(Improving) Muddled Information

Alex Frankel Navin Kartik

Chicago Booth Columbia

Muddled Information: *JPE* 2019
Improving Information from Manipulable Data: *In progress*

Muddled Information

Suppose our goal is to allocate **better goods** to agents with **higher types**

- Schooling: Give better college slot to higher ability student
- Credit: Loan more money / give lower interest rates to borrowers who are more likely to repay
- Web search: Give better search results to higher quality sites/products

Muddled Information

Suppose our goal is to allocate **better goods** to agents with **higher types**

- Schooling: Give better college slot to higher ability student
- Credit: Loan more money / give lower interest rates to borrowers who are more likely to repay
- Web search: Give better search results to higher quality sites/products

...But agent type is not directly observable

Evaluators only observe data that might be **manipulable** by agent

- Schooling: observe SAT score, grades
- Credit: observe credit history, FICO score
- Web search: observe keywords, incoming links, product reviews

Muddled Information

Suppose our goal is to allocate **better goods** to agents with **higher types**

- Schooling: Give better college slot to higher ability student
- Credit: Loan more money / give lower interest rates to borrowers who are more likely to repay
- Web search: Give better search results to higher quality sites/products

...But agent type is not directly observable

Evaluators only observe data that might be **manipulable** by agent

- Schooling: observe SAT score, grades
- Credit: observe credit history, FICO score
- Web search: observe keywords, incoming links, product reviews

Heterogeneity in gaming ability can lead to "muddled" information, and therefore undesirable or unfair allocations.

How to alleviate this information loss?

Two possible games

1. **Signaling game** (Nash eq): Allocation...
   - is made by competitive market (colleges, banks)
   - depends on belief about type given observables
   - is "ex post optimal" given available info
   - Look for equilibrium outcome given fundamentals

     $\rightarrow$ When is information better or worse?

(Fischer Verrecchia 2000; Benabou Tirole 2006; Ali Benabou 2016; Gesche 2017)

2. **Commitment problem** (Stackelberg): Allocation...
   - is made by designer (Google, Amazon, government)
   - can be arbitrary function of observables
   - need not be "ex post optimal" given available info
   - Look for outcome maximizing designer's objective

     $\rightarrow$ How to improve relative to signaling outcome?

(Hardt, Megiddo, Papadimitriou, Wootters 2016; Hu, Immorlica, Vaughan 2019)

2-dimensional types

Agent's cost of taking action $\boldsymbol{a}$ depends on a two-dimensional type $(\boldsymbol{\eta}, \boldsymbol{\gamma})$:

- Dimension of interest: *natural action* η
  What you do in the absence of incentives
- Other dimension: *gaming ability* γ
  Marginal cost of moving away from natural action
  (Equivalently, marginal benefit from higher beliefs)

Examples —

- Schooling:
  - Action is SAT score
  - Natural action is how well student would do without "test prep"
  - Gaming ability is money to pay for tutoring, quality of test prep material, how much one cares about getting a high score
- Web search: Gaming ability is skill at (or ability to pay for) SEO, incentive to improve ranking, willingness to do shady stuff

Formalizing the Signaling Game

- Agent of type $(\eta, \gamma) \in \Theta$ chooses action $a \in \mathbb{R}$ at cost $C(a; \eta, \gamma)$
- Given action a, belief on agent's natural action η of $\hat{\eta} = \mathbb{E}[\eta|a]$
- Signaling value $s \cdot v(\hat{\eta})$ for continuously increasing function v
  - Agent prefers higher beliefs
  - s represents "stakes"
- Agent payoff

$$sv(\hat{\eta}) - C(a; \eta, \gamma)$$

Formalizing the Signaling Game

- Agent of type $(\eta, \gamma) \in \Theta$ chooses action $a \in \mathbb{R}$ at cost $C(a; \eta, \gamma)$
- Given action a, belief on agent's natural action η of $\hat{\eta} = \mathbb{E}[\eta|a]$
- Signaling value $s \cdot v(\hat{\eta})$ for continuously increasing function v
  - Agent prefers higher beliefs
  - s represents "stakes"
- Agent payoff

$$sv(\hat{\eta}) - C(a; \eta, \gamma)$$

Observations:

1. Allocation is implicit – we've only modeled induced agent payoff
2. Agent payoff (and, implicitly, allocation) is mechanical given belief

   $\rightarrow$ Commitment game will make explicit the allocation problem, and allocation need not be optimal ex post given belief
3. Continuously increasing value $\longleftrightarrow$ continuous (or noisy) allocation

Types and Costs

Natural action η, Gaming ability γ

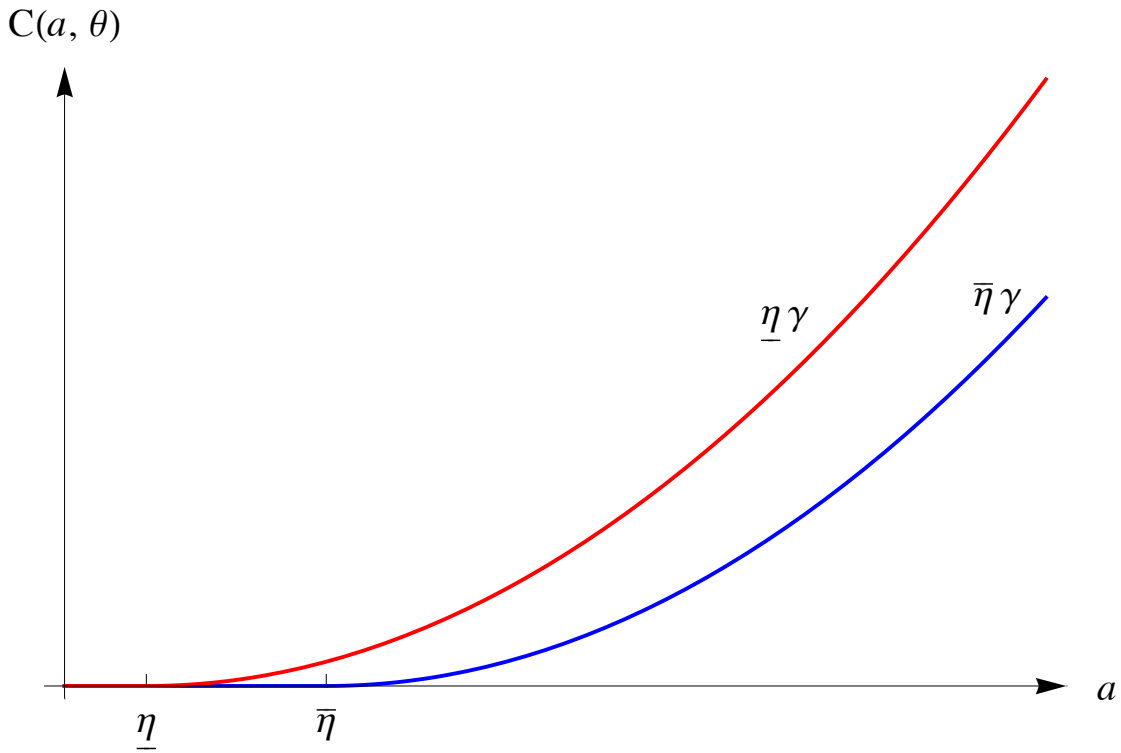
Assumptions on cost function $C(a; \eta, \gamma)$:

Types and Costs

Natural action η, Gaming ability γ

Assumptions on cost function $C(a; \eta, \gamma)$:

1. $C(a; \eta, \gamma) = 0$ for $a \leq \eta$
   - Ideal point is η, free downward deviations
2. C is differentiable; and for $a > \eta$, $C_{aa} > 0$
   - Costs are convex in actions
3. For $a > \eta$, $C_{a\eta} < 0$
   - Higher natural action $\Rightarrow$ lower MC of increasing a

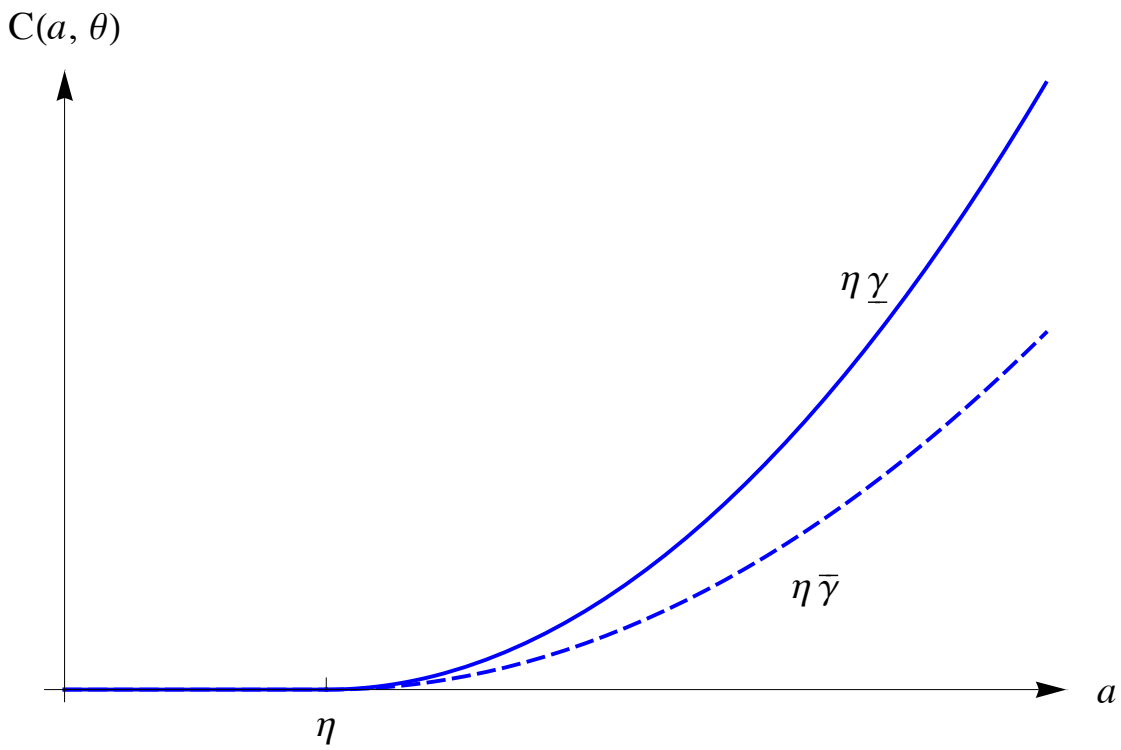


Types and Costs

Natural action η, Gaming ability γ

Assumptions on cost function $C(a; \eta, \gamma)$:

4. For $a > \eta$, $C_{a\gamma} < 0$
   - Higher gaming ability $\Rightarrow$ lower MC of increasing a



Types and Costs

Natural action η, Gaming ability γ

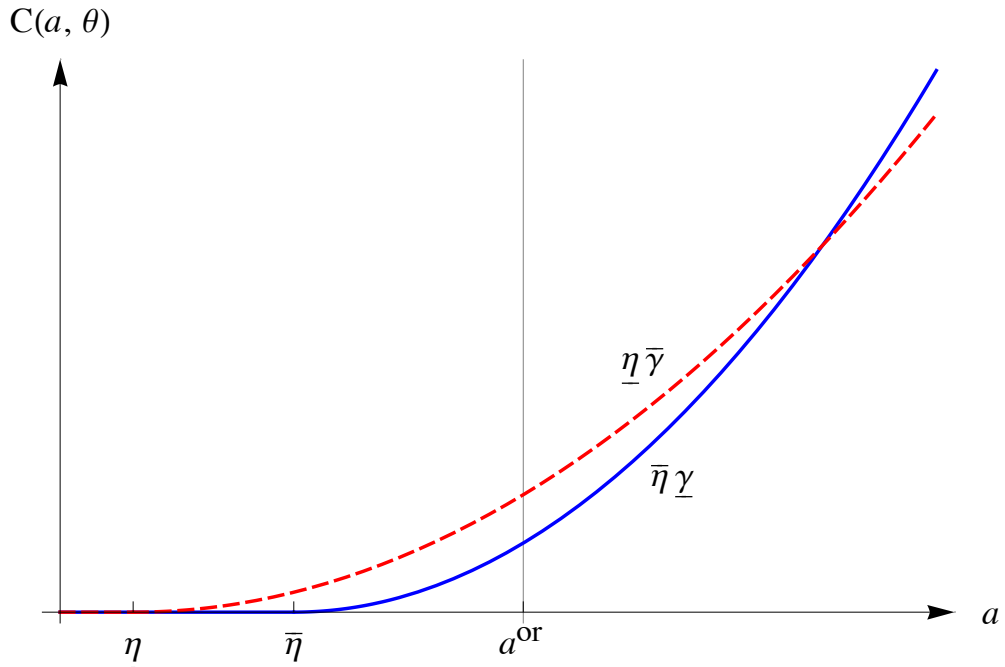
Assumptions on cost function $C(a;\eta,\gamma)$:

5. For any pair of cross types $(\underline{\eta},\overline{\gamma})$ and $(\overline{\eta},\underline{\gamma})$:

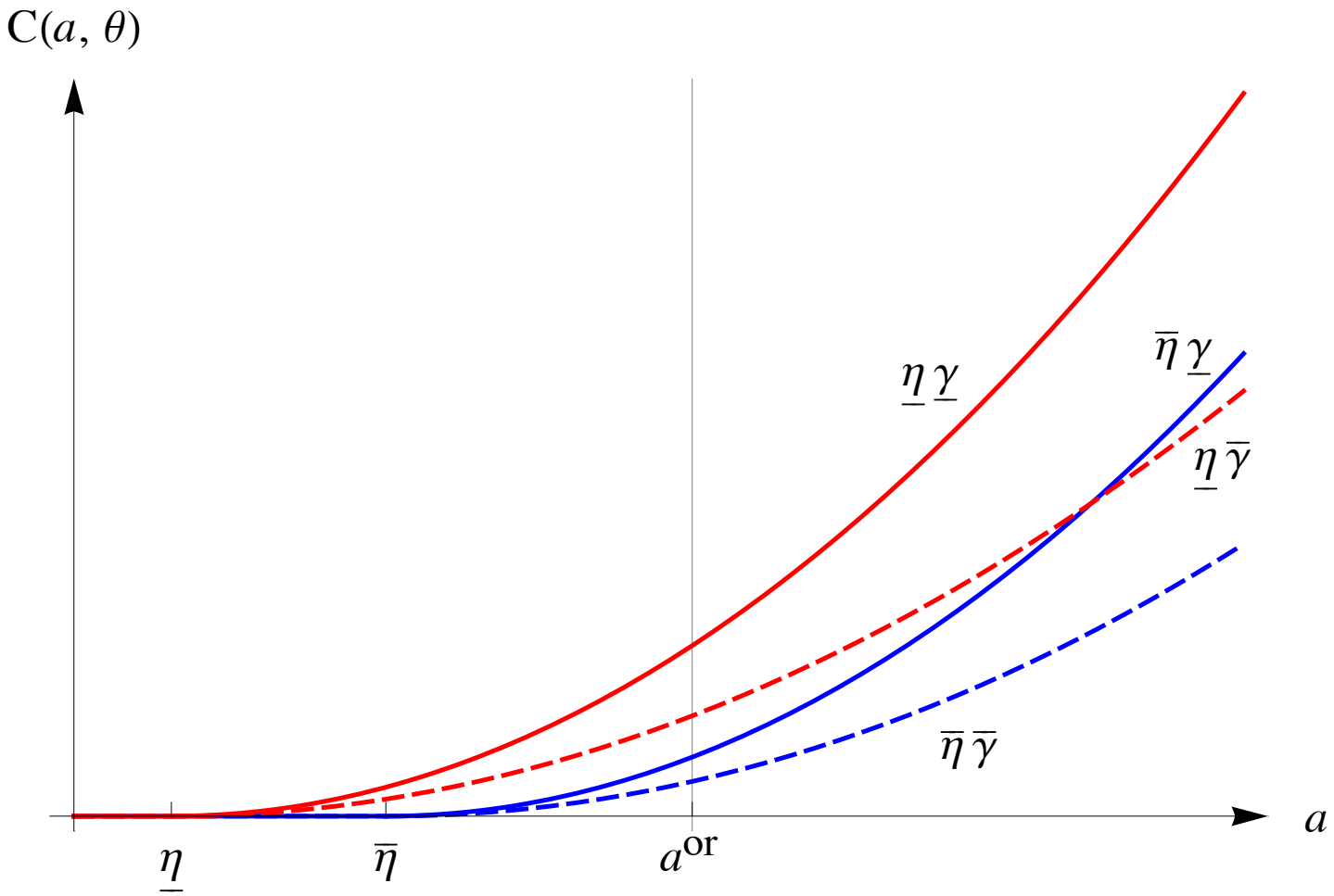
   $\frac{C_a(\cdot;\overline{\eta},\underline{\gamma})}{C_a(\cdot;\underline{\eta},\overline{\gamma})}$ is strictly increasing on $[\overline{\eta},\infty)$,

   and $\frac{C_a(\cdot;\overline{\eta},\underline{\gamma})}{C_a(\cdot;\underline{\eta},\overline{\gamma})} = 1$ at some **"order-reversing action"** $\boldsymbol{a^{or}} > \overline{\eta}$.

   - At low actions (below a^{or}), "the natural" $(\overline{\eta},\underline{\gamma})$ has lower MC
   - At high actions (above a^{or}), "the gamer" $(\underline{\eta},\overline{\gamma})$ has lower MC



Types and Costs



Leading example: $C(a;\eta,\gamma) = \begin{cases} \frac{(a-\eta)^2}{\gamma} & \text{if } a > \eta \\ 0 & \text{o/w} \end{cases}$

Completing definition of the signaling game

Equilibrium:

- Fix a joint distribution over types (η, γ)
- Look for Bayesian Nash equilibrium:
  - Market has beliefs $\hat{\eta}(a)$, Bayes consistent on-path
  - Agent of type (η, γ) maximizes $V(\hat{\eta}(a)) - C(a; \eta, \gamma)$ over choice of a

Completing definition of the signaling game

Equilibrium:

- Fix a joint distribution over types (η, γ)
- Look for Bayesian Nash equilibrium:
  - Market has beliefs $\hat{\eta}(a)$, Bayes consistent on-path
  - Agent of type (η, γ) maximizes $V(\hat{\eta}(a)) - C(a; \eta, \gamma)$ over choice of a
- Fully pooling equilibrium always exists
  - Free Downward Deviations $\Rightarrow$ can pool at lowest natural action
- We're interested in possibility of informative equilibria:
  Partially pooling, or separating on τ

Source of info loss

If agents all have the same gaming ability, then we have a separating eq

$\rightarrow$ Observers learn agent type perfectly from action

- Signaling game with single-crossing (Spence, 1973)
- There may be lots of wasteful effort ("rat race") but it washes out

Info loss caused by heterogeneous gaming ability

Source of info loss

First set of main results:

Higher stakes s in value function $sv(\hat{\eta})$ imply worse information

- 2×2 model: The set of eq at low stakes is Blackwell more informative than at high stakes, under weak set order

  For any eq at high s, we can find Blackwell more inf eq at low s

  For any eq at low s, we can find Blackwell less inf eq at high s
- General distributions / supports:
  - As $s \to 0$, approach full info
  - As $s \to \infty$, learn about η only through correlation with γ —

    If $\mathbb{E}[\eta|\gamma]$ non-increasing in γ, approach an uninformative limit
- Parametric results for "linear quadratic elliptical" LQE model

  Linear equilibria, with R^2 decreasing in s

Source of info loss

First set of main results:

Higher stakes s in value function $sv(\hat{\eta})$ imply worse information

- 2×2 model: The set of eq at low stakes is Blackwell more informative than at high stakes, under weak set order

  For any eq at high s, we can find Blackwell more inf eq at low s

  For any eq at low s, we can find Blackwell less inf eq at high s
- General distributions / supports:
  - As $s \to 0$, approach full info
  - As $s \to \infty$, learn about η only through correlation with γ —

    If $\mathbb{E}[\eta|\gamma]$ non-increasing in γ, approach an uninformative limit
- Parametric results for "linear quadratic elliptical" LQE model

  Linear equilibria, with R^2 decreasing in s

Higher incentives to manipulate $\Rightarrow$ more manipulation, worse information

So: Information quality is better the less the info is used

$\to$ "Goodhart's Law"

Interventions to improve information

How to improve equilibrium info:

- Reduce stakes – use info for fewer allocations
  - Currently, credit score used for loans, employment, insurance
  - Forbidding credit use in employment can improve info in loans

Interventions to improve information

How to improve equilibrium info:

- Reduce stakes – use info for fewer allocations
  - Currently, credit score used for loans, employment, insurance
  - Forbidding credit use in employment can improve info in loans

- Make mechanism "harder to manipulate"
  - Reinterpretation: lower s corresponds to higher manipulation costs

$$sv(\hat{\eta}) - C(a;\eta,\gamma) \iff v(\hat{\eta}) - \frac{C(a;\eta,\gamma)}{s}$$

  - Pagerank vs keyword analysis: incoming links hard to manipulate
  - Fair Isaac, Google: obscure details of algorithm
  - SAT prior to 1978-1980: don't reveal old test questions

Interventions to improve information

Transparency or not?

- Prior to 1978: SAT didn't reveal any old test Qs
- 1980: College Board starts selling books like "5 SATs", "10 SATs"
- Today: College Board releases free test prep with Khan Academy

Interventions to improve information

Transparency or not?

- Prior to 1978: SAT didn't reveal any old test Qs
  - Mid-70s: Kaplan advertises that they raise SAT scores by 100 points (Kaplan had built up private list of Qs from student reports)
- 1980: College Board starts selling books like "5 SATs", "10 SATs"
  - 2014, Khan Academy: 'We're thrilled to collaborate closely with the College Board to **level the playing field** by making truly world-class test-prep materials freely available to all students."
- Today: College Board releases free test prep with Khan Academy

Interventions to improve information

Transparency or not?

- Prior to 1978: SAT didn't reveal any old test Qs
  - Mid-70s: Kaplan advertises that they raise SAT scores by 100 points (Kaplan had built up private list of Qs from student reports)
- 1980: College Board starts selling books like "5 SATs", "10 SATs"
  - 2014, Khan Academy: 'We're thrilled to collaborate closely with the College Board to **level the playing field** by making truly world-class test-prep materials freely available to all students."
- Today: College Board releases free test prep with Khan Academy

What is the effect of revealing info on test / algorithm?

- Uniformly lowering manipulation costs $\Rightarrow$ bad for info
- "Leveling the playing field" (raising γ of low-γ types) $\Rightarrow$ maybe good
  - Less heterogeneity on γ leads to less info loss
  - Parametric LQE model: info varies inversely with σ_γ^2

Interventions to improve information

The commitment problem:

- Fair Isaac produces a credit score, banks decide on loans
  - Banks act competitively given info
  - If FICO score suggests I'm of borrower quality X, banks treat me as X
- Google or Amazon sets its own rankings
  - Info available to Amazon might suggest I have a product of quality X
  - Nothing stops Amazon from treating it as quality Y

If we can commit to an arbitrary mapping of action to allocation,
How would we distort signaling eq to improve info?
(Goal of "accuracy")

Interventions to improve information

The commitment problem:

- Fair Isaac produces a credit score, banks decide on loans
  - Banks act competitively given info
  - If FICO score suggests I'm of borrower quality X, banks treat me as X
- Google or Amazon sets its own rankings
  - Info available to Amazon might suggest I have a product of quality X
  - Nothing stops Amazon from treating it as quality Y

If we can commit to an arbitrary mapping of action to allocation,
How would we distort signaling eq to improve info?
(Goal of "accuracy")

Conclusion: Flatten the allocation rule, i.e., commit to put less weight on the manipulable data.

The commitment problem

I will now be explicit about designer's allocation problem

- Agents of type $(\eta, \gamma) \in \mathbb{R}^2$
- Designer wants to match agent allocation $y \in \mathbb{R}$ to natural action η:

$$\text{Utility} = -(y - \eta)^2$$

- Designer sets allocation $y = Y(x)$ based on observable (action) x
  - Assume linear allocation rule:

$$Y(x) = \beta x + \beta_0$$

- Agent chooses x based on type (η, γ) and allocation rule Y
  - Given linear allocation rule (β_0, β), assume agent response of

$$x = \eta + \gamma\beta$$

  - Optimal response for linear value, quadratic costs: $y - (x - \eta)^2/(2\gamma)$

Designer's best response

Say that when agents respond to allocation rule $\tilde{Y}(x) = \tilde{\beta}x + \tilde{\beta}_0$, designer's best linear estimate of η given x is $\hat{\eta}_{\tilde{\beta}}(x)$:

$$\hat{\eta}_{\tilde{\beta}}(x) = \hat{\beta}(\tilde{\beta})x + \hat{\beta}_0(\tilde{\beta})$$

Decomposition of welfare loss for allocation rule $Y(x) = \beta x + \beta_0$:

$$\text{Welfare Loss} = \underbrace{\mathbb{E}[(\hat{\eta}_\beta(x) - \eta)^2]}_{\text{Info loss of estimation of } \eta \text{ from } x} + \underbrace{\mathbb{E}[(Y(x) - \hat{\eta}_\beta(x))^2]}_{\text{Misallocation loss given estimation}}$$

Designer's best response

Say that when agents respond to allocation rule $\tilde{Y}(x) = \tilde{\beta}x + \tilde{\beta}_0$, designer's best linear estimate of η given x is $\hat{\eta}_{\tilde{\beta}}(x)$:

$$\hat{\eta}_{\tilde{\beta}}(x) = \hat{\beta}(\tilde{\beta})x + \hat{\beta}_0(\tilde{\beta})$$

Decomposition of welfare loss for allocation rule $Y(x) = \beta x + \beta_0$:

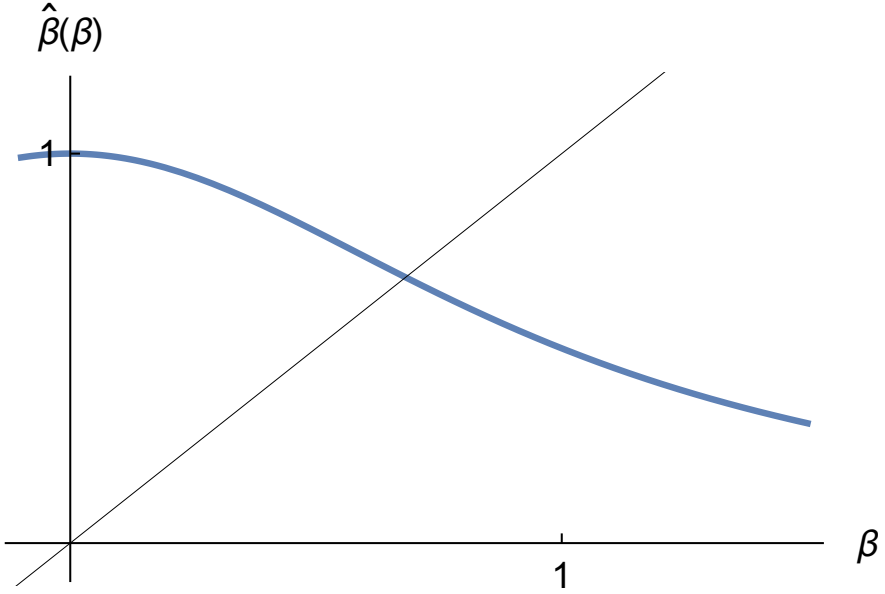
$$\text{Welfare Loss} = \underbrace{\mathbb{E}[(\hat{\eta}_\beta(x) - \eta)^2]}_{\text{Info loss of estimation of } \eta \text{ from } x} + \underbrace{\mathbb{E}[(Y(x) - \hat{\eta}_\beta(x))^2]}_{\text{Misallocation loss given estimation}}$$

Fixing agent's behavior (responding to $\tilde{Y}$ with coef $\tilde{\beta}$), designer's **best response** is to set $Y(x) = \hat{\eta}_{\tilde{\beta}}(x)$ with coef $\hat{\beta}(\tilde{\beta})$.

But then agent's behavior changes...

Designer best response $\hat{\beta}(\cdot)$

Welfare Loss = [Info loss of estimation] + [Misallocation loss given estimation]



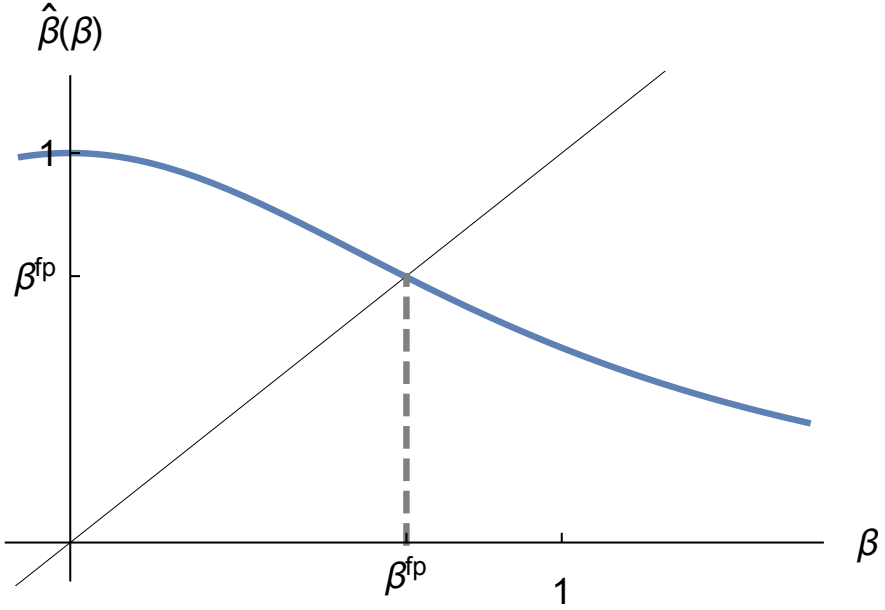
Loosely speaking, under policy $Y(x) = \beta x + \beta_0$...

Higher $\beta \Rightarrow$ More agent manipulation $\Rightarrow$

- Observable x less informative about type η
- Larger Info loss of estimation
- Lower $\hat{\beta}(\beta)$ from regressing η on x

Designer best response $\hat{\beta}(\cdot)$

Welfare Loss = [Info loss of estimation] + [Misallocation loss given estimation]

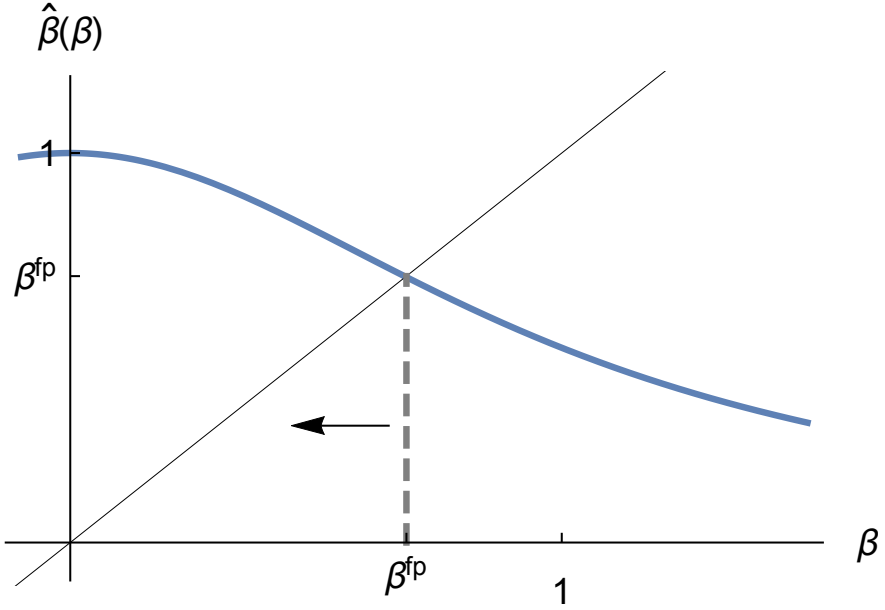


Policy $\beta = \beta^{\text{fp}}$: Gather data, best respond, gather data, ... $\rightarrow$ fixed point

- Designer best response $\hat{\beta}(\beta^{\text{fp}}) = \beta^{\text{fp}}$
- Some Info loss of estimation
- Zero Misallocation loss

Designer best response $\hat{\beta}(\cdot)$

Welfare Loss = [Info loss of estimation] + [Misallocation loss given estimation]



Starting from $\beta = \beta^{\text{fp}}$, misallocation loss $= 0$. Reducing β yields...

- First-order benefit from reducing info loss
- Second-order harm from increasing misallocation loss

  $\Rightarrow$ First-order benefit

Optimal policy

Policy $Y(x) = \beta x + \beta_0$

The commitment optimal β^* is in fact less than the fixed point β^{fp}.

Proposition

There is an optimum $\beta^* > 0$ such that for any $\beta^{\text{fp}} > 0$, $\beta^* \leq \beta^{\text{fp}}$ with $\beta^* < \beta^{\text{fp}}$ if $\rho \notin \{-1, 1\}$.

Intuition: First-order benefit from reducing β starting from β^{fp}

Actual proof: Show that the maximum of a quartic is the smallest positive extreme point

So – starting from fixed point, improve information by flattening the allocation rule, i.e., putting less weight on the manipulable predictor x.

How to generalize to other allocation problems?

Linear model: Use $Y(x) = \beta x + \beta_0$ with β below fixed point value

What does it mean to "flatten" an allocation rule, or "put less weight on" a predictor, in a general nonlinear problem?

- Highly dimensional data $\vec{x}$ including manipulable component x_m
- ML algorithm to predict η from $\vec{x}$
- Assign allocation y based on ML prediction of η

$\rightarrow$ How to make ML prediction "less sensitive" to x_m?

One idea: add artificial noise to x_m in estimation $\Rightarrow$ "attenuation bias"