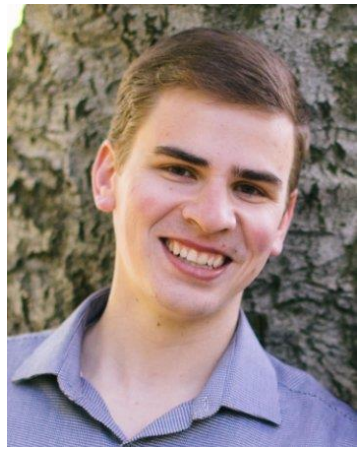
Performative Prediction in a Stateful World

Gavin Brown Shlomi Hod Iden Kalemaj

16 December 2021



BOSTON UNIVERSITY

Performativity

- ML models commonly used to make predictions that aid institutions in decision-making

Will a loan applicant default?	Will a user respond well to recommendations?	Will a candidate perform well in this job?	Is this email scam?

- Two key assumptions in supervised learning:
  - data is iid
  - data is obtained from a fixed distribution (future is like the past)
- Distribution shift and sample bias have been widely studied
- ML predictions themselves can affect behavior of the population they are trying to predict

Loan applicants strategically manipulate credit card usage	User preferences shift as they interact with recommended items.	Job applicants tailor resumes to resume-parsing algorithms.	Scamming becomes more sophisticated

- Such changes can manifest as a shift in the data-generating distribution
- Distributions are *decision-dependent,* or predictions are *performative*
- Common heuristic in practice: retraining. Does it converge? What does it converge to?

Overview

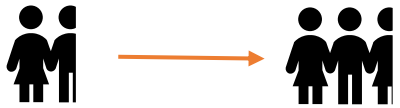
Performative prediction in a stateful world

- [Perdomo Zrnic Mendler-Dünner Hardt '20]: data distribution is a deterministic function of the currently public classifier $\theta \rightarrow f(\theta)$

  [Drusvyatskiy Xiao '20] [Izzo Ying Zou '21]
  [Mendler-Dünner Perdomo Zrnic Hardt '20]
  [Miller Perdomo Zrnic '21] [Dong Ratliff '21]
  [Maheshwari Chiu Mazumdar Sastry Ratliff '21]
  [Jagadeesan Mendler-Dünner Hardt '21]

- framework has no memory of previous classifiers / distributions
- once classifier is fixed, the distribution does not change

  e.g., unemployment benefits depend on number of children



- incorporate state: data distribution is function of the classifier and previous distribution
- decision-dependent distributions in a dynamic environment

[This work '20]
[Wood Bianchin Dall'Anese '21]
[Li Wai '21]
[Ray Ratliff Drusvyatskiy Fazel '21]

Commonalities:
- Supervised learning setting
- Institution uses empirical risk minimization
- Study retraining heuristics
- Provide formal guarantees about convergence of such heuristics

Plan

- Describe our performativity framework
- Describe our results within the framework
- Describe a simulation on credit loan applicants
- Open questions

Performativity without State

[Perdomo Zrnic Mendler-Dünner Hardt '20]

RRM
chose θ_{t+1} so that it minimizes loss on d_t

Institution

Train classifier θ_t with data drawn from distribution d_{t-1}
Suffer loss $\mathbb{E}[\ell(d_t; \theta_t)]$

Publish classifier θ_t →

← Receive sampes from d_t

World

Distribution shifts to $d_t = f(\theta_t)$

Performativity without State

[Perdomo Zrnic Mendler-Dünner Hardt '20]

RRM
chose θ_{t+1} so that it minimizes loss on d_t

Institution

Train classifier θ_t with data drawn from distribution d_{t-1}
Suffer loss $\mathbb{E}[\ell(d_t; \theta_t)]$

Publish classifier θ_t →

← Receive sampes from d_t

World

Distribution shifts to $d_t = f(\theta_t)$

This model is memoryless:

- the current distribution has no dependence on previously deployed classifiers
- no dependence on previous distributions
- if the classifier stays the same, the distribution never changes

Performativity with State

1. There are scenarios where to model updated distribution, we also need information about previous distribution
2. Information propagates at different rates, so it takes time for distribution to settle

Stateful

Consumers modify credit card usage to improve their credit score

Stateful

College admission criteria change over the years, but information propagates at different rates across high schools.

Stateless

Every day, commuters choose between train or car based on projected traffic.

Performativity with State

RRM
chose θ_{t+1} so that it minimizes loss on d_t

Assume $\theta \in \mathbb{R}^n$

Institution

Train classifier θ_t
with data from d_{t-1}
Suffer loss $\mathbb{E}[\ell(d_t; \theta_t)]$

Publish classifier θ_t

Receive sampes from d_t

World

Update
$d_t = \mathrm{Tr}(d_{t-1}, \theta_t)$

Transition map is fixed and unknown to institution

Transition map is deterministic

Performativity with State

RRM
chose θ_{t+1} so that it minimizes loss on d_t

Assume $\theta \in \mathbb{R}^n$

Institution

Train classifier θ_t
with data from d_{t-1}
Suffer loss $\mathbb{E}[\ell(d_t; \theta_t)]$

Publish classifier θ_t

Receive sampes from d_t

World

Update
$d_t = \mathrm{Tr}(d_{t-1}, \theta_t)$

Transition map is fixed and unknown to institution

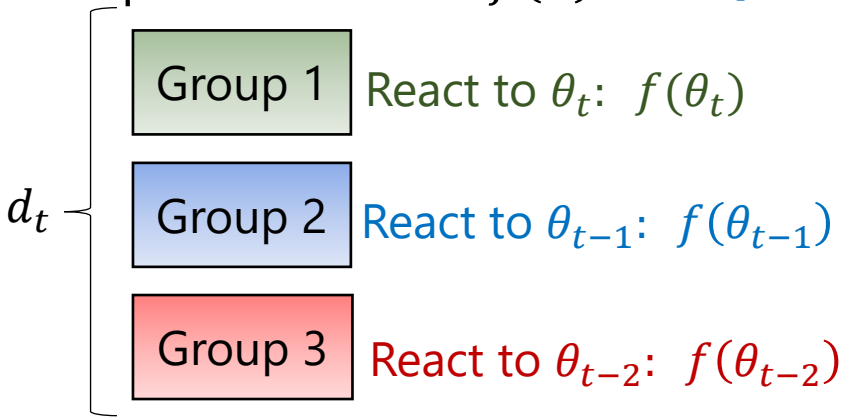
Transition map is deterministic

Questions about the framework?

Theoretical Examples

k Groups Respond Slowly

- Assume there is a best response function $f(\theta)$ as in [Perdomo Zrnic Mendler-Dünner Hardt '20]:



Geometric Decay

- Assume there is a best response function $f(\theta)$ as in [Perdomo Zrnic Mendler-Dünner Hardt '20]:

$$\mathrm{Tr}(d_{t-1}, \theta_t) = (1-\delta)d_{t-1} + \delta f(\theta_t) \text{ for } \delta \in [0,1]$$

- Setting studied in [Ray Ratliff Drusvyatskiy Fazel '21]

Markov transitions

- $\mathrm{Tr}(d, \theta) = A_\theta d$, where A_θ is a stochastic matrix.
- Studied in [Li Wai '21]

Plan

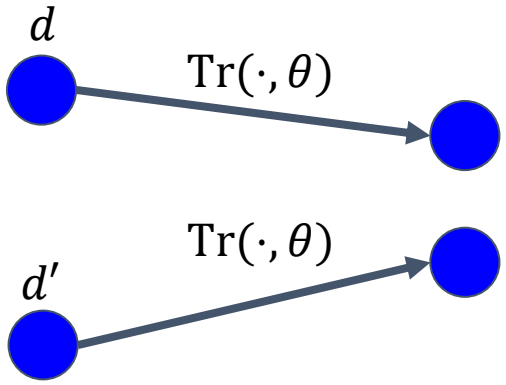
✓ Describe our performativity framework

- Describe our results within the framework
- Describe a simulation on credit loan applicants
- Open questions

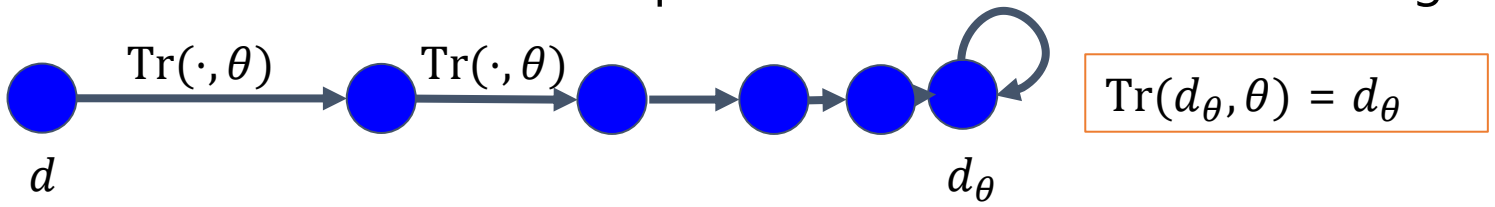
Optimal Pair

- How do we define an optimal strategy – ideally no dependence on initialal distribution d_0
- Applying the same classifier can cause e.g., alternation between two distributions
- Require the transition map to be contractive (ε-Lipschitz with $\varepsilon < 1$) w.r.t. to θ:

$$\mathcal{W}_1(\mathrm{Tr}(d,\theta),\mathrm{Tr}(d',\theta)) \leq \varepsilon \mathcal{W}_1(d,d')$$



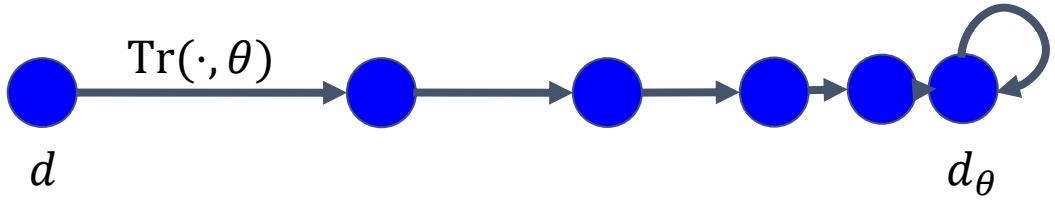
- Iterated application of the same classifier causes sequence of distributions to converge

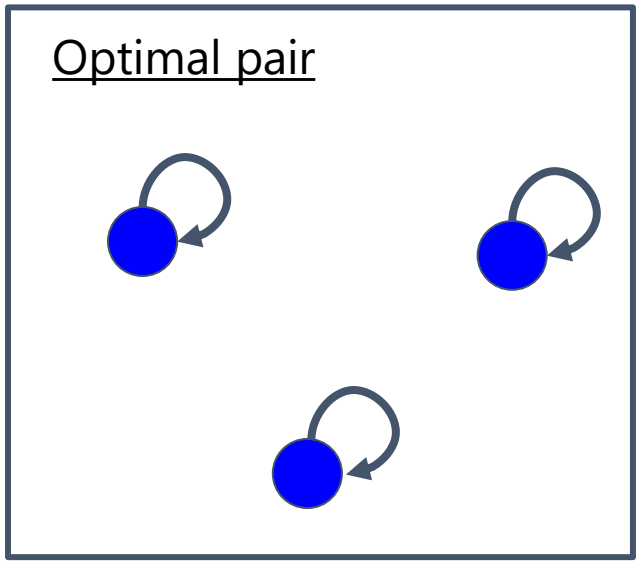


- Given θ, let d_θ be (unique) fixed-point distribution for θ (Banach's fixed point theorem)

Optimal Pair

- Given θ, let d_θ be (unique) fixed-point distribution for θ (Banach's fixed point theorem)



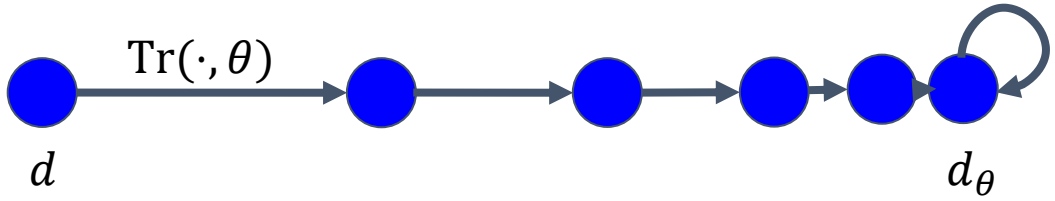


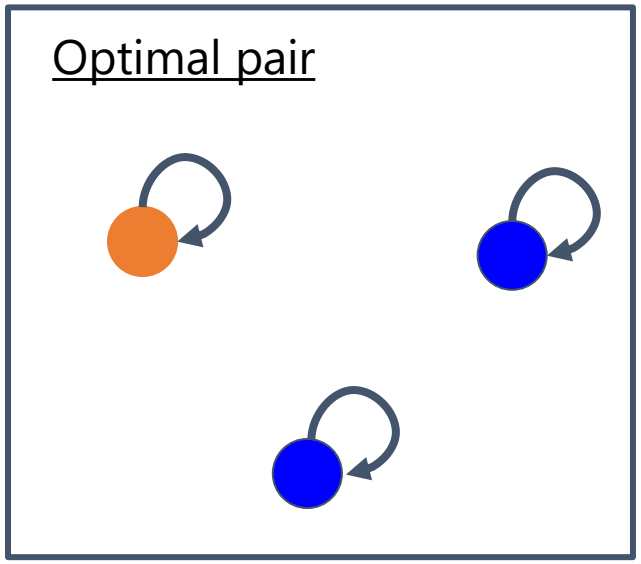
- For each (d_θ, θ) evaluate $\mathbb{E}[\ell(d_\theta; \theta)]$
- Choose pair that minimizes this quantity

$$\theta_{op} = \operatorname{argmin}_{\theta \in \Theta} \mathbb{E}[\ell(d_\theta; \theta)]$$

Optimal Pair

- Given θ, let d_θ be (unique) fixed-point distribution for θ (Banach's fixed point theorem)





- For each (d_θ, θ) evaluate $\mathbb{E}[\ell(d_\theta; \theta)]$
- Choose pair that minimizes this quantity

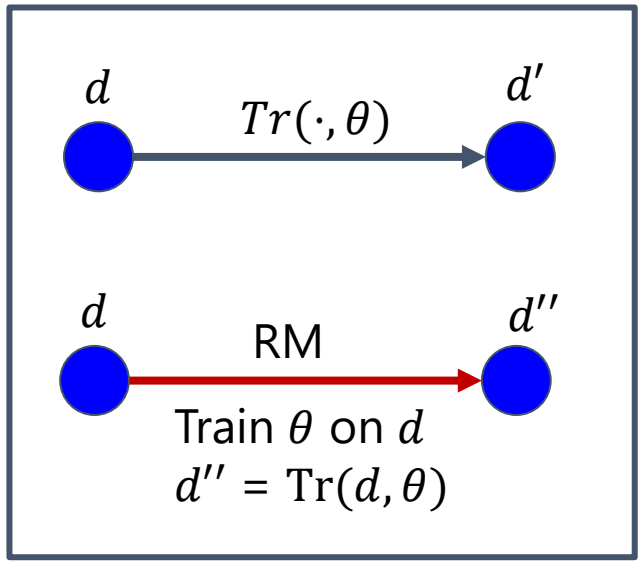
$$\theta_{op} = \operatorname{argmin}_{\theta \in \Theta} \mathbb{E}[\ell(d_\theta; \theta)]$$

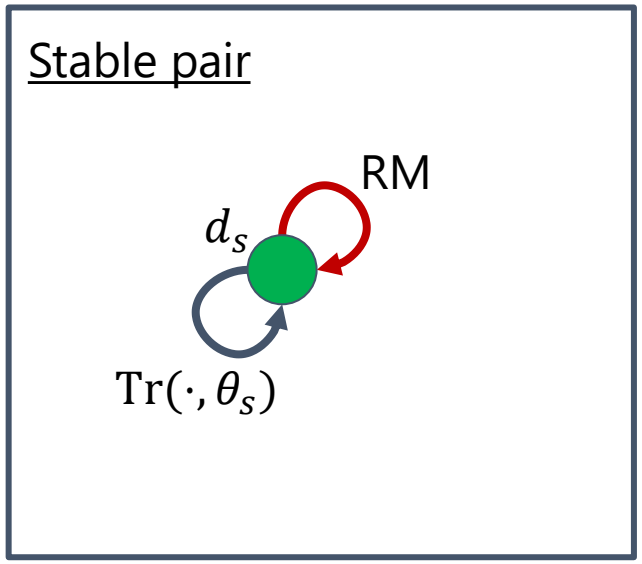
Questions about optimality?

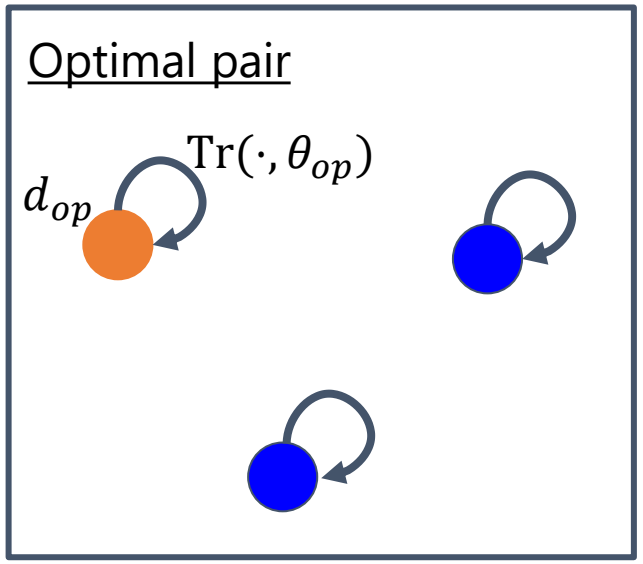
- If institution applies θ_{op} repeatedly from the beginning, it converges to $d_{\theta_{op}}$ at linear rate
- … and achieves smallest loss as compared to all single-classifier strategies
- Optimal points always exist if Tr map is contractive

Stable Pair

- Institution does not know Tr map a priori, so it cannot directly calculate θ_{op}
- Unfortunately, RRM does not converge to θ_{op}
- Instead, we show conditions so that RRM can converge to a pair that is near (d_{op}, θ_{op})
- Convergence is to a stable pair (d_s, θ_s)



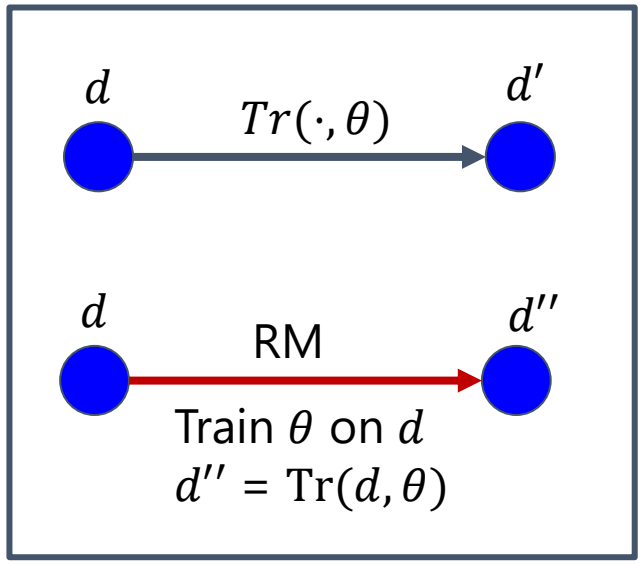


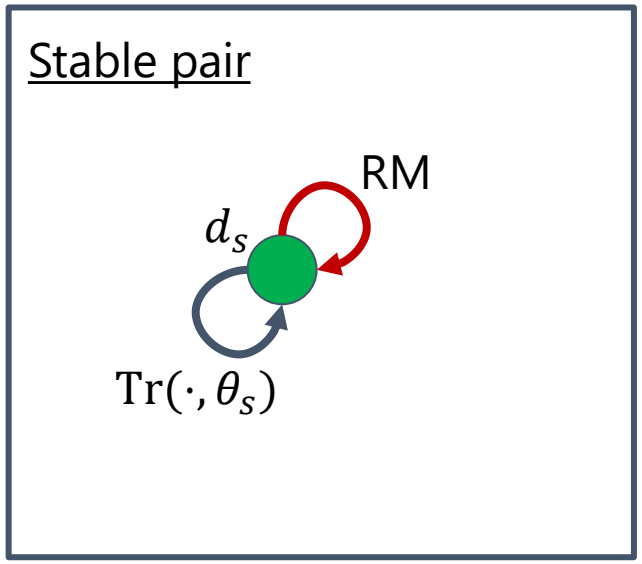


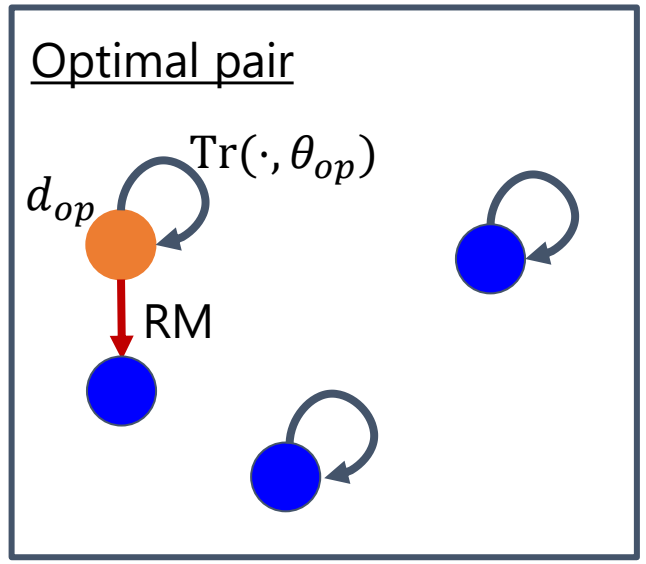
- θ_s is the best classifier for d_s (no need to retrain)

Stable Pair

- Institution does not know Tr map a priori, so it cannot directly calculate θ_{op}
- Unfortunately, RRM does not converge to θ_{op}
- Instead, we show conditions so that RRM can converge to a pair that is near (d_{op}, θ_{op})
- Convergence is to a stable pair (d_s, θ_s)







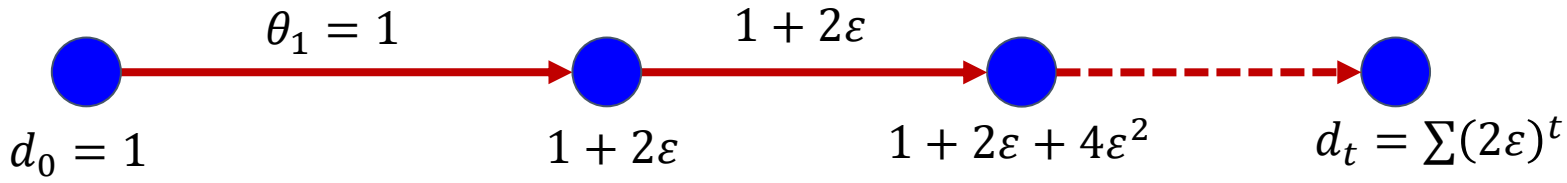
- θ_s is the best classifier for d_s (no need to retrain)
- Whereas θ_{op} is not necessarily the best classifier for d_{op}
- Stable pairs need not always exist (even if Tr map is contractive)

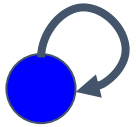
Questions about stability?

Example

Example. Distributions d are point mass over $[1, \infty)$. Suppose $d_0 = 1$.

$\mathrm{Tr}(d, \theta) = 1 + \varepsilon d + \varepsilon\theta$. Loss function is $(y - \theta)^2$. $\theta \in [1, \infty)$. $\varepsilon < 1$





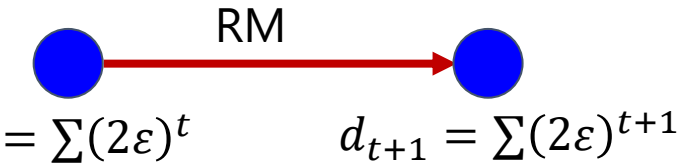
For each θ, the fixed-point distribution is $1 + \varepsilon d_\theta + \varepsilon\theta = d_\theta \Rightarrow d_\theta = \frac{1+\varepsilon\theta}{1-\varepsilon}$

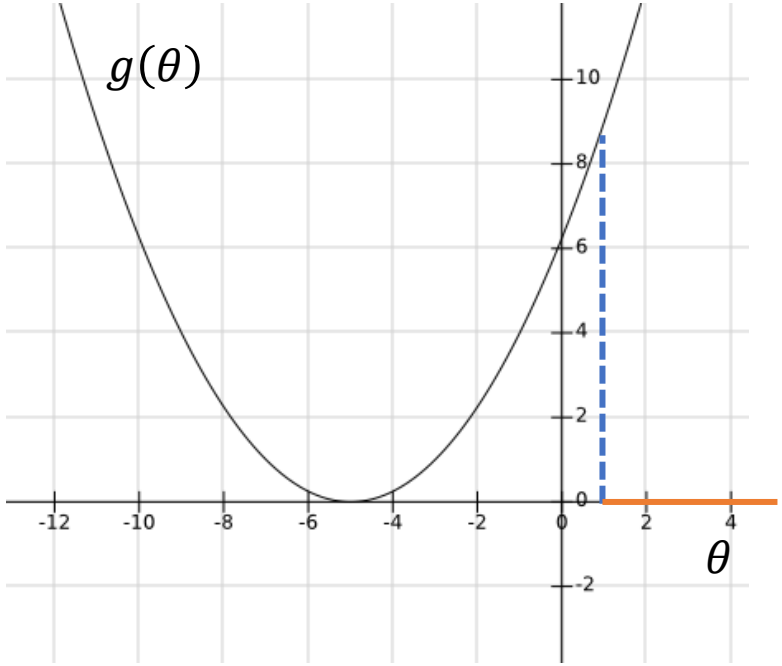
Optimality: Amongst all $\theta \in [1, \infty)$, choose the one that minimizes

$$g(\theta) = \mathbb{E}[(d_\theta - \theta)^2] = \left(\frac{1+\varepsilon\theta}{1-\varepsilon} - \theta\right)^2$$

Suppose $\varepsilon = 0.6$

- The optimal pair is $\theta_{op} = 1, d_{op} = \frac{1.6}{0.4}$
- A stable pair does not exist

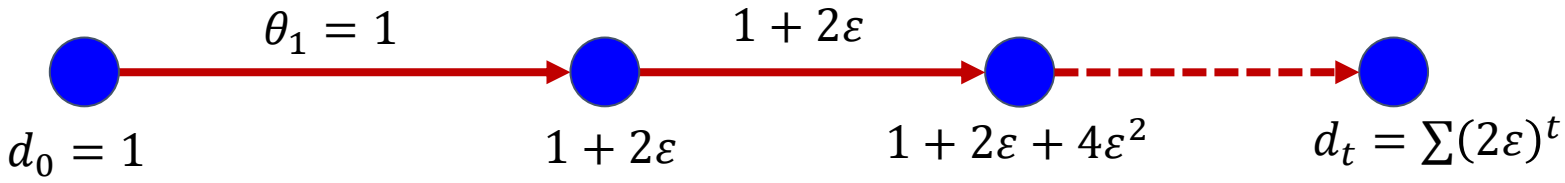


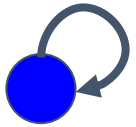


Example

Example. Distributions d are point mass over $[1, \infty)$. Suppose $d_0 = 1$.

$\mathrm{Tr}(d, \theta) = 1 + \varepsilon d + \varepsilon\theta$. Loss function is $(y - \theta)^2$. $\theta \in [1, \infty)$. $\varepsilon < 1$





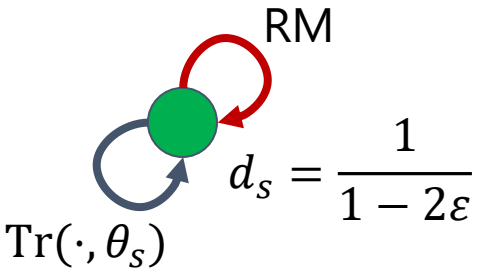
For each θ, the fixed-point distribution is $1 + \varepsilon d_\theta + \varepsilon\theta = d_\theta \Rightarrow d_\theta = \frac{1+\varepsilon\theta}{1-\varepsilon}$

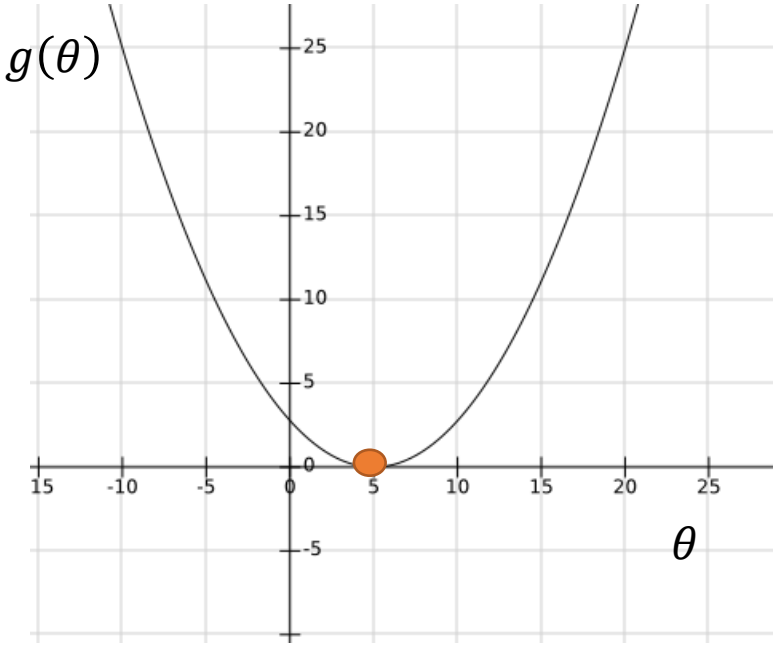
Optimality: Amongst all $\theta \in [1, \infty)$, choose the one that minimizes

$$g(\theta) = \mathbb{E}[(d_\theta - \theta)^2] = \left(\frac{1+\varepsilon\theta}{1-\varepsilon} - \theta\right)^2$$

Suppose $\varepsilon = 0.4$

- The stable pair is $d_s, \theta_s = \frac{1}{1-2\varepsilon}$
- Optimal pair is the same





Convergence to Stable Pair

Conditions for RRM to converge to a stable pair:

- Loss function is smooth (w. parameter β)

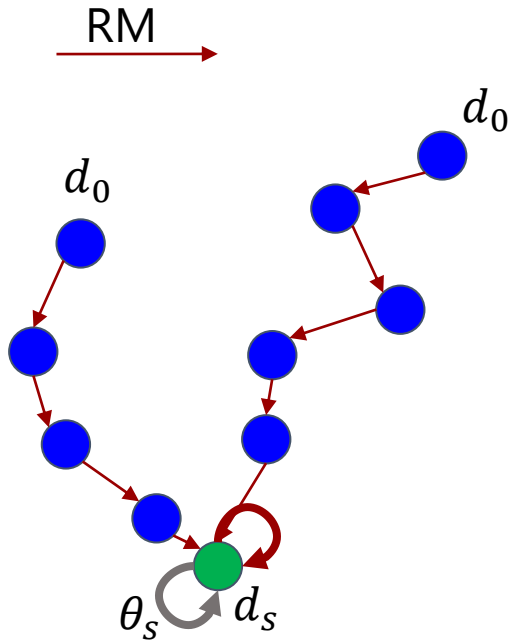
  Gradient $\nabla_\theta \ell(z,\theta)$ is β-Lipschitz in z

- Loss function is strongly convex in θ (w. parameter γ)

- Transition map is jointly ε-Lipschitz

  $$\mathcal{W}_1\big(\mathrm{Tr}(d,\theta),\mathrm{Tr}(d',\theta')\big) \le \varepsilon \mathcal{W}_1(d,d') + \varepsilon\|\theta-\theta'\|_2$$

- $\varepsilon\left(1+\frac{\beta}{\gamma}\right) < 1$



Thm 1. If above conditions hold, RRM converges to a stable pair at a linear rate, i.e., it comes within distance δ of a stable pair after

$$\left(1-\varepsilon\left(1+\frac{\beta}{\gamma}\right)\right)^{-1} \mathrm{O}(\log\frac{1}{\delta}) \text{ steps}$$

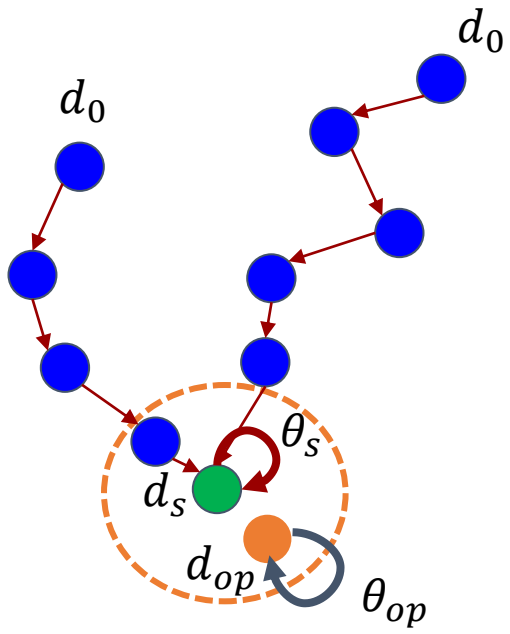
Convergence near Optimal Pair

Thm 2. Suppose:

- loss function $\ell(z, \theta)$ is L_z-Lipschitz in z and γ-strongly convex
- transition map is ε-jointly Lipschitz with $\varepsilon < 1$

For all θ_s and θ_{op} it holds

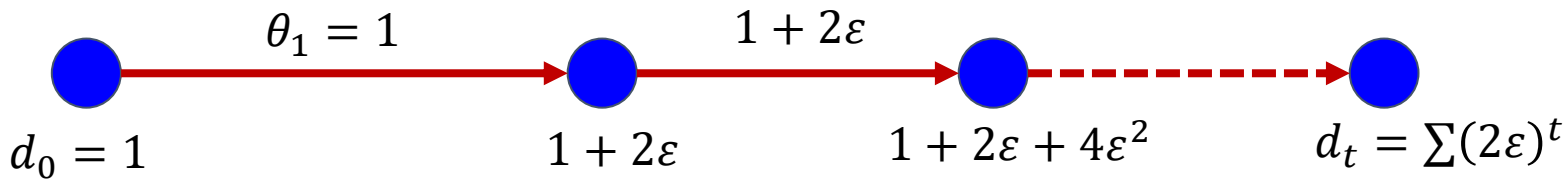
$$\|\theta_{op} - \theta_s\|_2 \leq \frac{2L_z\varepsilon}{\gamma(1-\varepsilon)}$$



Example

Example. Distributions d are point mass over $[1, \infty)$. Suppose $d_0 = 1$.

$\mathrm{Tr}(d, \theta) = 1 + \varepsilon d + \varepsilon\theta$. Loss function is $(y - \theta)^2$. $\theta \in [1, \infty)$. $\varepsilon < 1$



The loss function is β-jointly smooth and γ-strongly convex with $\beta = \gamma = 2$

If $\varepsilon < \frac{1}{1+\beta/\gamma} = \frac{1}{2}$, RRM converges to stable distribution and classifier $\frac{1}{1-2\varepsilon}$

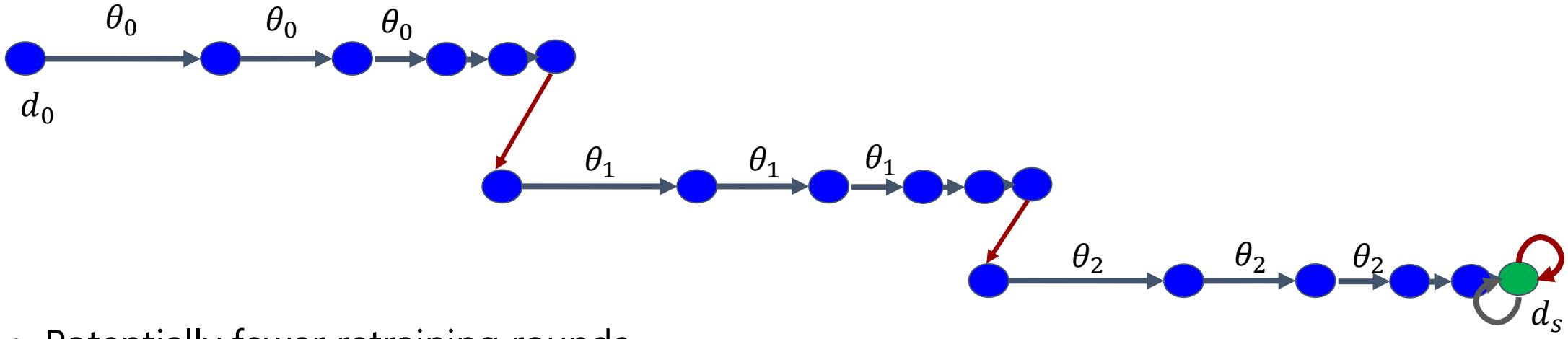
If $\varepsilon > \frac{1}{2}$, RRM diverges

RRM vs Delayed RRM

- Delayed RRM is a "lazier" algorithm for convergence to a stable point
- Uses the fact that the transition map is contractive

Repeat $O\left(\log\frac{1}{\delta}\right)$ times:

- Repeatedly deploy the same θ until distribution approaches d_θ within radius $O\left(\log\frac{1}{\delta}\right)$
- Perform empirical risk minimization on d



- Potentially fewer retraining rounds
- Speed of convergence is $O\left(\log^2\frac{1}{\delta}\right)$, i.e., square of speed of convergence of RRM
- RRM make progress towards a fixed-point distribution and a stable pair at the same time

Summary of Results

Thm 1. If the loss function is γ −strongly convex and β-smooth, and the transition map is ε-jointly Lipschitz, then for ε small enough, RRM converges to a stable point at linear rate.

Thm 2. If the loss function is γ −strongly convex and L_z-Lipschitz, and the transition map is ε-jointly Lipschitz with $\varepsilon < 1$, all stable points are near optimal points.

- Institution chooses the loss function
- No clear way to measure Lipschitzness of transition map
- Do provide some assurance that retraining converges (fast enough) to a desirable outcome

Plan

✓ Describe our performativity framework

✓ Describe our results within the framework

- Describe a simulation on credit loan applicants
- Open questions

Simulation on Loan Applicants

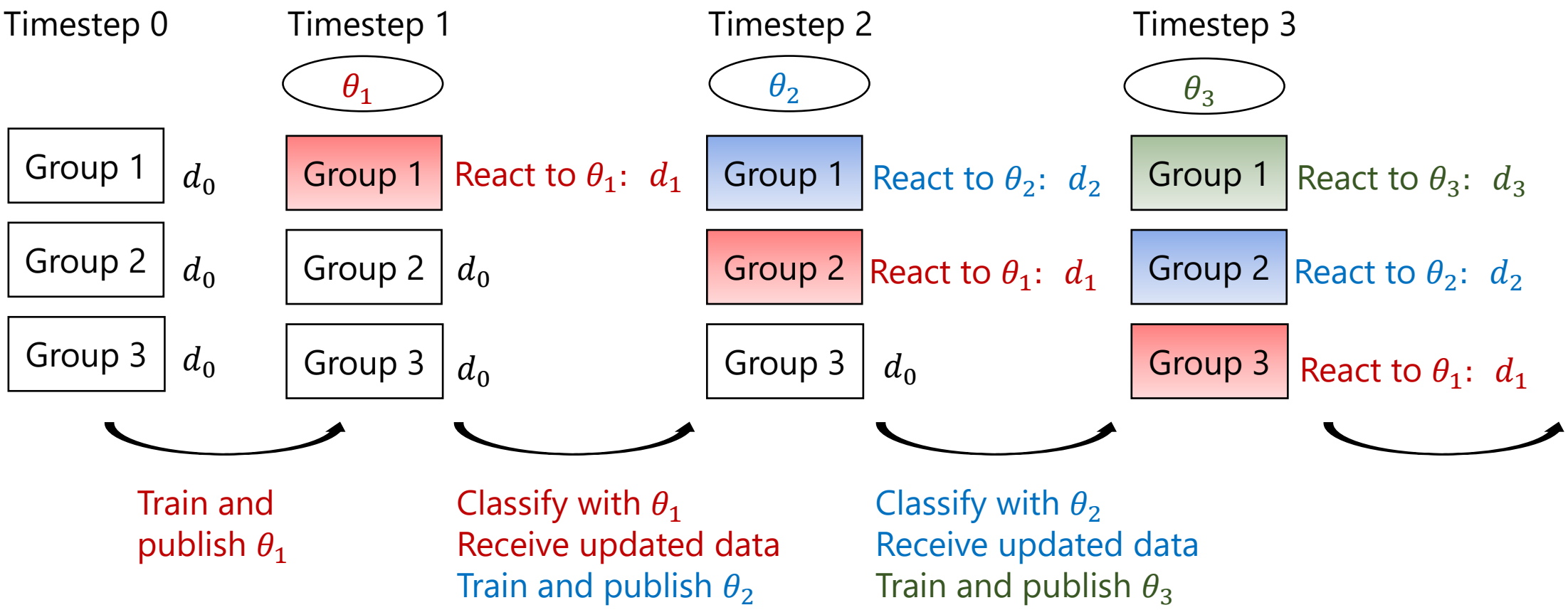
- Semi-synthetic data on loan applicants from Kaggle's GiveMeSomeCredit dataset
- Bank tries to predict whether applicant will default, using logistic regression
- Initial state of the world is determined by the dataset (11 features x 18k rows)
- Individuals update their features *strategically* but in the *delayed* fashion of k-Groups Respond Slowly
- Best-response of an individual is based on a utility and cost function

$$\text{maximize}\ u(x', \theta) - c(x', x)$$

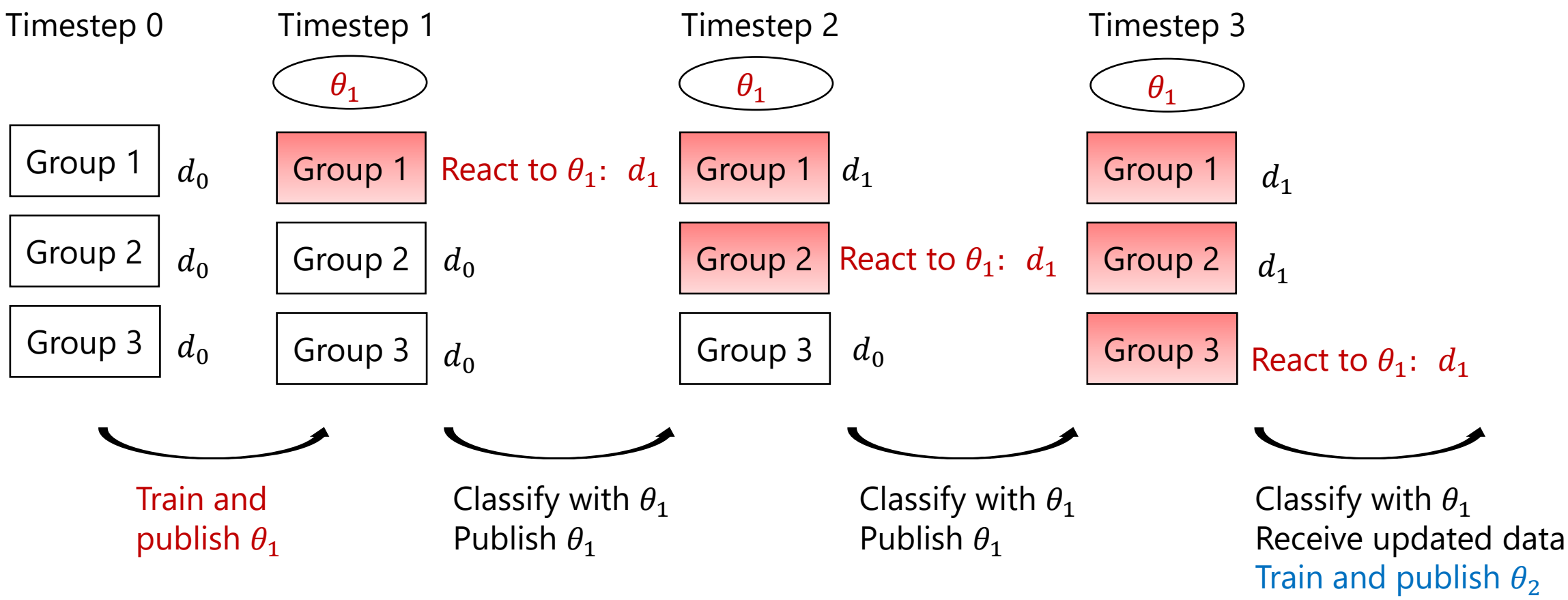
$$u(x) = -\langle \theta, x \rangle \qquad c(x, x') = \frac{1}{2\varepsilon}\|x' - x\|_2^2$$

ε controls strength of strategic response
Lower ε → Higher cost → Less performativity

Simulation on Loan Applicants - RRM

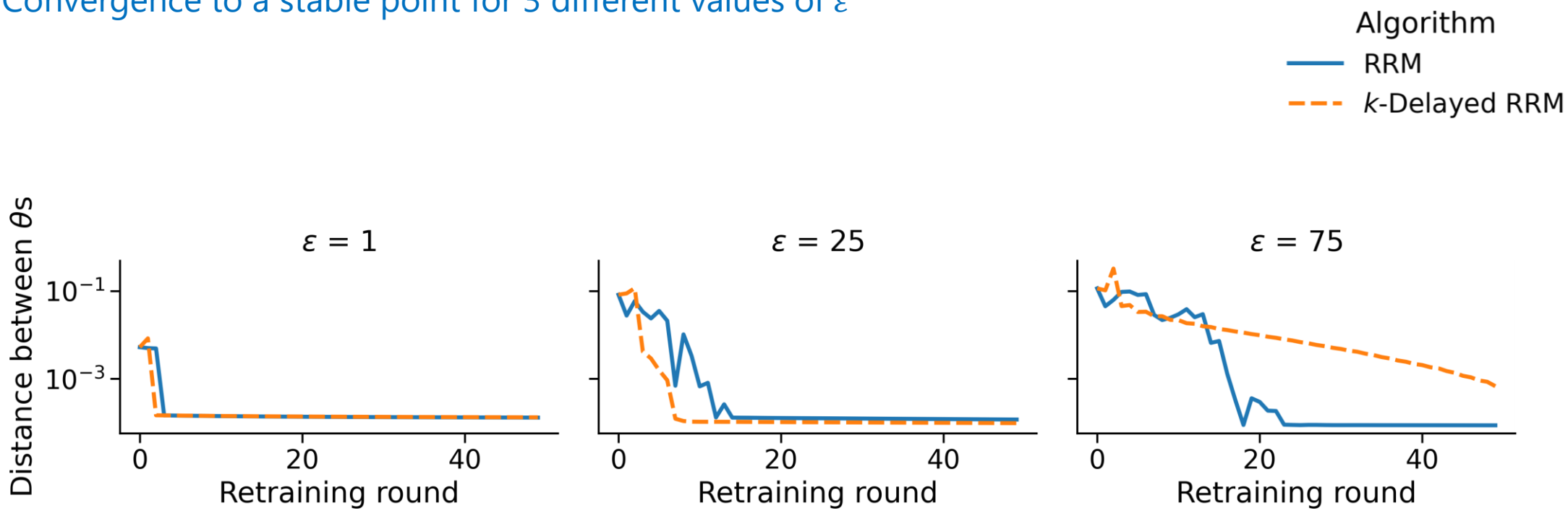


Simulation on Loan Applicants – Delayed RRM



Simulation on Loan Applicants

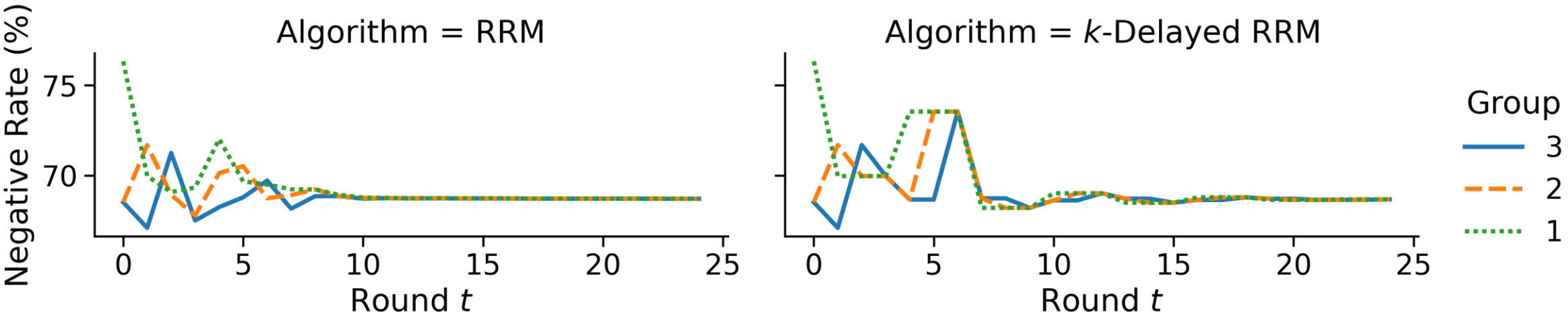
Convergence to a stable point for 3 different values of ε



In Delayed RRM the classifier is the same for 3 rounds, so we only consider retraining rounds

Simulation on Loan Applicants

From the perspective of the groups – desirable outcomes achieved



- Up until convergence, Group 1 benefits more from strategic response than other groups
- With Delayed RRM, the benefit of Group 1 (and other groups) is slightly higher
- After convergence, negative rate settles to the same value for all groups

Plan

✓ Describe our performativity framework

✓ Describe our results within the framework

✓ Describe a simulation on credit loan applicants

- Open questions

Open Questions

- We consider full access to distribution – what happens when sample is finite?

  [Perdomo Zrnic Mendler-Dünner Hardt '20] [Ray Ratliff Drusvyatskiy Fazel '21]

- Can we use stochastic methods to converge?

  [Drusvyatskiy Xiao '20] [Mendler-Dünner Perdomo Zrnic Hardt '20] [Wood Bianchin Dall'Anese '21]

- Can we converge *to* an optimal point (as opposed to near the optimal point)

  [Izzo Ying Zou '21][Miller Perdomo Zrnic '21] [Ray Ratliff Drusvyatskiy Fazel '21]

- Can we learn something about the transition function & use this information? [Miller Perdomo Zrnic '21]
- Reinforcement learning approaches?

- Is the optimal point desirable?
  - In a recommender system setting, retraining can push towards distributions where users have more extreme and less diverse preferences (fewer items are consumed)
- What if instead of fixed points, cycling between a few states is good enough?

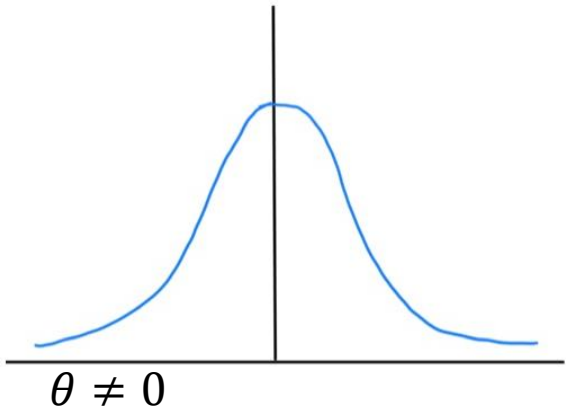
Thank you!

Stable Points vs Optimal Points

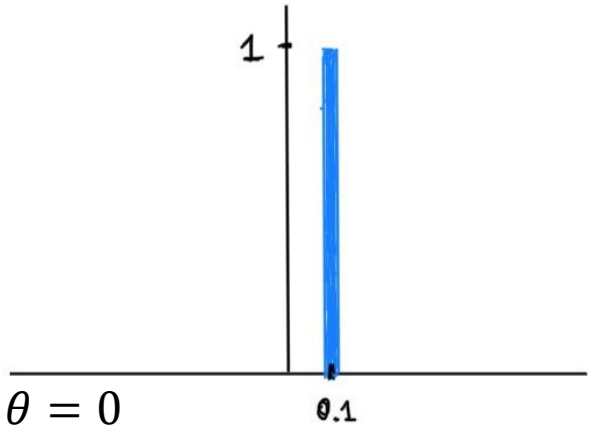
Example.

- $d_\theta = \mathcal{N}(0,1)$ if $\theta \neq 0$
- $d_\theta =$ point mass at 0.1 if $\theta = 0$
- Loss function is $(y - \theta)^2$

Distribution d_θ



Distribution d_θ



Optimal θ minimizes $\mathbb{E}[(d_\theta - \theta)^2] \Rightarrow \theta_{op} = 0$

But $\theta = 0$ is not stable, since $\theta = 0$ does not minimize loss on point mass distribution at 0.1

And any $\theta \neq 0$ is stable, since it achieves loss 1, which is minimum loss on $d_\theta = \mathcal{N}(0,1)$

$\Rightarrow$ Institution would prefer to play $\theta = 0$

$\Rightarrow$ Might have to settle for a stable point