{D}(\theta)}[\,\mathbb{1}\{y \neq f_{\theta'}(x)\}]$$

to denote the gradient of the decoupled performative risk with respect to the second argument. To prove Theorem 5 we show that the continuity of the derivatives of the decoupled performative risk guarantees the existence of stable points under mixtures with non-strategic agents.

Proof of Theorem 5. Our main technical ingredient is this proof is applying Brouwer's fixed point theorem on $G_{gd}(\theta) = \mathrm{Proj}_\Theta(\theta + \eta\,\mathsf{dPR}_\theta(\theta))$. It thus suffices to show that the map $\theta \mapsto \mathrm{Proj}_\Theta(\theta + \eta\,\mathsf{dPR}_\theta(\theta))$ is continuous.

First, we show that aggregate risk smoothness implies that $\theta \mapsto \mathrm{Proj}_\Theta(\theta + \eta\,\mathsf{dPR}_\theta(\theta))$ is a continuous map. By aggregate risk smoothness, we know that $\mathsf{dPR}_\theta(\theta)$ exists for all $\theta \in \Theta$. Moreover, for any $\theta \in \Theta$, aggregate risk smoothness tells us that:

$$\lim_{\theta' \to \theta} \left\|\mathsf{dPR}_\theta(\theta) - \mathsf{dPR}_{\theta'}(\theta')\right\| \leq \lim_{\theta' \to \theta} \|\mathsf{dPR}_\theta(\theta) - \mathsf{dPR}_{\theta'}(\theta)\| + \lim_{\theta' \to \theta} \left\|\mathsf{dPR}_{\theta'}(\theta) - \mathsf{dPR}_{\theta'}(\theta')\right\|.$$

Thus, $\mathsf{dPR}_\theta(\theta)$ is continuous in θ. Moreover, since the sum of continuous functions is continuous, this means that $\theta \mapsto \theta + \eta\,\mathsf{dPR}_\theta(\theta)$ is continuous. Now, since projection onto a convex set is a contraction map, we can conclude that $\theta \mapsto \mathrm{Proj}_\Theta(\theta + \eta\,\mathsf{dPR}_\theta(\theta))$ is continuous as desired.

□

C.4. Proof of Proposition 6

We now prove Proposition 6.

Proof of Proposition 6. Since derivatives are linear, we can break $\mathsf{dPR}_\theta(\theta)$ into a term for non-strategic agents and a term for strategic agents. Since the sum of two continuous function is continuous, it suffices to show that $\mathsf{dPR}_\theta(\theta)$ exists and is continuous for non-strategic agents and for strategic agents. For strategic agents, this follows from aggregate risk smoothness. For non-strategic agents, since the (non-performative) risk $R(\theta) := \mathbb{E}_{(x,y)\in\mathcal{D}_{XY}} \mathbb{1}\{f_\theta(x) = y\}$ is differentiable in θ and $\mathsf{dPR}_\theta(\theta) = \nabla_\theta R(\theta)$ is continuous in θ as desired. □

D. Proofs for Section 4

D.1. Proof of Proposition 7

Proof of Proposition 7. We use the following notation for this proof. Let's extend the cost function to be defined and valid on all of $\mathbb{R}$ rather than just X. For $x \in X$, let's use the notation $l_x \in \mathbb{R}$ to denote the unique value such that $l_x < x$ and $c(l_x, x) = 1$. Similarly, let $u_x \in \mathbb{R}$ denote the unique value such that $u_x > x$ and $c(x, u_x) = 1$. These values are unique by the definition of a valid cost function.

Fix $\sigma \in (0, \infty)$, and $x' \in X$. Let's characterize the agents who will change their features to x' when the threshold is θ. Either the agents' true features are equal to x' and their perception function $P(\theta) \notin (x', u_{x'}]$, or the agents' perception function $P(\theta) = x'$ and their true features x are in $[l_{x'}, x']$. Since the base distribution and the noise distribution are continuous, this means that there are no point masses in the distribution. To see that a probability density function exists everywhere and is continuous, let's compute the density. Let p_{base} denote the pdf of the base distribution (which is assumed to exist and be continuous since $\mathcal{D}_{XY}$ is a continuous distribution), and let p_{noise} denote the pdf of D (which is continuous since it is the pdf of a gaussian). Notice that the probability density of $\mathcal{D}(\theta)$ at (x', y') is

$$p_{\text{base}}((x', y')) \cdot \mathbb{P}_D[\eta \notin (x' - \theta, u_{x'} - \theta)] + p_{\text{noise}}(x' - \theta) \cdot \mathbb{P}_{\mathcal{D}_{XY}}[x \in [l_{x'}, x'], y = y'].$$

This is continuous in x' because $u_{x'}$ and $l_{x'}$ are continuous in x'. Moreover, this is nonzero on all x' because for all $x' \in X$, we see that $p_{\text{base}}((x', y')) > 0$ and $\mathbb{P}_D[\eta \notin (x' - \theta, u_{x'} - \theta)] > 0$ as well.

Now, we show aggregate smoothness. We see that the probability density $p_\theta((x', y'))$ at (x', y') is continuous in x' because each term is continuous in x'. Similarly, we see that this is continuous in θ because each term is continuous in θ. By Proposition 16, this implies aggregate smoothness. □

D.2. Social burden of noisy responses in general

We show that for any valid cost function, noisy responses results in an optimal point with no higher social burden than the optimal point deduced from standard microfoundations.

Proposition 17. *Let $\sigma \in (0, \infty)$, and let c be a valid cost function. Consider a 1-dimensional setting where $X \subseteq \mathbb{R}$ and Θ is a function class of threshold functions. Then, the following holds:*

$$\theta_{\text{PO}}(M_{SM}) \geq \theta_{\text{PO}}(M_\sigma)$$
$$\mathsf{Burden}(\theta_{\text{PO}}(M_{SM})) \geq \mathsf{Burden}(\theta_{\text{PO}}(M_\sigma)),$$

where M_{SM} is the mapping induced by standard microfoundations.

Proof. By Proposition 14, it suffices to show that M_σ satisfies expenditure monotonicity and Assumption 3. The fact that M_σ satisfies Assumption 3 follows from its definition. For expenditure monotonicity, note that the first condition follows from the fact that the optimization problem in (4) tells us that fuzzy perception agents never exceed their utility of a positive outcome from manipulation expenditure. We now show that the second condition is satisfied. Note that each agents' perception function takes the form $P(\theta) = \theta + \eta$ for some fixed η. Thus, any given agent either consistently overshoots or consistently undershoots the threshold. If $\eta < 0$, then the agent will only be positively classified if and only if $\theta \leq x$ where x are the agent's true features. If $\eta > 0$, then the agent will be positively classified if and only if $c(x, \theta + \eta) \leq 1$ or $\theta \leq x$. This proves the desired statement. □

D.3. Proof of Proposition 8

Proof of Proposition 8. By Proposition 17, we see that $\theta_{\mathrm{PO}}(\mathcal{D}_{\mathrm{SM}}) \geq \theta_{\mathrm{PO}}(\mathcal{D})$. It thus suffices to show that $\theta_{\mathrm{PO}}(\mathcal{D}_{\mathrm{SM}}) > \theta_{\mathrm{PO}}(\mathcal{D})$. To show this, it suffices to show that the derivative of the performative risk exists and is nonzero at $\theta_{\mathrm{PO}}(\mathcal{D}_{\mathrm{SM}})$.

Like in the proof of Theorem 7, we use the notation l_x, u_x, p_{base}, and $\mathbb{P}_D$. By the expenditure constraint, we know that agents with true features $x \leq l_\theta$ will all be classified as 0, and so their net contribution to the performative risk is $\int_{-\infty}^{l_\theta} p_{\mathrm{base}}((x,1))dx$. By the properties of noisy response, we know that agents with true features $x \geq \theta$ will be classified as 1, so their net contribution to the performative risk is $\int_{-\infty}^{l_\theta} p_{\mathrm{base}}((x,1))dx$. For agents with true features $x \in (\theta-1,\theta)$, agents will be classified as 1 if and only if they manipulate features to $x' \geq x$ if and only if $\eta \in [0, u_x - \theta]$. Putting this all together, we see that the performative risk is thus equal to:

$$
\begin{aligned}
\mathrm{PR}(\theta) &= \int_{\theta}^{\infty} p_{\mathrm{base}}((x,0))dx + \int_{-\infty}^{\theta-1} p_{\mathrm{base}}((x,1))dx + \int_{\theta-1}^{\theta} p_{\mathrm{base}}((x',0))\mathbb{P}_D[\eta \in [0, u_{x'} - \theta]]dx' \\
&+ \int_{\theta-1}^{\theta} p_{\mathrm{base}}((x',1))\mathbb{P}_D[\eta \notin [0, u_{x'} - \theta]]dx' \\
&= \int_{\theta}^{\infty} p_{\mathrm{base}}((x,0))dx + \int_{-\infty}^{\theta-1} p_{\mathrm{base}}((x,1))dx + \int_{\theta-1}^{\theta} p_{\mathrm{base}}((x',0))\mathbb{P}_D[\eta \in [0, x' + 1 - \theta]]dx' \\
&+ \int_{\theta-1}^{\theta} p_{\mathrm{base}}((x',1))\mathbb{P}_D[\eta \notin [0, x' + 1 - \theta]]dx' \\
&= \int_{\theta}^{\infty} p_{\mathrm{base}}((x,0))dx + \int_{-\infty}^{\theta-1} p_{\mathrm{base}}((x,1))dx + \int_{0}^{1} p_{\mathrm{base}}((\theta - 1 + x,0))\mathbb{P}_D[\eta \in [0,x]]dx \\
&+ \int_{0}^{1} p_{\mathrm{base}}((\theta - 1 + x,1))\mathbb{P}_D[\eta \notin [0,x]]dx \\
&= \int_{\theta}^{\infty} p_{\mathrm{base}}((x,0))dx + \int_{-\infty}^{\theta-1} p_{\mathrm{base}}((x,1))dx + \mathbb{P}_{\mathcal{D}_{XY}}[x \in (\theta-1,\theta), y = 0] \\
&+ \int_{0}^{1} (p_{\mathrm{base}}((\theta - 1 + x,1)) - p_{\mathrm{base}}((\theta - 1 + x,0)))\mathbb{P}_D[\eta \notin [0,x]]dx.
\end{aligned}
$$

Let's write $\int_0^1 (p_{\mathrm{base}}((\theta-1+x,1)) - p_{\mathrm{base}}((\theta-1+x,0)))\mathbb{P}_D[\eta \notin [0,x]]dx$ in a slightly different form.

$$
\begin{aligned}
&\int_{0}^{1} (p_{\mathrm{base}}((\theta - 1 + x,1)) - p_{\mathrm{base}}((\theta - 1 + x,0)))\mathbb{P}_D[\eta \notin [0,x]]dx \\
&= (\mathbb{P}_D[\eta \in [-\infty, 0]] + \mathbb{P}_D[\eta \in [1,\infty]]) \int_{0}^{1} (p_{\mathrm{base}}((\theta - 1 + x,1)) - p_{\mathrm{base}}((\theta - 1 + x,0)))dx \\
&+ \int_{0}^{1} (p_{\mathrm{base}}((\theta - 1 + x,1)) - p_{\mathrm{base}}((\theta - 1 + x,0)))\mathbb{P}_D[\eta \in [x,1]]dx \\
&= (\mathbb{P}_D[\eta \in [-\infty, 0]] + \mathbb{P}_D[\eta \in [1,\infty]])(\mathbb{P}_{\mathcal{D}_{XY}}[x \in (\theta-1,\theta), y = 1] - \mathbb{P}_{\mathcal{D}_{XY}}[x \in (\theta-1,\theta), y = 0]) \\
&+ \int_{0}^{1} (p_{\mathrm{base}}((\theta - 1 + x,1)) - p_{\mathrm{base}}((\theta - 1 + x,0)))\mathbb{P}_D[\eta \in [x,1]]dx.
\end{aligned}
$$

We can rewrite:

$$\begin{aligned}
&\int_0^1 (p_{\text{base}}((\theta-1+x,1)) - p_{\text{base}}((\theta-1+x,0)))\mathbb{P}_D[\eta \in [x,1]]dx \\
&= \int_0^1 \int_x^1 (p_{\text{base}}((\theta-1+x,1)) - p_{\text{base}}((\theta-1+x,0)))p_{\text{noise}}(z)dzdx \\
&= \int_0^1 p_{\text{noise}}(z) \int_0^z (p_{\text{base}}((\theta-1+x,1)) - p_{\text{base}}((\theta-1+x,0)))dxdz \\
&= \int_0^1 p_{\text{noise}}(z)(\mathbb{P}_{\mathcal{D}_{XY}}[x \in ((\theta-1,\theta-1+z)), y=1] - \mathbb{P}_{\mathcal{D}_{XY}}[x \in ((\theta-1,\theta-1+z))), y=0])dz
\end{aligned}$$

When we take a derivative with respect to θ, we obtain:

$$\begin{aligned}
\frac{\partial \text{PR}(\theta)}{\partial \theta} &= -p_{\text{base}}((\theta,0)) + p_{\text{base}}((\theta-1,1)) - p_{\text{base}}((\theta-1,0)) + p_{\text{base}}((\theta,0)) \\
&+ (\mathbb{P}_D[\eta \in [-\infty,0]] + \mathbb{P}_D[\eta \in [1,\infty]])(p_{\text{base}}((\theta,1)) - p_{\text{base}}((\theta-1,1)) - p_{\text{base}}((\theta,0)) + p_{\text{base}}((\theta-1,0))) \\
&+ \int_0^1 p_{\text{noise}}(z)(p_{\text{base}}((\theta-1+z,1)) - p_{\text{base}}((\theta-1,1)) - p_{\text{base}}((\theta-1+z,0)) + p_{\text{base}}((\theta-1,0)))dz.
\end{aligned}$$

Let's analyze this expression at $\theta = \theta_{\text{PO}}(\mathcal{D}_{\text{SM}})$. By the assumptions on the cost function, and using that $\theta_{\text{SL}} + 1 \in \Theta \cap X$, we see that $\theta_{\text{PO}}(\mathcal{D}_{\text{SM}}) = \theta_{\text{SL}} + 1$, so $\theta - 1 = \theta_{\text{SL}}$. This means that $p_{\text{base}}((\theta-1,1)) - p_{\text{base}}((\theta-1,0)) = p_{\text{base}}((\theta_{\text{SL}},1)) - p_{\text{base}}((\theta_{\text{SL}},0)) = 0$. Thus, the expression simplifies to:

$$\begin{aligned}
\frac{\partial \text{PR}(\theta)}{\partial \theta} &= (\mathbb{P}_D[\eta \in [-\infty,0]] + \mathbb{P}_D[\eta \in [1,\infty]])(p_{\text{base}}((\theta,1)) - p_{\text{base}}((\theta,0))) \\
&+ \int_0^1 p_{\text{noise}}(z)(p_{\text{base}}((\theta-1+z,1)) - p_{\text{base}}((\theta-1+z,0)))dz.
\end{aligned}$$

We see that $p_{\text{base}}((\theta',1)) > p_{\text{base}}((\theta',0))$ for all $\theta' \geq \theta_{\text{SL}}$ by the assumption on μ in Setup 1. This implies that the first term is positive and the second term is nonnegative, so $\frac{\partial \text{PR}(\theta)}{\partial \theta}$ is positive as desired. □

E. Reducing the complexity of estimating the distribution map

Apart from defining a natural class of feasible microfoundations models, an additional advantage of Property 3 is that it naturally constrains each agent's range of manipulations. This can significantly reduce the complexity of estimating the distribution map for a decision-maker who wants to compute a strategy robust classifier offline.

Assume the decision maker follows a two-stage estimation procedure to estimate a performatively optimal point, similar to (Miller et al., 2021). First, they compute an estimate $\tilde{M}$ of the true mapping M and infer $\mathcal{D}(\cdot;\tilde{M})$ from the base distribution $\mathcal{D}_{XY}$. Second, they assume the model reflects the true decision dynamics and approximate optimal points as follows:

$$\theta_{\text{PO}}(\tilde{M}) := \text{argmin}_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim\mathcal{D}(\theta;\tilde{M})}[\mathbb{1}\{y \neq f_\theta(x)\}]. \tag{7}$$

The naive approach is to compute an estimate $\tilde{M}$ of M, such that $\sup_\theta \text{TV}(\mathcal{D}(\theta;\tilde{M}), \mathcal{D}(\theta;M)) \leq \xi$. This guarantee that $\text{PR}(\theta_{\text{PO}}(M)) - \text{PR}(\theta_{\text{PO}}(\tilde{M})) \leq 2\xi$.

Lemma 18. *Let $\tilde{M}$ be an estimate of the true distribution map M. Then the suboptimality of the performative risk of $\theta_{\text{PO}}(M)$ as per (7) is bounded by: $\text{PR}(\theta_{\text{PO}}(\tilde{M})) - \text{PR}(\theta_{\text{PO}}(M)) \leq 2\sup_\theta \left\{\text{TV}\left(\mathcal{D}(\theta;M), \mathcal{D}(\theta;\tilde{M})\right)\right\}$, where $\text{PR}(\theta) := \mathbb{E}_{(x,y)\sim\mathcal{D}(\theta)}[\mathbb{1}\{y \neq f_\theta(x)\}]$ denotes the performative risk with respect to M.*

Proof of Lemma 18. Let $\xi = \left\{\text{TV}\left(\mathcal{D}(\theta;M), \mathcal{D}(\theta;\tilde{M})\right)\right\}$. Let $\text{PR}(\theta;M)$ denote the performative risk at θ on $\mathcal{D}(\theta;M)$ and let $\text{PR}(\theta;\tilde{M})$ denote the performative risk at θ on $\mathcal{D}(\theta;\tilde{M})$. It suffices to show that $|\text{PR}(\theta;M) - \text{PR}(\theta;\tilde{M})| \leq \xi$ (since

this would mean that $\mathrm{PR}(\theta_{\mathrm{PO}}(\tilde{M}); M) \leq \mathrm{PR}(\theta_{\mathrm{PO}}(\tilde{M}); \tilde{M}) + \xi \leq \mathrm{PR}(\theta_{\mathrm{PO}}(M); \tilde{M}) + \xi \leq \mathrm{PR}(\theta_{\mathrm{PO}}(M); M) + 2\xi$, as desired). Notice that:

$$|\mathrm{PR}(\theta; M) - \mathrm{PR}(\theta; \tilde{M})| = \left|\mathbb{E}_{(x,y)\sim\mathcal{D}(\theta;M)}[\,\mathbb{1}\{y \neq f_\theta(x)\}] - \mathbb{E}_{(t,x,y)\sim\mathcal{D}(\theta;\tilde{M})}[\,\mathbb{1}\{y \neq f_\theta(x)\}]\right|.$$

Since the indicator variables are always constrained between 0 and 1, we can immediately obtain an upper bound of $TV(\mathcal{D}(\theta; M), \mathcal{D}(\theta; \tilde{M}))$. □

However, achieving a sufficient level of accuracy for the distribution map in terms of TV distance fundamentally requires a full specification of the response types for every agent in the population.

The expenditure constraint helps to make this task more tractable, in that the decision-maker only needs to estimate responses for a small fraction of the agents to achieve the same bound on the suboptimality of the obtained performative risk. To formalize this, let's assume the decision maker can define a set $\Theta_0 \subseteq \Theta$ that contains the performatively optimal classifier $\theta_{\mathrm{PO}}(M)$. Then, given the implied restriction in the search space in (7), the expenditure constraint enables us to restrict the set of covariates that are relevant for the optimization problem to

$$S(\Theta_0, c) := \cup_{\theta\in\Theta_0}\{x \in X : \exists x' \in X : f_\theta(x') \neq f_\theta(x) \wedge c(x, x') \leq \gamma\}. \tag{8}$$

The salient part $S(\Theta_0, c) \subseteq X$ captures all agents who are sufficiently close to the decision boundary for some $\theta \in \Theta_0$ so they are able to cross it without expending more than γ units of cost. The subset $S(\Theta_0, c)$ can be entirely specified by the cost function c and can be much smaller than X.

We now describe the implications of constraining to the salient part for a 1-dimensional setting where $X \subseteq \mathbb{R}$ and f_θ is a threshold function.[9] Let us define an *agent response oracle* that given x and θ, outputs a draw x' from the response distribution $(\mathcal{R}_t(x, \theta), y)$ where $(x, y) \sim \mathcal{D}_{XY}$. We show with few calls to the oracle, the decision-maker can build an sufficiently precise estimate of M.

Proposition 19. *Let $X \subseteq \mathbb{R}$, let $\Theta \subseteq \mathbb{R}$ be the function class of threshold functions. Suppose that M satisfies the expenditure constraint, the distribution map $\mathcal{D}(\cdot; M)$ is 1-Lipschitz with respect to TV distance, and $\Theta_0 \subseteq \Theta : \theta_{\mathrm{PO}}(M) \in \Theta_0$. We further assume that an agent's type does not depend on their label, i.e., $M(x, 0) = M(x, 1)$ for all $x \in X$. Then, with $O\left(\zeta^2 \frac{\ln(1/\epsilon)}{2\epsilon^3}\right)$ calls to the agent response oracle, where $\zeta := \mathbb{P}_{\mathcal{D}_{XY}}[x \in S(\Theta_0, c)]$, the decision maker can create an estimate $\tilde{M}$ so that:*

$$\mathrm{PR}(\theta_{\mathrm{PO}}(\tilde{M})) \leq \mathrm{PR}(\theta_{\mathrm{PO}}(M)) + \epsilon.$$

with probability 0.9.

The number of necessary calls to the response function oracle for estimating M decays with $\zeta := \mathbb{P}_{\mathcal{D}_{XY}}[x \in S(\Theta_0, c)]$. Without any assumption on agent responses we have $S(\Theta_0, c) = X$ and the value of ζ is equal to 1. However, when the decision-maker is able to constrain $S(\Theta_0, c)$ to a small part of the input space by relying on the expenditure constraint, domain knowledge, or stronger assumptions on agent behavior, ζ and thus the number of oracle calls can be reduced significantly.

The concept of a salient part bears resemblance to the approaches by Zhang and Conitzer (2021); Zhang et al. (2021), which directly specify the set of feature changes that an agent may make, rather than implicitly specifying agent actions through a cost function. While these models assume that agents best-respond, our key finding is that constraining agent behavior alone can lessen the empirical burden on the decision-maker.

For the remainder of the section, we prove Proposition 19.

E.1. Proof of Proposition 19

To prove Proposition 19, first we show a bound on the performative risk in terms of the Kolmogrov-Sminorff (KS) distance between the true distribution map and estimated distribution map. To state this bound, we introduce the following notation. We use a subscript notation $\mathcal{D}_{S(\Theta_0,c)}(\theta; M)$ to denote the aggregate response distribution $\mathcal{D}(\theta; M)$ restricted to agents with true features $x \in S(\Theta_0, c)$, where $S(\Theta_0, c)$ is defined as in (8). Let $\mathcal{D}^0_{S(\Theta_0,c)}(\theta; M)$ be the marginal distribution over x of the conditional distribution of $(x, y) \sim \mathcal{D}_{S(\Theta_0,c)}(\theta; M)$ conditional on $y = 0$. We define $\mathcal{D}^1_{S(\Theta_0,c)}(\theta; M)$, $\mathcal{D}^0_{S(\Theta_0,c)}(\theta; \tilde{M})$, and $\mathcal{D}^1_{S(\Theta_0,c)}(\theta; \tilde{M})$ analogously.

[9] Proposition 19 directly extends to *posterior threshold functions* (Milli et al., 2019).

Lemma 20. *Let Θ be a function class of posterior threshold functions, and c be an outcome-valid cost function. Suppose that $M, \tilde{M}$ restricted to the domain $(X \setminus S(\Theta_0, c)) \times Y$ are expenditure-constrained. Then, for any $\Theta_0 \subseteq \Theta : \theta_{\mathrm{PO}}(M) \in \Theta_0$, the predicted performative optima $\theta_{\mathrm{PO}}(\tilde{M})$ satisfies:*

$$\mathrm{PR}_M(\theta_{\mathrm{PO}}(\tilde{M}) \leq \mathrm{PR}_M(\theta_{\mathrm{PO}}(M))) + 2\xi$$

where ξ is defined to be

$$\sup_{\theta}(A_\theta + B_\theta)$$

where

$$A(\theta) := \mathbb{P}[x \in S(\Theta_0, c) \ \ \& \ \ y = 0]\mathrm{KS}\left(\mathcal{D}^0_{S(\Theta_0,c)}(\theta; M), \mathcal{D}^0_{S(\Theta_0,c)}(\theta; \tilde{M})\right)$$
$$B(\theta) := \mathbb{P}[x \in S(\Theta_0, c) \ \ \& \ \ y = 1]\mathrm{KS}\left(\mathcal{D}^1_{S(\Theta_0,c)}(\theta; M), \mathcal{D}^1_{S(\Theta_0,c)}(\theta; \tilde{M})\right).$$

Proof. Let $\mathrm{PR}(\theta; M)$ denote the performative risk at θ on $\mathcal{D}(\theta; M)$ and let $\mathrm{PR}(\theta; \tilde{M})$ denote the performative risk at θ on $\mathcal{D}(\theta; \tilde{M})$. It suffices to show that $|\mathrm{PR}(\theta; M) - \mathrm{PR}(\theta; \tilde{M})| \leq \xi$ for all $\theta \in \Theta_0$ (since this would mean that $\mathrm{PR}(\theta_{\mathrm{PO}}(\tilde{M}); M) \leq \mathrm{PR}(\theta_{\mathrm{PO}}(\tilde{M}); \tilde{M}) + \xi \leq \mathrm{PR}(\theta_{\mathrm{PO}}(M); \tilde{M}) + \xi \leq \mathrm{PR}(\theta_{\mathrm{PO}}(M); M) + 2\xi$, as desired). Notice that:

$$|\mathrm{PR}(\theta; M) - \mathrm{PR}(\theta; \tilde{M})| = \left|\mathbb{E}_{(x,y)\sim\mathcal{D}(\theta;M)}[\ \mathbb{1}\{y \neq f_\theta(x)\}] - \mathbb{E}_{(x,y)\sim\mathcal{D}(\theta;\tilde{M})}[\ \mathbb{1}\{y \neq f_\theta(x)\}]\right|.$$

Let's let $\mathcal{D}_{TXY}$ be the distribution of (t, x, y) where $(x, y) \sim \mathcal{D}_{XY}$ and $t \sim M(x, y)$. Similarly, let $\tilde{\mathcal{D}}_{TXY}$ be the distribution of (t, x, y) where $(x, y) \sim \mathcal{D}_{XY}$ and $t \sim \tilde{M}(x, y)$. Notice that:

$$|\mathrm{PR}(\theta; M) - \mathrm{PR}(\theta; \tilde{M})| = \left|\mathbb{E}_{(t,x,y)\sim\mathcal{D}_{TXY}}[\ \mathbb{1}\{y \neq f_\theta(\mathcal{R}_t(x, \theta))\}] - \mathbb{E}_{(t,x,y)\sim\tilde{\mathcal{D}}_{TXY}}[\ \mathbb{1}\{y \neq f_\theta(\mathcal{R}_t(x, \theta))\}]\right|.$$

Now, we claim that for any agent (t, x) where $x \notin S(\Theta_0, c)$ and for $t \in \mathrm{supp}(\mathcal{D}_{TXY}) \cup \mathrm{supp}(\tilde{\mathcal{D}}_{TXY})$, it holds that $f_\theta(\mathcal{R}_t(x, \theta)) = f_\theta(x)$ for every $\theta \in \Theta_0$. Note that since M satisfies the expenditure constraint with respect to c, then we know that if $x \notin S_\theta$, it holds that $f_\theta(\mathcal{R}_t(x, \theta)) = f_\theta(x)$. Moreover, note that since $S_\theta \subseteq S(\Theta_0, c)$ by definition, this yields the desired statement. Thus we have that:

$$\begin{aligned}
&\left|\mathbb{E}_{(t,x,y)\sim\mathcal{D}_{TXY}}[\ \mathbb{1}\{y \neq f_\theta(\mathcal{R}_t(x, \theta))\}] - \mathbb{E}_{(t,x,y)\sim\tilde{\mathcal{D}}_{TXY}}[\ \mathbb{1}\{y \neq f_\theta(\mathcal{R}_t(x, \theta))\}]\right| \\
&\leq \left|\mathbb{E}_{(t,x,y)\sim\mathcal{D}_{TXY}}[\ \mathbb{1}\{y \neq f_\theta(\mathcal{R}_t(x, \theta))\}\ \mathbb{1}\{x \notin S(\Theta_0, c)\}] - \mathbb{E}_{(t,x,y)\sim\tilde{\mathcal{D}}_{TXY}}[\ \mathbb{1}\{y \neq f_\theta(\mathcal{R}_t(x, \theta))\}\ \mathbb{1}\{x \notin S(\Theta_0, c)\}]\right| \\
&+ \left|\mathbb{E}_{(t,x,y)\sim\mathcal{D}_{TXY}}[\ \mathbb{1}\{y \neq f_\theta(\mathcal{R}_t(x, \theta))\}\ \mathbb{1}\{x \in S(\Theta_0, c)\}] - \mathbb{E}_{(t,x,y)\sim\tilde{\mathcal{D}}_{TXY}}[\ \mathbb{1}\{y \neq f_\theta(\mathcal{R}_t(x, \theta))\}\ \mathbb{1}\{x \in S(\Theta_0, c)\}]\right| \\
&= \left|\mathbb{E}_{(t,x,y)\sim\mathcal{D}_{TXY}}[\ \mathbb{1}\{y \neq f_\theta(\mathcal{R}_t(x, \theta))\}\ \mathbb{1}\{x \in S(\Theta_0, c)\}] - \mathbb{E}_{(t,x,y)\sim\tilde{\mathcal{D}}_{TXY}}[\ \mathbb{1}\{y \neq f_\theta(\mathcal{R}_t(x, \theta))\}\ \mathbb{1}\{x \in S(\Theta_0, c)\}]\right| \\
&= \left|\mathbb{E}_{(x,y)\sim\mathcal{D}(\theta)}[\ \mathbb{1}\{y \neq f_\theta(x)\}\ \mathbb{1}\{x \in S(\Theta_0, c)\}] - \mathbb{E}_{(x,y)\sim\tilde{\mathcal{D}}(\theta)}[\ \mathbb{1}\{y \neq f_\theta(x)\}\ \mathbb{1}\{x \in S(\Theta_0, c)\}]\right|.
\end{aligned}$$

We can break this into terms where $y = 0$ and terms where $y = 1$. Thus, it suffices to bound:

$$\left|\mathbb{E}_{(x,y)\sim\mathcal{D}(\theta;M)}[\ \mathbb{1}\{y \neq f_\theta(x)\}\ \mathbb{1}\{x \in S(\Theta_0, c)\}\ \mathbb{1}\{y = 0\}] - \mathbb{E}_{(x,y)\sim\mathcal{D}(\theta;\tilde{M})}[\ \mathbb{1}\{y \neq f_\theta(x)\}\ \mathbb{1}\{x \in S(\Theta_0, c)\}\ \mathbb{1}\{y = 0\}]\right|$$

and

$$\left|\mathbb{E}_{(x,y)\sim\mathcal{D}(\theta;M)}[\ \mathbb{1}\{y \neq f_\theta(x)\}\ \mathbb{1}\{x \in S(\Theta_0, c)\}\ \mathbb{1}\{y = 1\}]\mathbb{E}_{(x,y)\sim\mathcal{D}(\theta;\tilde{M})}[\ \mathbb{1}\{y \neq f_\theta(x)\}\ \mathbb{1}\{x \in S(\Theta_0, c)\}\ \mathbb{1}\{y = 1\}]\right|.$$

It suffices to show that the first term is upper bounded by $A(\theta)$ and the second term is upper bounded by $B(\theta)$. Since these two bounds follow from analogous arguments, we only present the proof of the first bound.

$$\begin{aligned}
&\left|\mathbb{E}_{(x,y)\sim\mathcal{D}(\theta;M)}[\ \mathbb{1}\{y \neq f_\theta(x)\}\ \mathbb{1}\{x \in S(\Theta_0, c)\}\ \mathbb{1}\{y = 0\}] - \mathbb{E}_{(x,y)\sim\mathcal{D}(\theta;\tilde{M})}[\ \mathbb{1}\{y \neq f_\theta(x)\}\ \mathbb{1}\{x \in S(\Theta_0, c)\}\ \mathbb{1}\{y = 0\}]\right| \\
&= \mathbb{P}[x \in S(\Theta_0, c) \ \ \& \ \ y = 0]\left|\mathbb{E}_{(x,y)\sim\mathcal{D}^0_{S(\Theta_0,c)}(\theta;M)}[\ \mathbb{1}\{p(x) \geq \theta\}] - \mathbb{E}_{(x,y)\sim\mathcal{D}^0_{S(\Theta_0,c)}(\theta;\tilde{M})}[\ \mathbb{1}\{p(x) \geq \theta\}]\right| \\
&= \mathbb{P}[x \in S(\Theta_0, c) \ \ \& \ \ y = 0]\left|\mathbb{E}_{l\sim\mathcal{D}^0_{S(\Theta_0,c)}(\theta;M)}[\ \mathbb{1}\{l \geq \theta\}] - \mathbb{E}_{l\sim\mathcal{D}^{p,0}_{S(\Theta_0,c)}(\theta;\tilde{M})}[\ \mathbb{1}\{l \geq \theta\}]\right| \\
&\leq \mathbb{P}[x \in S(\Theta_0, c) \ \ \& \ \ y = 0]\mathrm{KS}\left(\mathcal{D}^{p,0}_{S(\Theta_0,c)}(\theta; M), \mathcal{D}^0_{S(\Theta_0,c)}(\theta; \tilde{M})\right).
\end{aligned}$$

□

Now, we are ready to prove Proposition 19.

Proof of Proposition 19. Let Θ_{net} be an ϵ net of Θ_0. The decision-maker uses the agent response oracle as follows. For each $\theta \in \Theta_{\text{net}}$, they can generate n_0 samples as follows: draw a sample $(x, y) \sim \mathcal{D}_{XY}$ conditioned on $y = 0$. If $x \in S(\Theta_0, c)$, then query the agent response oracle on x at θ. It is easy to see that these samples are distributed as n_0 independent samples from $\mathcal{D}^{p,0}_{S(\Theta_0,c)}(\theta)$. Similarly, the decision-maker uses the agent response oracle to draw n_1 samples that are distributed as n_1 independent samples from $\mathcal{D}^{p,1}_{S(\Theta_0,c)}(\theta)$. (We will specify the values of n_0 and n_1 later.)

First, we define a distribution map $\tilde{\mathcal{D}}$ using these samples and the base distribution. Let's define $D_0(\theta)$ to be the empirical distribution of the n_0 samples, and let $D_1(\theta)$ be the empirical distribution of the n_1 samples. Let $D'(\theta)$ be the distribution given by a mixture of $(x, 0)$ where $x \sim D_0(\theta)$ with probability $\mathbb{P}_{\mathcal{D}_{XY}}[y = 0 \mid x \in S(\Theta_0, c)]$ and $x \sim D_1(\theta)$ with probability $\mathbb{P}_{\mathcal{D}_{XY}}[y = 1 \mid x \in S(\Theta_0, c)]$. Let $D''(\theta)$ be the distribution given by (x, y) drawn from the conditional distribution of $\mathcal{D}_{XY}$ given $x \notin S(\Theta_0, c)$. We let $\tilde{\mathcal{D}}$ be the distribution given by a mixture of $D'(\theta)$ with probability $\mathbb{P}_{\mathcal{D}_{XY}}[x \in S(\Theta_0, c)]$ and $D''(\theta)$ with probability $1 - \mathbb{P}_{\mathcal{D}_{XY}}[x \in S(\Theta_0, c)]$.

We can microfound $\tilde{\mathcal{D}}$ with a map $\tilde{M}$ as follows. Let $\tilde{M}(x, y) = x$ when $x \notin S(\Theta_0, c)$. Let $\tilde{M}$ on $S(\Theta_0, c) \times Y$ be defined in such any way it microfounds $D'(\theta)$ (this is possible because of Proposition 15). It is easy to see that $\tilde{M}$ microfounds $\tilde{\mathcal{D}}$ and that $\tilde{M}$ restricted to the domain $(X \setminus S(\Theta_0, c)) \times Y$ is expenditure-constrained. This means that we can apply Lemma 20.

Now, we bound the performative risk $\text{PR}(\theta_{\text{PO}}(\tilde{M}))$, where:

$$\theta_{\text{PO}}(\tilde{M}) = \text{argmin}_{\theta \in \Theta_0} \mathbb{E}_{(x,y)\sim\mathcal{D}(\theta;\tilde{M})}[\mathbb{1}\{y \neq f_\theta(x)\}] = \text{argmin}_{\theta \in \Theta_{\text{net}}} \mathbb{E}_{(x,y)\sim\mathcal{D}(\theta;\tilde{M})}[\mathbb{1}\{y \neq f_\theta(x)\}].$$

In order to apply Lemma 20, we need to bound:

$$\sup_{\theta \in \Theta_{\text{net}}} \{A(\theta) + B(\theta)\} \tag{9}$$

where:

$$A(\theta) := \mathbb{P}[x \in S(\Theta_0, c) \ \ \& \ \ y = 0] \cdot \text{KS}\left(\mathcal{D}^{p,0}_{S(\Theta_0,c)}(\theta; M), \mathcal{D}^{p,0}_{S(\Theta_0,c)}(\theta; \tilde{M})\right)$$
$$B(\theta) := \mathbb{P}[x \in S(\Theta_0, c) \ \ \& \ \ y = 1] \cdot \text{KS}\left(\mathcal{D}^{p,1}_{S(\Theta_0,c)}(\theta; M), \mathcal{D}^{p,1}_{S(\Theta_0,c)}(\theta; \tilde{M})\right).$$

To bound (9), we union bound over Θ_{net}. This set has cardinality $O(1/\epsilon)$. Notice that with probability $\geq 1 - \alpha$, we know that:

$$\text{KS}\left(\mathcal{D}^{p,0}_{S(\Theta_0,c)}(\theta; M), \mathcal{D}^{p,0}_{S(\Theta_0,c)}(\theta; \tilde{M})\right) \leq \sqrt{\frac{\ln(2/\alpha)}{2n_0}}.$$
$$\text{KS}\left(\mathcal{D}^{p,1}_{S(\Theta_0,c)}(\theta; M), \mathcal{D}^{p,1}_{S(\Theta_0,c)}(\theta; \tilde{M})\right) \leq \sqrt{\frac{\ln(2/\alpha)}{2n_0}}.$$

We can now set $\alpha = \Theta(\epsilon/100)$ in the previous result to obtain that with probability $\geq 99/100$, the expression in (9) is bounded by:

$$E := O\left(\mathbb{P}_{\mathcal{D}_{XY}}[x \in S(\Theta_0, c) \wedge y = 0]\sqrt{\frac{\ln(2/\epsilon)}{2n_0}}\right) + O\left(\mathbb{P}_{\mathcal{D}_{XY}}[x \in S(\Theta_0, c) \wedge y = 1]\sqrt{\frac{\ln(2/\epsilon)}{2n_1}}\right).$$

We can now apply Lemma 20 to Θ_{net} to see that:

$$\text{PR}(\theta_{\text{PO}}(\tilde{M})) \leq E + \min_{\theta \in \Theta_{\text{net}}} \mathbb{E}_{(x,y)\sim\mathcal{D}(\theta;M)}[\mathbb{1}\{y \neq f_\theta(x)\}],$$

Now, let's use the Lipschitz requirement on the distribution map to move to the set Θ_0. Let's consider a distribution map $\mathcal{D}'$ that is defined as follows: for $\theta \in \Theta_{\text{net}}$, we take $\mathcal{D}'(\theta) := \mathcal{D}(\theta)$, and for $\theta \notin \Theta_{\text{net}}$, we take $\mathcal{D}'(\theta) := \mathcal{D}(\theta')$ where θ' is the closest element in Θ_{net} to θ. Now, let's apply Lemma 18 to $\mathcal{D}$ and $\mathcal{D}'$ on Θ_0 to obtain that:

$$\begin{aligned}\min_{\theta \in \Theta_{\text{net}}} \mathbb{E}_{(x,y)\sim\mathcal{D}(\theta;M)}[\mathbb{1}\{y \neq f_\theta(x)\}]) &\leq \epsilon + \min_{\theta \in \Theta_0} \mathbb{E}_{(x,y)\sim\mathcal{D}(\theta;M)}[\mathbb{1}\{y \neq f_\theta(x)\}] \\ &= \epsilon + \min_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim\mathcal{D}(\theta;M)}[\mathbb{1}\{y \neq f_\theta(x)\}].\end{aligned}$$

This means that

$$\mathrm{PR}(\theta_{\mathrm{PO}}(\tilde{M})) \leq E + \epsilon + \mathrm{PR}(\theta_{\mathrm{PO}}(M)).$$

Thus, it suffices to bound E and set n_0 and n_1 appropriately. Suppose that

$$n_0 = \Theta\left(\mathbb{P}_{\mathcal{D}_{XY}}[x \in S(\Theta_0, c) \ \ \& \ \ y = 0]^2 \frac{\ln(1/\epsilon)}{2\epsilon^2}\right)$$

and

$$n_1 = \Theta\left(\mathbb{P}_{\mathcal{D}_{XY}}[x \in S(\Theta_0, c) \ \ \& \ \ y = 1]^2 \frac{\ln(1/\epsilon)}{2\epsilon^2}\right).$$

Plugging in these expressions into the expression for E, we obtain the desired bounds. Moreover, notice that the total number of queries to the oracle is $\Theta(1/\epsilon) \cdot (n_0 + n_1) \leq \Theta\left(\mathbb{P}_{\mathcal{D}_{XY}}[x \in S(\Theta_0, c)]^2 \frac{\ln(1/\epsilon)}{2\epsilon^3}\right)$.

□