Maximizing Welfare with Incentive-Aware Evaluation Mechanisms

Nka Haghtalab
UC Berkeley

Brendan Lucier
Microsoft Research

Nicole Immorlica
Microsoft Research

Jack Wang
Cornell University

Roles of Classification Mechanisms

Classification:
Identify qualification

Incentivization:
Encourage qualification

Goodhart's Law

"When a measure becomes a target, it ceases to be a good measure."
-- Goodhart

Assumption for applying Goodhart's Law:

A person's true features and quality are immutable.

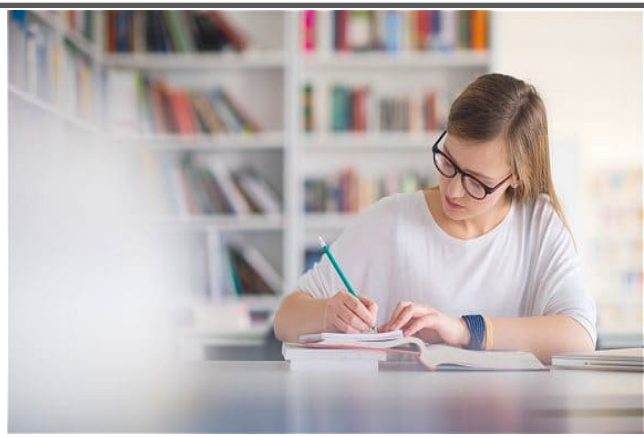
Changes in the feature as results of incentives don't impact one's quality.

An example of a Goodhart's law: Teacher's pay affected by how well their students do on tests has led to teachers tampering with tests.

Effective Change

A person's features represent their current qualifications.

People can exert effort to improve their qualifications.



Find a classifiers that incentivize distributions of agents to improve their qualification.

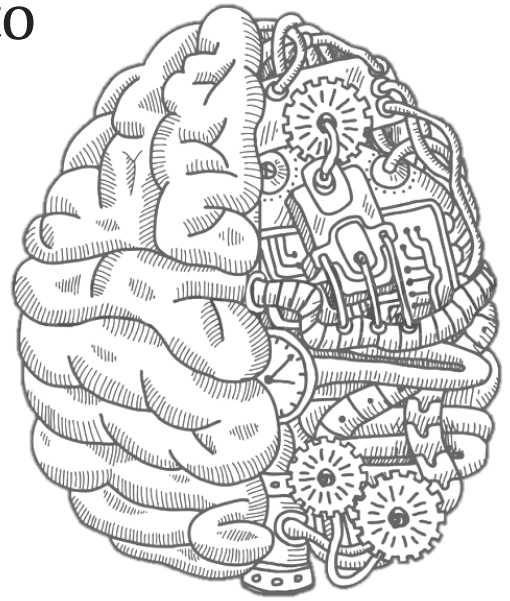
Kleinberg-Raghavan'19: Similar perspective for incentivize a single agent.

Questions

1. How do we model the problem incentivizing distributions of agents to improve their qualification?

2. How much information do we need for welfare maximization?

3. How much computational power do we need for welfare maximization?



Model

Underlying features and quality

HW/exam/SAT score,
hrs studying/volunteering

Underlying features:
$\vec{x} = (x_1, x_2, \ldots, x_n)$

$quality($  $)$

Underlying quality
e.g., linear function or its monotone transformation

Model

Underlying features and quality

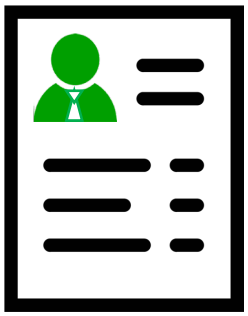
HW/exam/SAT score,
hrs studying/volunteering

Underlying features:
$\vec{x} = (x_1, x_2, \dots, x_n)$

$quality($  $)$

Underlying quality
e.g., linear function or its monotone transformation

Visible features and classification Mechanism



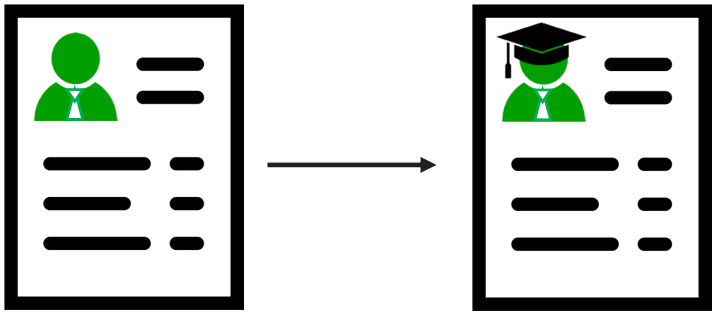
Visible features:
projection on a subspace $P\vec{x}$
0.3 HW + 0.7 Exam, SAT score, class rank

$M($  $)$

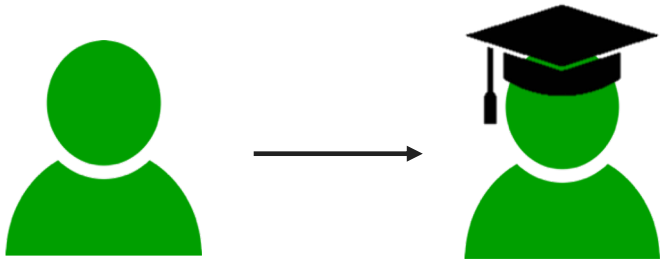
Classification mechanism for accepting/rejecting a candidate.
Choose $M \in \mathcal{M}$.

Incentive-Aware Classification

Improving visible feature



Observable change to increase $M(.)$



Underlying change

- **Cost:** Going from $\vec{x}$ to $\vec{x}'$, $\text{cost}(\vec{x}, \vec{x}') = \| \vec{x} - \vec{x}' \|_2$
- **Best Response:** Agent $\vec{x}$ changes their features to

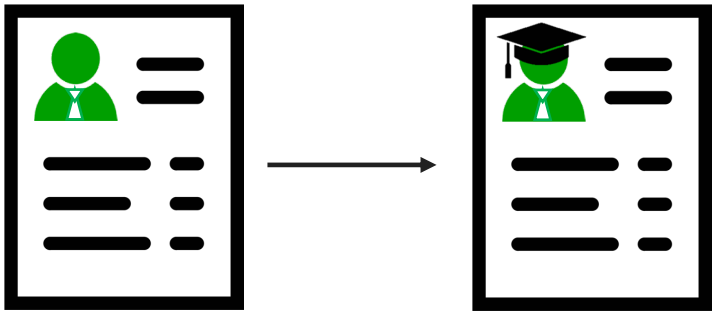
$$\text{Response}_M(\vec{x}) = \text{argmax}_{\vec{x}'} \; M(\vec{x}') - \text{cost}(\vec{x}, \vec{x}').$$

- **Goal:**

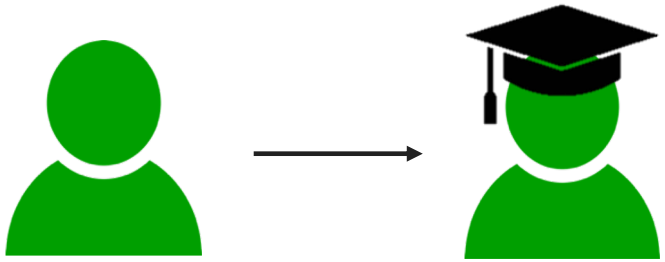
$$\text{argmax}_{M \in \mathcal{M}}$$ Expected quality of the improved features

Incentive-Aware Classification

Improving visible feature



Observable change to increase $M(.)$



Underlying change

- **Cost:** Going from $\vec{x}$ to $\vec{x}'$, $\text{cost}(\vec{x}, \vec{x}') = \| \vec{x} - \vec{x}' \|_2$
- **Best Response:** Agent $\vec{x}$ changes their features to

$$\text{Response}_M(\vec{x}) = \text{argmax}_{\vec{x}'} \; M(\vec{x}') - \text{cost}(\vec{x}, \vec{x}').$$

- **Goal:**

Expected quality of the improved features

$$\text{argmax}_{M \in \mathcal{M}} \; \mathbb{E}_{\vec{x} \sim \mathcal{D}} \left[quality\left(\text{Response}_M(\vec{x}) \right) \right]$$

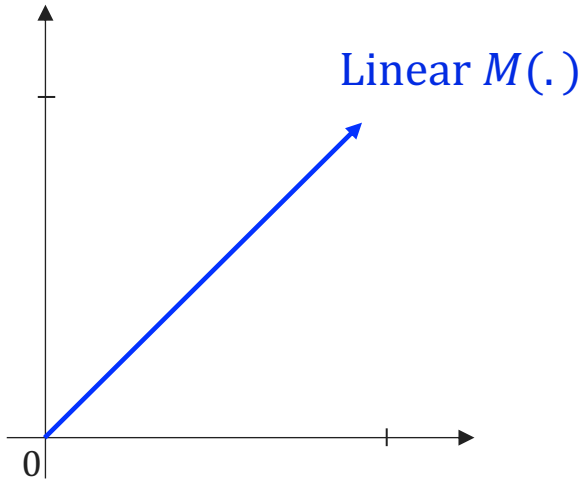
Dist. of people — true quality — resulting agent type

From Quality to Classification

Question: Can we find the optimal mechanism $M(.)$ from $quality(.)$, alone? Independently of the distribution of agent?

From Quality to Classification

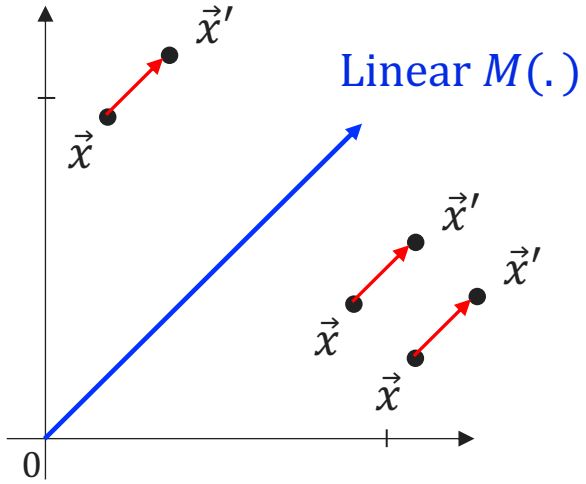
Question: Can we find the optimal mechanism $M(.)$ from $quality(.)$, alone? Independently of the distribution of agent?



$\mathcal{M}$: set of linear function

From Quality to Classification

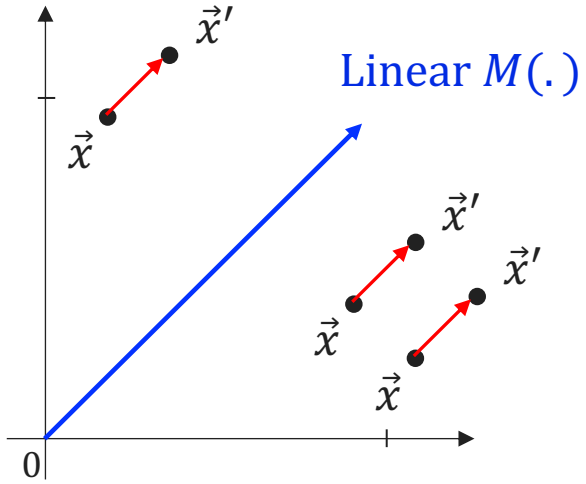
Question: Can we find the optimal mechanism $M(.)$ from $quality(.)$, alone? Independently of the distribution of agent?



$\mathcal{M}$: set of linear function

From Quality to Classification

Question: Can we find the optimal mechanism $M(.)$ from $quality(.)$, alone? Independently of the distribution of agent?



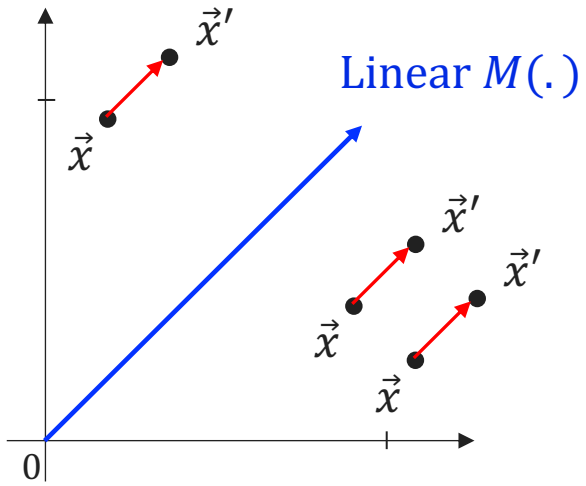
$\mathcal{M}$: set of linear function

Observation

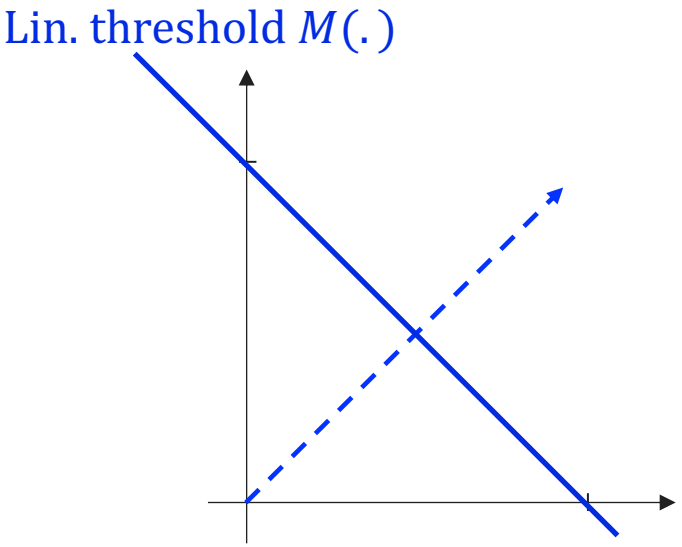
$\mathcal{M}$ **Linear:** Projection $quality(.)$ on the visible features.

From Quality to Classification

Question: Can we find the optimal mechanism $M(.)$ from $quality(.)$, alone? Independently of the distribution of agent?



$\mathcal{M}$: set of linear function



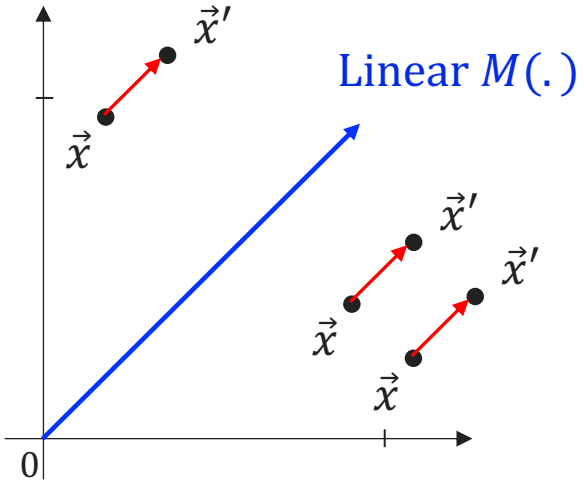
$\mathcal{M}$: set of linear **threshold** function

Observation

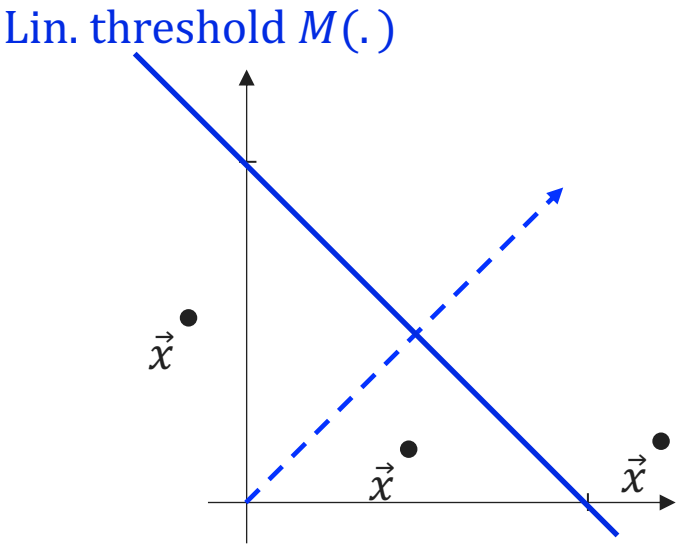
$\mathcal{M}$ **Linear:** Projection $quality(.)$ on the visible features.

From Quality to Classification

Question: Can we find the optimal mechanism $M(.)$ from $quality(.)$, alone? Independently of the distribution of agent?



$\mathcal{M}$: set of linear function



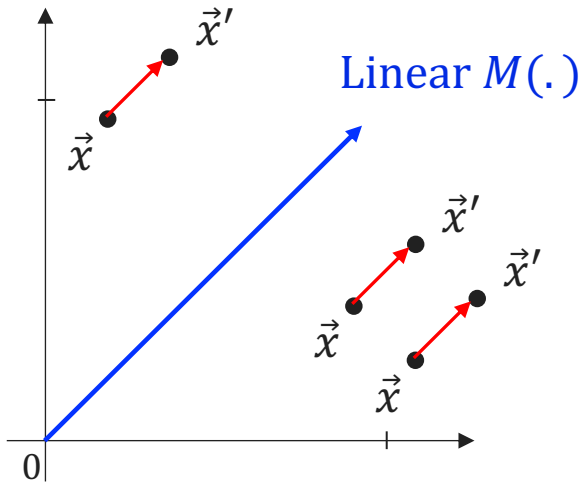
$\mathcal{M}$: set of linear **threshold** function

Observation

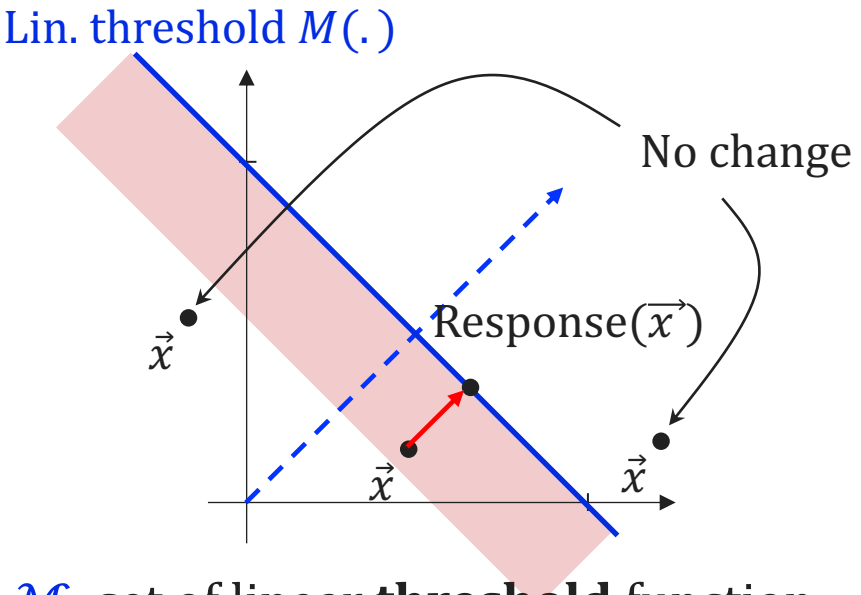
$\mathcal{M}$ **Linear:** Projection $quality(.)$ on the visible features.

From Quality to Classification

Question: Can we find the optimal mechanism $M(.)$ from $quality(.)$, alone? Independently of the distribution of agent?



$\mathcal{M}$: set of linear function



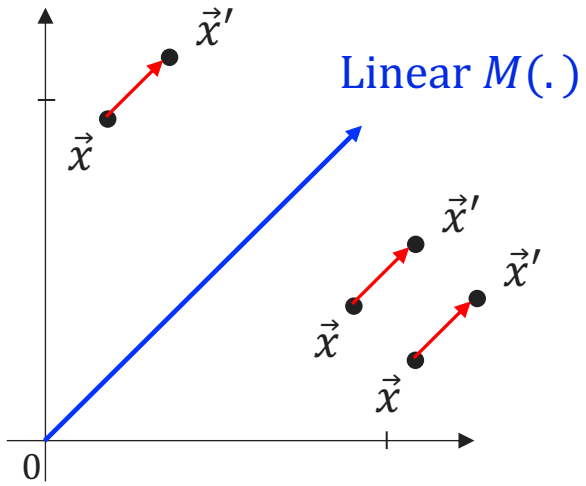
$\mathcal{M}$: set of linear **threshold** function

Observation

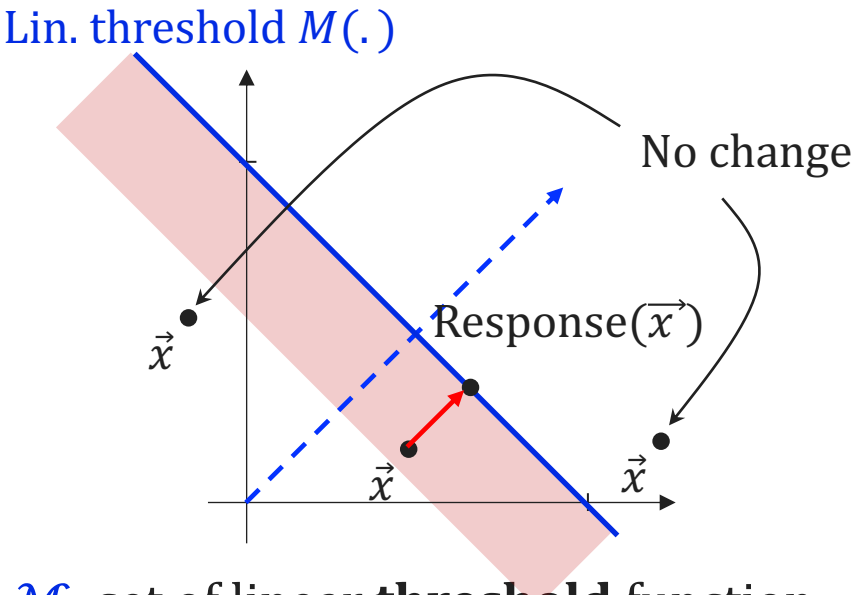
$\mathcal{M}$ **Linear:** Projection $quality(.)$ on the visible features.

From Quality to Classification

Question: Can we find the optimal mechanism $M(.)$ from $quality(.)$, alone? Independently of the distribution of agent?



$\mathcal{M}$: set of linear function



$\mathcal{M}$: set of linear **threshold** function

Observation

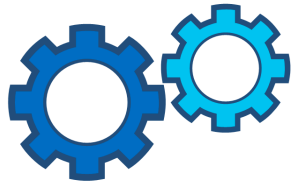
$\mathcal{M}$ **Linear:** Projection $quality(.)$ on the visible features.

$\mathcal{M}$ **Linear threshold**: Depends also on the **distribution of people.**

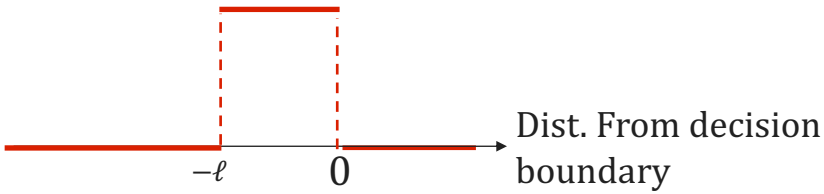
	Linear Mechanisms	**Linear Threshold Mechanisms**
Computation	projection step	$\left(\frac{1}{4} - \epsilon\right)$ approximation (using routine opt oracles)
Information	0 samples	$O\left(\frac{k}{\epsilon^2}\right)$ samples, k = # visible features

Comparable computational power and sample complexity to optimization of simple functions without incentives.

How Much Computation?

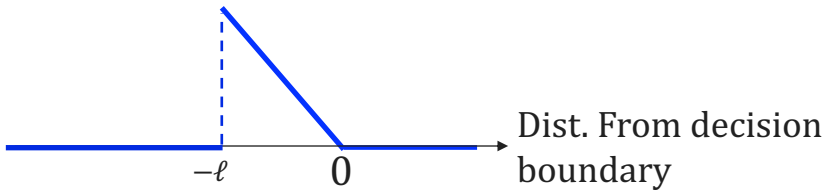


Those who fall in the margin of M

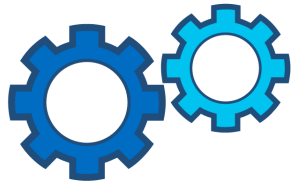


Improvement in quality induced by M

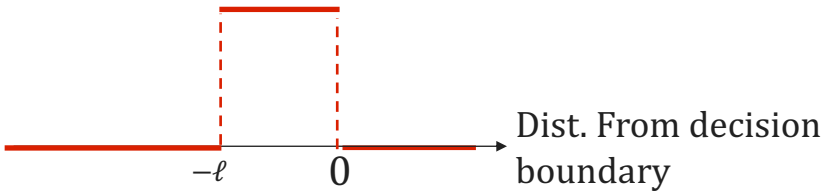
$\text{quality}(Response_M(x)) - quality(x)$:



How Much Computation?

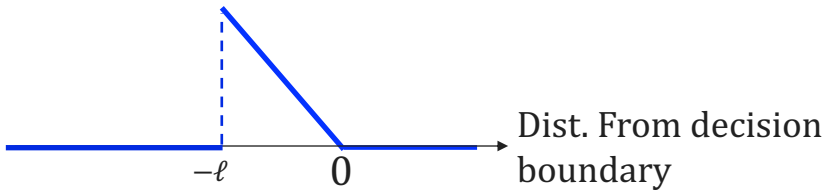


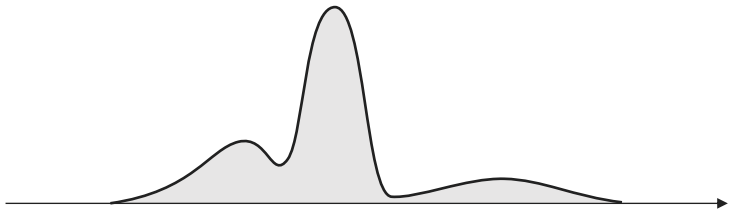
Those who fall in the margin of M



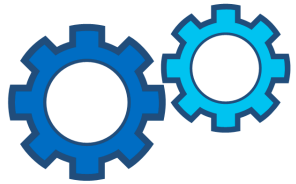
Improvement in quality induced by M

$\text{quality}(Response_M(x)) - quality(x)$:

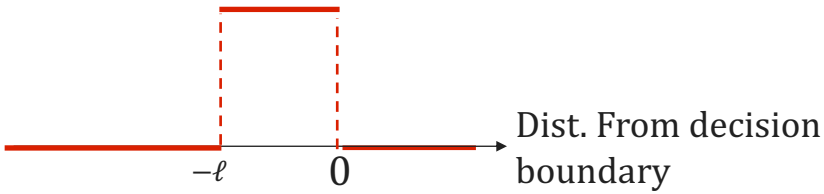




How Much Computation?

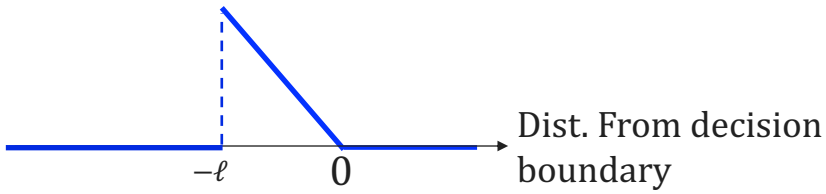


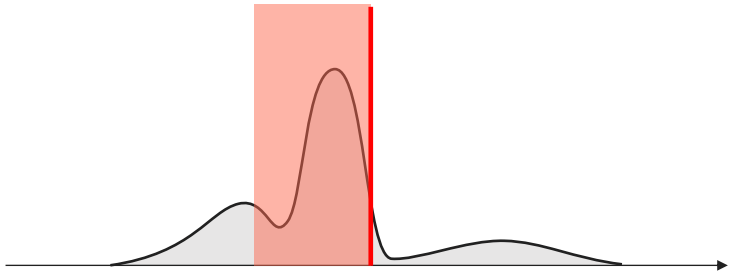
Those who fall in the margin of M



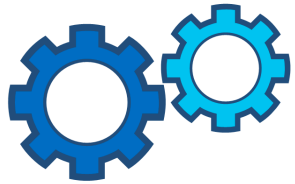
Improvement in quality induced by M

$\text{quality}(Response_M(x)) - quality(x)$:

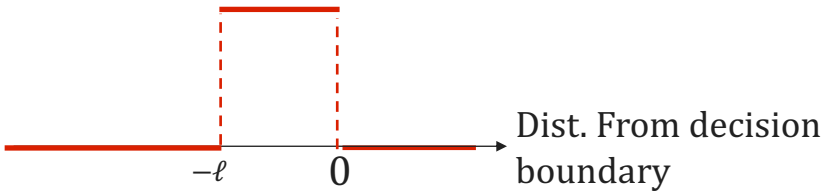




How Much Computation?

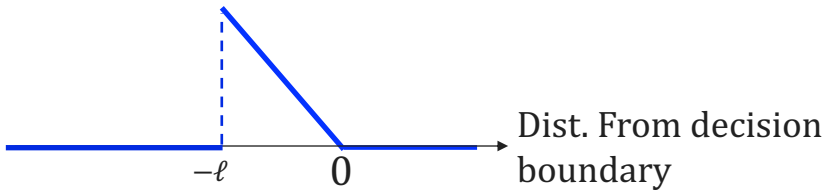


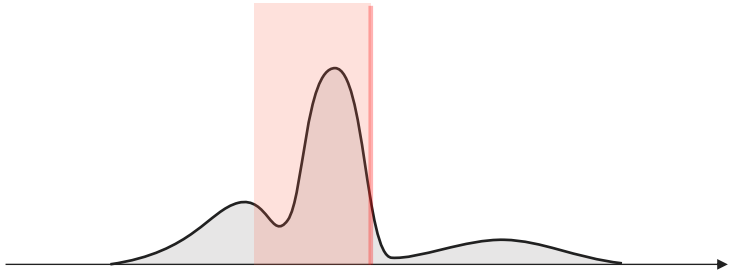
Those who fall in the margin of M



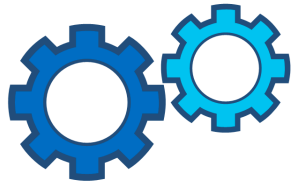
Improvement in quality induced by M

$\text{quality}(Response_M(x)) - quality(x)$:

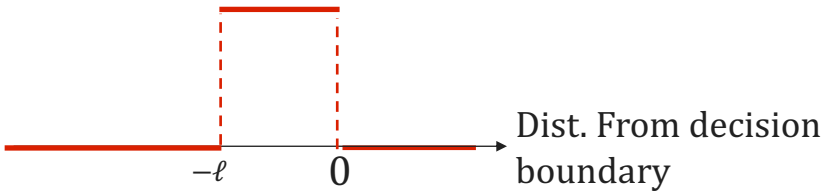




How Much Computation?

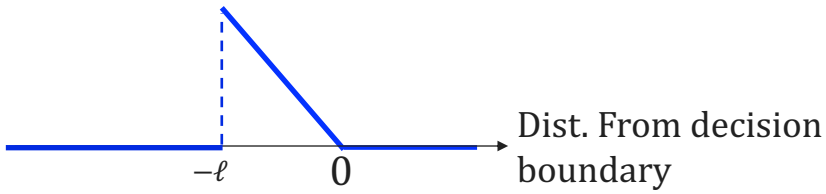


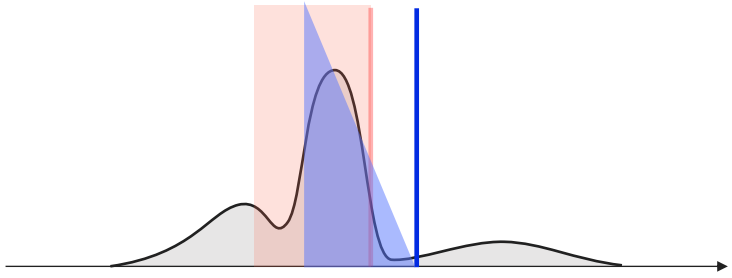
Those who fall in the margin of M



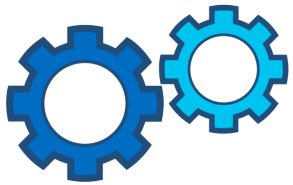
Improvement in quality induced by M

$\text{quality}(Response_M(x)) - quality(x)$:

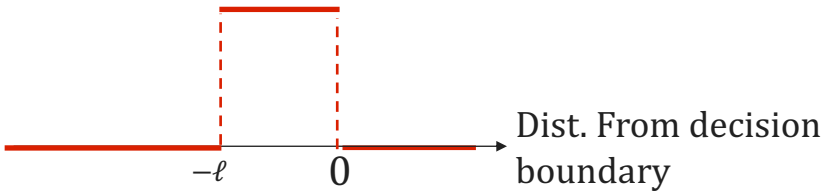




How Much Computation?

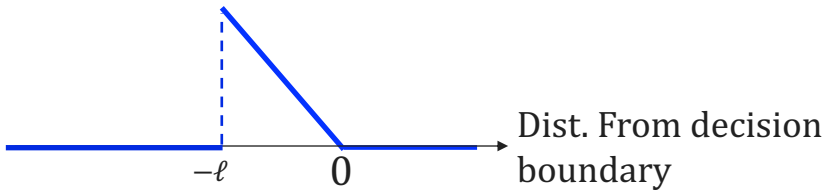


Those who fall in the margin of M



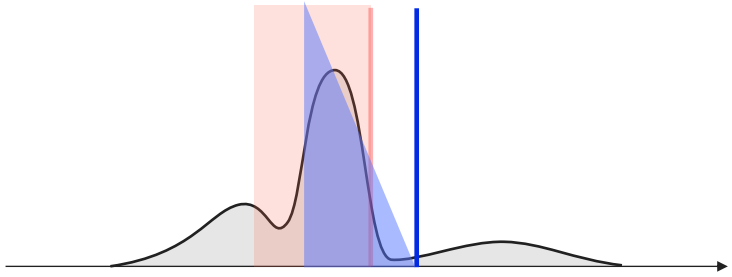
Improvement in quality induced by M

$\text{quality}(Response_M(x)) - quality(x)$:

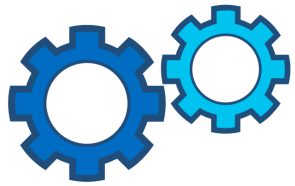


Maximizing # people in the margin and close to the direction of quality()

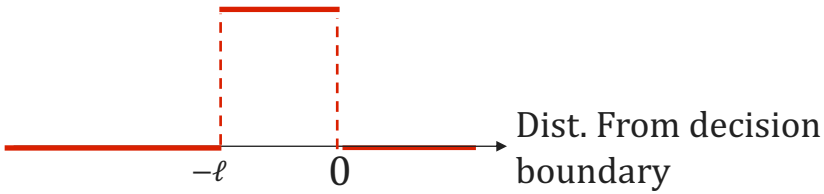
Move it and get $\geq \frac{1}{4} \times$ optimal improvement in the quality



How Much Computation?

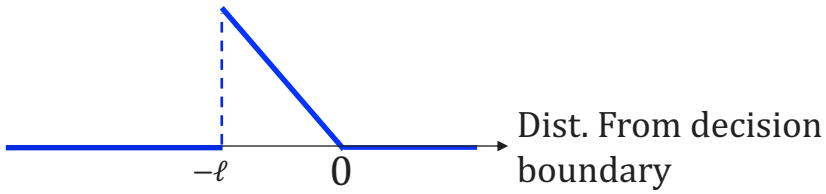


Those who fall in the margin of M



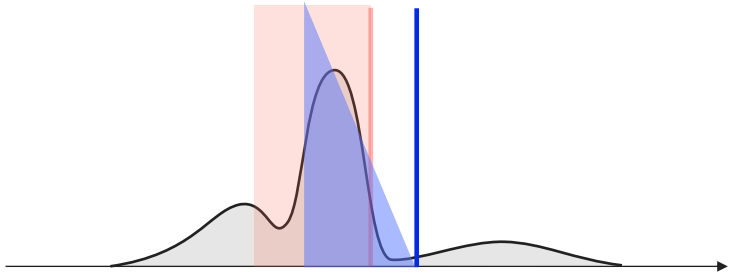
Improvement in quality induced by M

$\text{quality}(Response_M(x)) - quality(x)$:



Maximizing # people in the margin and close to the direction of quality()

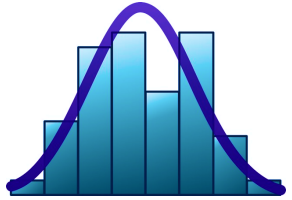
Move it and get $\geq \frac{1}{4} \times$ optimal improvement in the quality



Max margin density: NP-hard but a routine task in optimization and machine learning, even without incentives.

Computational power is the same as optimizing margin density.

How Much Information?

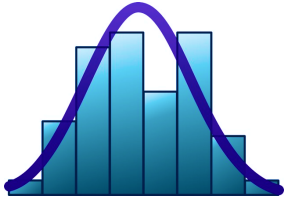


Do we need to know the distribution of features?

Just the visible features of candidates.
→ Projections on the visible subspace.

Just samples from these projections.
→ If mechanism M has low *VC dim*.
→ The $quality\left(Response_M(.)\right)$ has low *Pseudo-dim.*

How Much Information?



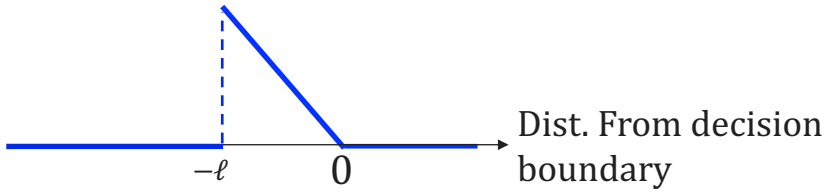
Do we need to know the distribution of features?

Just the visible features of candidates.
→ Projections on the visible subspace.

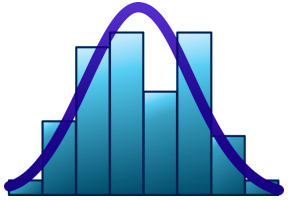
Just samples from these projections.
→ If mechanism M has low *VC dim*.
→ The $quality\left(Response_M(.)\right)$ has low *Pseudo-dim.*

Who to admit?
$M(x)$:

How much the quality improves?
$quality(Response_M(x)) - quality(x)$



How Much Information?



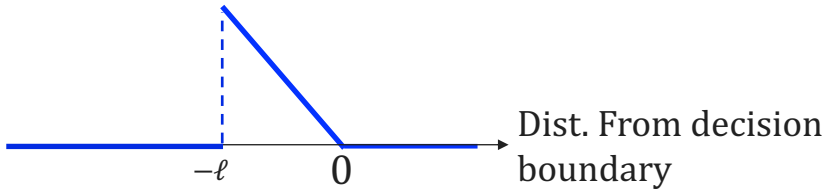
Do we need to know the distribution of features?

Just the visible features of candidates.
→ Projections on the visible subspace.

Just samples from these projections.
→ If mechanism M has low *VC dim*.
→ The $quality\left(Response_M(.)\right)$ has low *Pseudo-dim.*

Who to admit?
$M(x)$:

How much the quality improves?
$quality(Response_M(x)) - quality(x)$



→ At most $O\left(\frac{k}{\epsilon^2}\right)$ samples.

visible features

Main Message

The welfare maximizing mechanism depends on the distribution.

Comparable computation power and sample complexity to optimization of functions without incentives.

Designing classification mechanisms that optimize welfare is good for society and computationally and statistically doable.