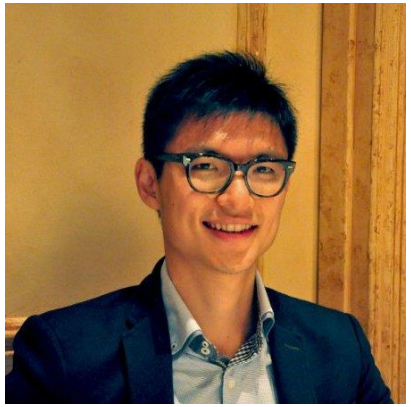
Gaming Helps! Learning From Strategic Interactions in Natural Dynamics

Conference on Artificial Intelligence and Statistics (AISTATS)

April 2021

Yahav Bechavod
Hebrew University

Katrina Ligett
Hebrew University



Steven Wu
Carnegie Mellon University

Juba Ziani
University of Pennsylvania

clark.com

5 Sneaky Ways to Improve Your Credit Score

Beverly Harzog | January 1st, 2021

clark.com

5 Sneaky Ways to Improve Your Credit Score

Beverly Harzog | January 1st, 2021

1. Find Out When Your Issuer Reports Payment History
2. Pay Down Debt Strategically
3. Pay Twice a Month
4. Raise Your Credit Limits
5. Mix It Up

Gaming

Strategic modifications to measurements, which individuals anticipate would positively affect the outcome of the decision rule.

1. College admissions
2. Credit
3. Insurance
4. Hiring
5. …

Machine Learning algorithms are increasingly involved.

Gaming may have negative consequences

May make individuals appear better than they actually are.

Goodhart's Law: "When a measure becomes a target, it ceases to be a good measure."

Approaches in prior work

1. Obfuscation of decision rule.
    - May leak over time.
    - Individuals can learn from past examples.

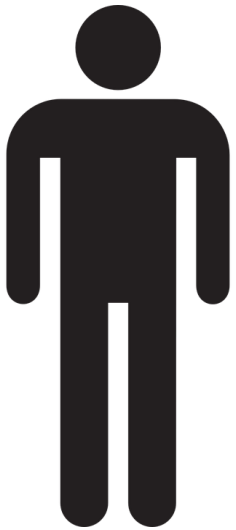
2. Robustness to gaming.
    - Additional burden on qualified individuals.
    - Cripples ability to recover or improve.

Our Approach

Gaming could actually be **helpful**!

Idea: Distinguish **false feature manipulation** from **improvement**.

Manipulation Vs. Improvement



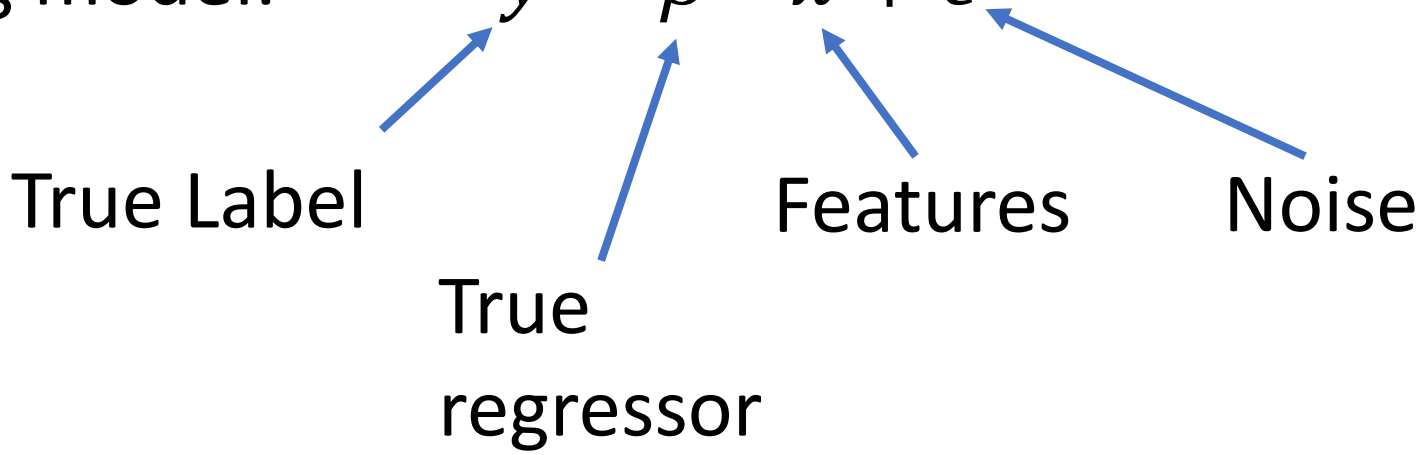
Obtain additional credit cards
Raise your credit limits
…

Reduce your debt
Increase your income
…

Our Model

Online. Linear Regression.

Underlying model: $y = \beta^{*T} x + \epsilon$



Meaningful/non-meaningful features: $\beta^* \in R^d$

Meaningful: $\beta^*_i \neq 0$

Non-Meaningful: $\beta^*_i = 0$

Motivation

If distribution over X is not full-rank, recovery of β^* is **impossible**.

Optimizing for $\hat{\beta}$ over a rank-deficient space implies:

1. Non-zero weight on non-meaningful features -> Susceptibility to false manipulations.

2. Less weight on meaningful features -> Reduced utility.

Results

An algorithm, which utilizes individuals' gaming, that guarantees:

1. Recovery of the true underlying model $\hat{\beta}$.

2. Achieving recovery within the confinements of **natural dynamics**.

   At any point, deployed scoring rule projected to the recovered subspace is optimal.

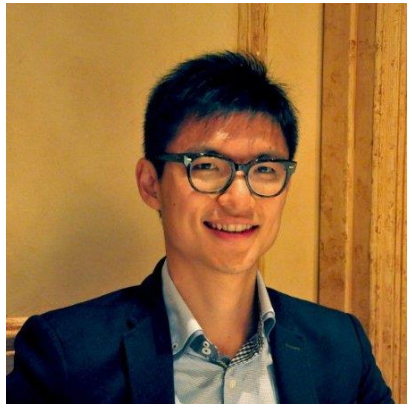
Thank you!

Yahav Bechavod
Hebrew University

yahav.bechavod@cs.huji.ac.il

Katrina Ligett
Hebrew University

katrina@cs.huji.ac.il



Steven Wu
Carnegie Mellon University

zstevenwu@cmu.edu

Juba Ziani
University of Pennsylvania

jziani@seas.upenn.edu