Optimal Incentives under Moral Hazard: From Theory to Practice

George Georgiadis

Michael Powell

Northwestern Kellogg
Motivation

Imagine you have to design an employee performance pay plan.

- If you know all payoff-relevant parameters (i.e., agent preferences, production function, etc), you can find optimal contract (in principle).
- Otherwise, agency theory gives us guiding principles (trade-offs, CS)

This paper: How to improve an existing PPP?

1. What information do you need?
2. And how should you use that information?
Motivation

- Imagine you have to design an employee performance pay plan.
- If you know all payoff-relevant parameters (*i.e.*, agent preferences, production function, etc), you can find optimal contract (in principle).
- Otherwise, agency theory gives us guiding principles (trade-offs, CS)

This paper: How to improve an existing PPP?

1. What information do you need?
2. And how should you use that information?
Imagine you have to design an employee performance pay plan.

If you know all payoff-relevant parameters (i.e., agent preferences, production function, etc), you can find optimal contract (in principle).

Otherwise, agency theory gives us guiding principles (trade-offs, CS)

**This paper:** How to improve an existing PPP?

1. What information do you need?
2. And how should you use that information?
Framework: Static agency model with a risk-averse agent

- Principal knows only distribution of output following $w_A(\cdot)$ and $w_B(\cdot)$.
- Goal: Find a new contract that raises profits as much as possible.

Key Lemma:
If the principal takes a stance on the agent’s marginal utility for money, she can predict the distribution of output corresponding to any contract.

- Then, the principal can find an optimal perturbation.

Application using real-effort experiment of DellaVigna and Pope (‘17)

Predictions: Use any pair of treatments to predict the other 5

Counterfactuals: Estimate model and evaluate optimal perturbations
Preview

- **Framework**: Static agency model with a risk-averse agent
  - Principal knows only distribution of output following $w_A(\cdot)$ and $w_B(\cdot)$.
  - **Goal**: Find a new contract that raises profits as much as possible.

**Key Lemma:**
If the principal *takes a stance on* the agent’s marginal utility for money, she can predict the distribution of output corresponding to *any* contract.

- Then, the principal can find an *optimal perturbation*.

- Application using real-effort experiment of DellaVigna and Pope (’17)
  1. **Predictions**: Use any pair of treatments to predict the other 5
  2. **Counterfactuals**: Estimate model and evaluate optimal perturbations
Framework: Static agency model with a risk-averse agent

- Principal knows only distribution of output following $w_A(\cdot)$ and $w_B(\cdot)$.
- Goal: Find a new contract that raises profits as much as possible.

Key Lemma:
If the principal takes a stance on the agent’s marginal utility for money, she can predict the distribution of output corresponding to any contract.

Then, the principal can find an optimal perturbation.

Application using real-effort experiment of DellaVigna and Pope (’17)

1. Predictions: Use any pair of treatments to predict the other 5
2. Counterfactuals: Estimate model and evaluate optimal perturbations
Preview

*Framework:* Static agency model with a risk-averse agent

- Principal knows *only* distribution of output following $w_A(\cdot)$ and $w_B(\cdot)$.
- *Goal:* Find a new contract that raises profits as much as possible.

**Key Lemma:**
If the principal *takes a stance on* the agent’s marginal utility for money, she can predict the distribution of output corresponding to *any* contract.

- Then, the principal can find an *optimal perturbation*.

- **Application using real-effort experiment of DellaVigna and Pope ('17)**
  1. *Predictions:* Use any pair of treatments to predict the other 5
  2. *Counterfactuals:* Estimate model and evaluate optimal perturbations
Related Literature

- **Agency problems — Theory:**
  - Mirrlees (1976), Holmström (1979), ...
  - Gibbons (1998), Murphy (1999), ...

- **Agency problems — Empirics:**
  - Lazear (2000), Shearer (2004), Bandiera et al. (2007, 2009), ...
  - Chiappori & Salanie (2002), Prendergast (2002), ...

- **Sufficient statistics:**
  - Monopoly pricing: Lerner (1934), Tirole (1988), ...
  - Optimal taxation: Saez (2001), Golosov et al. (2014), Chetty (2009), ..
Environment

Model

- Principal-agent model with the following timing:
  1. Principal offers a contract $w(\cdot)$.
  2. Agent observes $w(\cdot)$ and chooses effort $a(w) \in \mathbb{R}$.
  3. Output $x \sim f(\cdot|a(w))$ and payoffs are realized. (Normalize $\mathbb{E}[x|a] = a$.)

- Preferences:
  - Agent’s utility: $\int v(w(x))f(x|a)dx - c(a)$
  - Principal’s profit: $\pi(w) := ma(w) - \int w(x)f(x|a)dx$.

- Information:
  - Agent knows all payoff-relevant parameters
  - Principal knows (only) $f(\cdot|a(w_A))$, $f(\cdot|a(w_B))$, and
    
    $$f_a(\cdot|a(w_A)) \approx \frac{f(\cdot|a(w_B)) - f(\cdot|a(w_A))}{a(w_B) - a(w_A)}$$

Georgiadis and Powell
Optimal Incentives under Moral Hazard
Northwestern Kellogg
Model

- Principal-agent model with the following timing:
  1. Principal offers a contract \( w(\cdot) \).
  2. Agent observes \( w(\cdot) \) and chooses effort \( a(w) \in \mathbb{R} \).
  3. Output \( x \sim f(\cdot|a(w)) \) and payoffs are realized. (Normalize \( \mathbb{E}[x|a] = a \).)

- Preferences:
  - Agent’s utility: \( \int v(w(x))f(x|a)dx - c(a) \)
  - Principal’s profit: \( \pi(w) := ma(w) - \int w(x)f(x|a)dx \).

- Information:
  - Agent knows all payoff-relevant parameters
  - Principal knows (only) \( f(\cdot|a(w_A)), f(\cdot|a(w_B)) \), and
  \[
  f_a(\cdot|a(w_A)) = \frac{f(\cdot|a(w_B)) - f(\cdot|a(w_A))}{a(w_B) - a(w_A)}
  \]
Model

- Principal-agent model with the following timing:
  1. Principal offers a contract $w(\cdot)$.
  2. Agent observes $w(\cdot)$ and chooses effort $a(w) \in \mathbb{R}$.
  3. Output $x \sim f(\cdot|a(w))$ and payoffs are realized. (Normalize $\mathbb{E}[x|a] = a$.)

- Preferences:
  - Agent’s utility: $\int v(w(x)) f(x|a) dx - c(a)$
  - Principal’s profit: $\pi(w) := ma(w) - \int w(x) f(x|a) dx$.

- Information:
  - Agent knows all payoff-relevant parameters.
  - Principal knows (only) $f(\cdot|a(w_A))$, $f(\cdot|a(w_B))$, and
  
  $$f_a(\cdot|a(w_A)) \approx \frac{f(\cdot|a(w_B)) - f(\cdot|a(w_A))}{a(w_B) - a(w_A)}$$

Georgiadis and Powell
Optimal Incentives under Moral Hazard
Northwestern Kellogg
The Canonical Principal-Agent Problem

- In the canonical formulation (Holmström, 1979), the principal solves

\[
\max_{w(\cdot), a} \int [mx - w(x)] f(x|a)dx \\
\text{s.t. } \int v(w(x))f(x|a)dx - c(a) \geq u
\]

(IR)

\[
a \in \arg \max_{\tilde{a}} \left\{ \int v(w(x))f(x|\tilde{a})dx - c(\tilde{a}) \right\}
\]

(IC)

- To do so, she must know \(v(\cdot), u, c(a),\) and \(f(\cdot|a)\) for all \(a\).

- In our setting, only knows \(f(\cdot|a(w_i))\) for \(i \in \{A, B\}\), and \(f_a(\cdot|a(w_A))\)

- **Notations:**

\[
\hat{a} := a(w_A) \quad \hat{f} := f(\cdot|a(w_A)) \quad \text{and} \quad \hat{f}_a := f_a(\cdot|a(w_A))
\]
The Canonical Principal-Agent Problem

- In the canonical formulation (Holmström, 1979), the principal solves

\[
\max_{w(\cdot), a} \int [mx - w(x)] f(x|a) \, dx \\
\text{s.t. } \int v(w(x)) f(x|a) \, dx - c(a) \geq u
\]  

(IR)

\[
a \in \arg \max \int v(w(x)) f(x|\tilde{a}) \, dx - c(\tilde{a})
\]  

(IC)

- To do so, she must know \( v(\cdot), u, c(a), \) and \( f(\cdot|a) \) for all \( a \).

- In our setting, only knows \( f(\cdot|a(w_i)) \) for \( i \in \{A, B\} \), and \( f_a(\cdot|a(w_A)) \)

- Notations:

\[
\hat{a} := a(w_A) \, , \, \hat{f} := f(\cdot|a(w_a)) \, , \, \text{and} \, \hat{f}_a := f_a(\cdot|a(w_a))
\]
The Canonical Principal-Agent Problem

- In the canonical formulation (Holmström, 1979), the principal solves

\[
\max_{w(\cdot), a} \int [mx - w(x)] f(x|a) \, dx \\
\text{s.t.} \int v(w(x)) f(x|a) \, dx - c(a) \geq u
\]

(\text{IR})

\[
a \in \arg \max \left\{ \int v(w(x)) f(x|\tilde{a}) \, dx - c(\tilde{a}) \right\}
\]

(\text{IC})

- To do so, she must know \( v(\cdot), u, c(a), \) and \( f(\cdot|a) \) for all \( a \).

- In our setting, only knows \( f(\cdot|a(w_i)) \) for \( i \in \{A, B\} \), and \( f_a(\cdot|a(w_A)) \)

- **Notations:**

\[
\tilde{a} := a(w_A), \quad \tilde{f} := f(\cdot|a(w_a)), \quad \text{and} \quad \tilde{f}_a := f_a(\cdot|a(w_a))
\]
Agent’s Problem

- Assume optimal effort $a(w)$ satisfies the first-order condition
  \[ \int v(w(x)) f_a(x|a(w)) \, dx = c'(a(w)) \]  
  \hspace{1cm} \text{(IC)}

- Suppose $w(\cdot)$ is replaced by (some) contract $w(\cdot) + \theta t(\cdot)$, $\theta$ small.

- Define the directional (Gateaux) derivative
  \[ D a(w, t) := \frac{d a(w + \theta t)}{d \theta} \bigg|_{\theta=0}, \]
  interpreted as the MC of $a$ when $w$ perturbed in the direction of $w + t$.

- Assume the principal knows
  \[ D a(w_A, w_B - w_A) \approx a(w_B) - a(w_A). \]

- Implicitly assuming $\|w_B - w_A\| \approx 0$ and $|a(w_B) - a(w_A)| \approx 0$. 

Georgiadis and Powell
Optimal Incentives under Moral Hazard
Northwestern Kellogg
Agent’s Problem

- Assume optimal effort $a(w)$ satisfies the first-order condition
  \[ \int v(w(x)) f_a(x|a(w)) \, dx = c'(a(w)) \tag{IC} \]
- Suppose $w(\cdot)$ is replaced by (some) contract $w(\cdot) + \theta t(\cdot)$, $\theta$ small.
- Define the directional (Gateaux) derivative
  \[ \mathcal{D}a(w, t) := \left. \frac{da(w + \theta t)}{d\theta} \right|_{\theta=0}, \]
  interpreted as the MC of $a$ when $w$ perturbed in the direction of $w + t$.
- Assume the principal knows
  \[ \mathcal{D}a(w_A, w_B - w_A) \simeq a(w_B) - a(w_A). \]
- Implicitly assuming $\|w_B - w_A\| \simeq 0$ and $|a(w_B) - a(w_A)| \simeq 0$.
Agent’s Problem

- Assume optimal effort $a(w)$ satisfies the first-order condition

$$\int \nu(w(x)) f_a(x|a(w)) dx = c'(a(w)) \quad (IC)$$

- Suppose $w(\cdot)$ is replaced by (some) contract $w(\cdot) + \theta t(\cdot)$, $\theta$ small.

- Define the directional (Gateaux) derivative

$$D a(w, t) := \frac{d a(w + \theta t)}{d \theta} \bigg|_{\theta = 0},$$

interpreted as the MC of $a$ when $w$ perturbed in the direction of $w + t$.

- Assume the principal knows

$$D a(w_A, w_B - w_A) \simeq a(w_B) - a(w_A).$$

- Implicitly assuming $\|w_B - w_A\| \simeq 0$ and $|a(w_B) - a(w_A)| \simeq 0$
**Principal’s Problem**

- If $w(\cdot)$ is replaced by (some) $w(\cdot) + \theta t(\cdot)$, then the principal’s profit

  $$\pi(w + \theta t) \approx \pi(w) + \theta D\pi(w, t),$$

  where $D\pi(w, t)$ is the derivative of $\pi(w)$ in direction of $w + t$, and

  $$D\pi(w, t) := \frac{d\pi(w + \theta t)}{d\theta} \bigg|_{\theta=0} = \left( m - \int wf_a dx \right) Da(w, t) - \int t f dx$$

- Assume the principal’s goal is to maximize $D\pi(w_A, t)$ subject to $w_A + \theta t$ giving the agent at least as much utility as $w_A$.

- Using (IC), this (participation) constraint can be rewritten as

  $$\int tv'(w_A) \tilde{f} dx \geq 0$$

- **Info Requirements**: $Da(w_A, t)$ for all $t$ & marg. utility function $v'(\cdot)$
Principal’s Problem

- If \( w(\cdot) \) is replaced by (some) \( w(\cdot) + \theta t(\cdot) \), then the principal’s profit

\[
\pi(w + \theta t) \simeq \pi(w) + \theta D\pi(w, t),
\]

where \( D\pi(w, t) \) is the derivative of \( \pi(w) \) in direction of \( w + t \), and

\[
D\pi(w, t) := \frac{d\pi(w + \theta t)}{d\theta} \bigg|_{\theta=0} = \left( m - \int wf_a dx \right) Da(w, t) - \int tf dx
\]

- Assume the principal’s goal is to maximize \( D\pi(w_A, t) \) subject to \( w_A + \theta t \) giving the agent at least as much utility as \( w_A \).

- Using (IC), this (participation) constraint can be rewritten as

\[
\int tv'(w_A)\hat{f} dx \geq 0
\]

- Info Requirements: \( Da(w_A, t) \) for all \( t \) & marg. utility function \( v'(\cdot) \)
Principal’s Problem

- If $w(\cdot)$ is replaced by (some) $w(\cdot) + \theta t(\cdot)$, then the principal’s profit
  \[ \pi(w + \theta t) \approx \pi(w) + \theta \mathcal{D} \pi(w, t), \]
  where $\mathcal{D} \pi(w, t)$ is the derivative of $\pi(w)$ in direction of $w + t$, and
  \[ \mathcal{D} \pi(w, t) := \frac{d\pi(w + \theta t)}{d\theta} \bigg|_{\theta=0} = \left( m - \int w f_a dx \right) \mathcal{D} a(w, t) - \int tf dx \]

- Assume the principal’s goal is to maximize $\mathcal{D} \pi(w_A, t)$ subject to $w_A + \theta t$ giving the agent at least as much utility as $w_A$.

- Using (IC), this (participation) constraint can be rewritten as
  \[ \int tv'(w_A) \widehat{f} dx \geq 0 \]

- Info Requirements: $\mathcal{D} a(w_A, t)$ for all $t$ & marg. utility function $v'(\cdot)$
Principal’s Problem

- If $w(\cdot)$ is replaced by (some) $w(\cdot) + \theta t(\cdot)$, then the principal’s profit
  \[ \pi(w + \theta t) \simeq \pi(w) + \theta D\pi(w, t), \]
  where $D\pi(w, t)$ is the derivative of $\pi(w)$ in direction of $w + t$, and
  \[ D\pi(w, t) := \frac{d\pi(w + \theta t)}{d\theta} \bigg|_{\theta=0} = \left( m - \int wf_{a}dx \right) Da(w, t) - \int tfdx \]

- Assume the principal’s goal is to maximize $D\pi(w_{A}, t)$ subject to $w_{A} + \theta t$ giving the agent at least as much utility as $w_{A}$.

- Using (IC), this (participation) constraint can be rewritten as
  \[ \int tv'(w_{A})\hat{f}dx \geq 0 \]

- **Info Requirements**: $Da(w_{A}, t)$ for all $t$ & marg. utility function $v'(\cdot)$
Simplifying the Informational Requirements

- Using (IC), we can write $\mathcal{D}a(w, t)$ in terms of primitives as

$$\mathcal{D}a(w, t) = \frac{\int tv'(w)f_a dx}{c''(a(w)) - \int v(w)f_{aa} dx}$$

**Remark 1.** For any (upper semi-continuous) $t$:

$$\mathcal{D}a(w_A, t) = \frac{\mathcal{D}a(w_A, w_B - w_A)}{\int (w_B - w_A)v'(w_A)\widehat{f}_a dx} \left( \frac{\int tv'(w_A)\widehat{f}_a dx}{\mathcal{D}M(w_A, t)} \right)$$

- Perturbation leads to a change in the agent's marginal incentives, $\mathcal{D}M(w_A, t)$, which is predictable given $v'$ and $\widehat{f}_a$. Locally,

$$\mathcal{D}a(w_A, t) = C \times \mathcal{D}M(w_A, t)$$

where $C = \frac{\mathcal{D}a(w_A, w_B - w_A)}{\mathcal{D}M(w_A, w_B - w_A)}$.

- If the principal takes a stance on $v'$, she can predict $\mathcal{D}a(w_A, t)$ $\forall t$. 
Simplifying the Informational Requirements

- Using (IC), we can write $D_a(w, t)$ in terms of primitives as
  \[ D_a(w, t) = \frac{\int tv'(w)f_a dw}{c''(a(w)) - \int v(w)f_{aa} dw} \]

Remark 1. For any (upper semi-continuous) $t$:

\[ D_a(w_A, t) = \frac{D_a(w_A, w_B - w_A)}{\int (w_B - w_A)v'(w_A)\hat{f}_a dw} \left( \int tv'(w_A)\hat{f}_a dw \right) / DM(w_A, t) \]

- Perturbation leads to a change in the agent’s marginal incentives, $DM(w_A, t)$, which is predictable given $v'$ and $\hat{f}_a$. Locally,
  \[ D_a(w_A, t) = C \times DM(w_A, t) \text{, where } C = \frac{D_a(w_A, w_B - w_A)}{DM(w_A, w_B - w_A)}. \]

- If the principal takes a stance on $v'$, she can predict $D_a(w_A, t)$ for all $t$. 
Simplifying the Informational Requirements

Using (IC), we can write $\mathcal{D}a(w, t)$ in terms of primitives as

$$\mathcal{D}a(w, t) = \frac{\int tv'(w)f_a dx}{c''(a(w)) - \int v(w)f_{aa} dx}$$

Remark 1. For any (upper semi-continuous) $t$:

$$\mathcal{D}a(w_A, t) = \frac{\mathcal{D}a(w_A, w_B - w_A)}{\int (w_B - w_A)v'(w_A)\tilde{f}_a dx} \underbrace{\int tv'(w_A)\tilde{f}_a dx}_{\mathcal{D}M(w_A, t)}$$

Perturbation leads to a change in the agent’s marginal incentives, $\mathcal{D}M(w_A, t)$, which is predictable given $v'$ and $\tilde{f}_a$. Locally,

$$\mathcal{D}a(w_A, t) = C \times \mathcal{D}M(w_A, t), \text{ where } C = \frac{\mathcal{D}a(w_A, w_B - w_A)}{\mathcal{D}M(w_A, w_B - w_A)}.$$ 

If the principal takes a stance on $v'$, she can predict $\mathcal{D}a(w_A, t) \forall t.$
Principal’s Problem (Cont’d)

- The principal solves

$$\max_{t \in \mathcal{C}} \mu \int tv'(w_A)\hat{f}_a \, dx - \int t\hat{f} \, dx$$

s.t. $$\int tv'(w_A)\hat{f} \, dx \geq 0$$

$$\int |t|^p \, dx \leq 1$$

where $$p \in \{1, 2, \ldots\}$$ normalizes the length of $$t$$.

- Problem is convex, so it can be solved using standard techniques.

  - Necessary & sufficient condition for $$w_A$$ to be optimal
  - Opt. Perturbation: Replace $$w_A$$ with $$w \equiv w_A + \theta t$$ for some $$\theta > 0$$ small
Principal’s Problem (Cont’d)

- The principal solves
  \[
  \max_{t \text{ u.s.c}} \mu \int t v'(w_A) \hat{f}_a \, dx - \int t \hat{f} \, dx \\
  \text{s.t } \int t v'(w_A) \hat{f} \, dx \geq 0 \\
  \int |t|^p \, dx \leq 1
  \]
  where \( p \in \{1, 2, \ldots\} \) normalizes the length of \( t \).

Problem is convex, so it can be solved using standard techniques.

- Necessary & sufficient condition for \( w_A \) to be optimal
- Opt. Perturbation: Replace \( w_A \) with \( w = w_A + \theta t \) for some \( \theta > 0 \) small
Principal’s Problem (Cont’d)

- The principal solves

\[
\max_{t \text{ u.s.c}} \mu \int tv'(w_A)\hat{f}a dx - \int t\hat{f} dx \\
\text{s.t} \int tv'(w_A)\hat{f} dx \geq 0 \\
\int |t|^p dx \leq 1
\]

where \( p \in \{1, 2, \ldots\} \) normalizes the length of \( t \).

- Problem is convex, so it can be solved using standard techniques.
  1. Necessary & sufficient condition for \( w_A \) to be optimal
  2. Opt. Perturbation: Replace \( w_A \) with \( w = w_A + \theta t \) for some \( \theta > 0 \) small
Principal’s Problem (Cont’d)

- The principal solves

\[
\max_{t \ u.s.c \ \mu} \int tv'(w_A) \tilde{f_a} dx - \int t \tilde{f} dx \\
\text{s.t. } \int tv'(w_A) \tilde{f} dx \geq 0 \\
\int |t|^p dx \leq 1
\]

where \( p \in \{1, 2, \ldots\} \) normalizes the length of \( t \).

- Problem is convex, so it can be solved using standard techniques.
  
  1. Necessary & sufficient condition for \( w_A \) to be optimal
  2. Opt. Perturbation: Replace \( w_A \) with \( w \equiv w_A + \theta t \) for some \( \theta > 0 \) small
Non-Local Perturbations

Goal: Develop algorithm for finding optimal non-local perturbations

A.1. For all $a$ in some interval that contains $\tilde{a}$, $f_a(\cdot|a) \equiv \tilde{f}_a$

- Hence, the marginal incentive of effort corresponding to $w$,

$$M(w) = \int v(w)\tilde{f}_a dx$$

does not depend on $a$ itself – agent’s FOC: $M(w) = c'(a)$

A.2. For any $w$, effort and marginal incentives are related by

$$\log a(w) = \beta + \epsilon \log M(w),$$

where $\beta$ and $\epsilon$ estimated using A-B test data and assumed $v'($.

- Implicitly assuming the agent has isoelastic cost function.
Non-Local Perturbations

- **Goal**: Develop algorithm for finding optimal *non-local* perturbations

A.1. For all \(a\) in some interval that contains \(\hat{a}\), \(f_a(\cdot | a) \equiv \hat{f}_a\)

- Hence, the marginal incentive of effort corresponding to \(w\),

\[
M(w) = \int v(w)\hat{f}_a \, dx
\]

does not depend on \(a\) itself – agent’s FOC: \(M(w) = c'(a)\)

A.2. For any \(w\), effort and marginal incentives are related by

\[
\log a(w) = \beta + \epsilon \log M(w),
\]

where \(\beta\) and \(\epsilon\) estimated using A-B test data and assumed \(v'(\cdot)\)

- Implicitly assuming the agent has isoelastic cost function.
Non-Local Perturbations

- **Goal:** Develop algorithm for finding optimal *non-local* perturbations

A.1. For all $a$ in some interval that contains $\hat{a}$, $f_a(\cdot|a) \equiv \hat{f}_a$

- Hence, the marginal incentive of effort corresponding to $w$,

$$M(w) = \int v(w) \hat{f}_a dx$$

does not depend on $a$ itself – *agent’s FOC:* $M(w) = c'(a)$

A.2. For any $w$, effort and marginal incentives are related by

$$\log a(w) = \beta + \epsilon \log M(w),$$

where $\beta$ and $\epsilon$ estimated using A-B test data and assumed $v'(\cdot)$

- Implicitly assuming the agent has isoelastic cost function.
Non-Local Perturbations

- **Goal:** Develop algorithm for finding optimal *non-local* perturbations

A.1. For all $a$ in some interval that contains $\widehat{a}$, $f_a(\cdot|a) \equiv \widehat{f_a}$

- Hence, the marginal incentive of effort corresponding to $w$,

$$M(w) = \int v(w)\widehat{f_a}dx$$

does not depend on $a$ itself – agent’s FOC: $M(w) = c'(a)$

A.2. For any $w$, effort and marginal incentives are related by

$$\log a(w) = \beta + \epsilon \log M(w),$$

where $\beta$ and $\epsilon$ estimated using A-B test data and assumed $v'($)

- Implicitly assuming the agent has isoelastic cost function.
Towards an Optimal non-local Perturbation

Claim: Principal should solve

$$\max_{w(\cdot), \Delta a} \ m(\hat{a} + \Delta a) - \int w(\hat{f} + \Delta a\hat{f}_a)$$  \hspace{1cm} (P)

s.t. \hspace{1cm} \int v(w)\hat{f}_a = \left(\frac{\hat{a} + \Delta a}{\hat{a}}\right)^{1/\epsilon} \int v(w_A)\hat{f}_a \hspace{1cm} (IC)

$$\int v(w)(\hat{f} + \Delta a\hat{f}_a) \geq \int v(w_A)(\hat{f} + \Delta a\hat{f}_a) \hspace{1cm} (IR)$$

- Suppose \( a(w) = \hat{a} + \Delta a \). Using a first-order approximation:

  $$f(\cdot|\hat{a} + \Delta a) \approx \hat{f} + \Delta a\hat{f}_a \quad \text{and} \quad c(\hat{a} + \Delta a) \approx c(\hat{a}) + \Delta a \int v(w_A)\hat{f}_a$$

- It follows from \( \log a(w) = \beta + \epsilon \log M(w) \) that \( w \) must satisfy (IC).

- Constraint that \( w \) gives at least as much utility as \( w_A \):

  $$\int v(w(x))f(x|\hat{a} + \Delta a) - c(\hat{a} + \Delta a) \geq \int v(w_A)\hat{f} - c(\hat{a}) \quad \Rightarrow (IR)$$
Towards an Optimal non-local Perturbation

Claim: Principal should solve

\[
\max_{\mathcal{w}(\cdot), \Delta a} \mathcal{m}(\hat{a} + \Delta a) - \int \mathcal{w}(\hat{f} + \Delta a \hat{f}_a) \]

s.t.

\[
\int \mathcal{v}(w) \hat{f}_a = \left( \frac{\hat{a} + \Delta a}{\hat{a}} \right)^{1/\epsilon} \int \mathcal{v}(w_A) \hat{f}_a \quad (\text{IC})
\]

\[
\int \mathcal{v}(w) \left( \hat{f} + \Delta a \hat{f}_a \right) \geq \int \mathcal{v}(w_A) \left( \hat{f} + \Delta a \hat{f}_a \right) \quad (\text{IR})
\]

Suppose \( a(w) = \hat{a} + \Delta a \). Using a first-order approximation:

\[
f(\cdot|\hat{a} + \Delta a) \simeq \hat{f} + \Delta a \hat{f}_a \quad \text{and} \quad c(\hat{a} + \Delta a) \simeq c(\hat{a}) + \Delta a \int \mathcal{v}(w_A) \hat{f}_a
\]

It follows from \( \log a(w) = \beta + \epsilon \log M(w) \) that \( w \) must satisfy (IC).

Constraint that \( w \) gives at least as much utility as \( w_A \):

\[
\int \mathcal{v}(w(x)) f(x|\hat{a} + \Delta a) - c(\hat{a} + \Delta a) \geq \int \mathcal{v}(w_A) \hat{f} - c(\hat{a}) \quad \Rightarrow (\text{IR})
\]
Towards an Optimal non-local Perturbation

Claim: Principal should solve

$$\max_{w(\cdot), \Delta a} \ m(\hat{a} + \Delta a) - \int w(\hat{f} + \Delta a\hat{f}_a) \quad \text{(P)}$$

$$\text{s.t.} \quad \int v(w)\hat{f}_a = \left(\frac{\hat{a} + \Delta a}{\hat{a}}\right)^{1/\epsilon} \int v(w_A)\hat{f}_a \quad \text{(IC)}$$

$$\int v(w) (\hat{f} + \Delta a\hat{f}_a) \geq \int v(w_A) (\hat{f} + \Delta a\hat{f}_a) \quad \text{(IR)}$$

- Suppose $a(w) = \hat{a} + \Delta a$. Using a first-order approximation:
  $$f(\cdot|\hat{a} + \Delta a) \simeq \hat{f} + \Delta a\hat{f}_a \quad \text{and} \quad c(\hat{a} + \Delta a) \simeq c(\hat{a}) + \Delta a \int v(w_A)\hat{f}_a$$

- It follows from $\log a(w) = \beta + \epsilon \log M(w)$ that $w$ must satisfy (IC).

- Constraint that $w$ gives at least as much utility as $w_A$:
  $$\int v(w(x))f(x|\hat{a} + \Delta a) - c(\hat{a} + \Delta a) \geq \int v(w_A)\hat{f} - c(\hat{a}) \quad \Longrightarrow \text{(IR)}$$
Towards an Optimal non-local Perturbation

Claim: Principal should solve

\[
\max_{w(\cdot), \Delta a} m(\hat{a} + \Delta a) - \int w(\hat{f} + \Delta a \hat{f}_a) \tag{P}
\]

\[
\text{s.t. } \int v(w) \hat{f}_a = \left(\frac{\hat{a} + \Delta a}{\hat{a}}\right)^{1/\epsilon} \int v(w_A) \hat{f}_a \quad \tag{IC}
\]

\[
\int v(w) (\hat{f} + \Delta a \hat{f}_a) \geq \int v(w_A) (\hat{f} + \Delta a \hat{f}_a) \quad \tag{IR}
\]

- Suppose \( a(w) = \hat{a} + \Delta a \). Using a first-order approximation:

\[
f(\cdot | \hat{a} + \Delta a) \simeq \hat{f} + \Delta a \hat{f}_a \quad \text{and} \quad c(\hat{a} + \Delta a) \simeq c(\hat{a}) + \Delta a \int v(w_A) \hat{f}_a
\]

- It follows from \( \log a(w) = \beta + \epsilon \log M(w) \) that \( w \) must satisfy (IC).

- Constraint that \( w \) gives at least as much utility as \( w_A \):

\[
\int v(w(x)) f(x|\hat{a} + \Delta a) - c(\hat{a} + \Delta a) \geq \int v(w_A) \hat{f} - c(\hat{a}) \quad \Longrightarrow \text{(IR)}
\]
Solving for the Optimal non-local Perturbation

- **Stage 1**: For every $\Delta a$, solve

  \[ \hat{\Pi}(\Delta a) = \max_{w(\cdot)} m(\hat{a} + \Delta a) - \int w(\hat{f} + \Delta a\hat{f}_a) \]

  subject to

  \[ \int v(w)\hat{f}_a = \left(\frac{\hat{a} + \Delta a}{\hat{a}}\right)^{1/\epsilon} \int v(w_A)\hat{f}_a \]

  \[ \int v(w)(\hat{f} + \Delta a\hat{f}_a) \geq \int v(w_A)(\hat{f} + \Delta a\hat{f}_a) \]

  - Optimization program is convex as long as $\hat{f} + \Delta a\hat{f}_a > 0$ for all $x$.

- **Stage 2**: Solve

  \[ \hat{\Pi}^* = \max_{\Delta a} \hat{\Pi}(\Delta a) \]

  - **Info. requirements**: Must know $\hat{f}$, $\hat{f}_a$, and $v'(\cdot)$ (using $\int \hat{f}_a = 0$)

  - **Alternative**: Can approximate $v(w) \approx v(w_A) + (w - w_A)v'(w_A)$ to make constraints linear in $w$—then stage 1 program is convex $\forall \Delta a$. 
Solving for the Optimal non-local Perturbation

- **Stage 1:** For every $\Delta a$, solve

  \[
  \hat{\Pi}(\Delta a) = \max_{w(\cdot)} m(\hat{a} + \Delta a) - \int w(\hat{f} + \Delta a\hat{f}_a)
  \]

  s.t. \[
  \int v(w)\hat{f}_a = \left(\frac{\hat{a} + \Delta a}{\hat{a}}\right)^{1/\epsilon} \int v(w_A)\hat{f}_a \]

  \[
  \int v(w) (\hat{f} + \Delta a\hat{f}_a) \geq \int v(w_A) (\hat{f} + \Delta a\hat{f}_a)
  \]

  Optimization program is convex as long as $\hat{f} + \Delta a\hat{f}_a > 0$ for all $x$.

- **Stage 2:** Solve

  \[
  \hat{\Pi}^* = \max_{\Delta a} \hat{\Pi}(\Delta a)
  \]

  - **Info. requirements:** Must know $\hat{f}$, $\hat{f}_a$, and $v'(\cdot)$ (using $\int \hat{f}_a = 0$)
  - **Alternative:** Can approximate $v(w) \simeq v(w_A) + (w - w_A)v'(w_A)$ to make constraints linear in $w$—then stage 1 program is convex $\forall \Delta a$. 

Georgiadis and Powell

Optimal Incentives under Moral Hazard

Northwestern Kellogg
Solving for the Optimal non-local Perturbation

- **Stage 1:** For every $\Delta a$, solve

  \[
  \hat{\Pi}(\Delta a) = \max_{w(\cdot)} m(\hat{a} + \Delta a) - \int w(\hat{f} + \Delta a\hat{f}_a) \\
  \text{s.t.} \quad \int v(w)\hat{f}_a = \left(\frac{\hat{a} + \Delta a}{\hat{a}}\right)^{1/\epsilon} \int v(w_A)\hat{f}_a \\
  \int v(w)(\hat{f} + \Delta a\hat{f}_a) \geq \int v(w_A)(\hat{f} + \Delta a\hat{f}_a)
  \]

  Optimization program is convex as long as $\hat{f} + \Delta a\hat{f}_a > 0$ for all $x$.

- **Stage 2:** Solve

  \[
  \hat{\Pi}^* = \max_{\Delta a} \hat{\Pi}(\Delta a)
  \]

  **Info. requirements:** Must know $\hat{f}$, $\hat{f}_a$, and $v'(\cdot)$ (using $\int \hat{f}_a = 0$)

  **Alternative:** Can approximate $v(w) \approx v(w_A) + (w - w_A)v'(w_A)$ to make constraints linear in $w$—then stage 1 program is convex $\forall \Delta a$. 
**Stage 1:** For every $\Delta a$, solve

$$\widehat{\Pi}(\Delta a) = \max_{w(\cdot)} m(\widehat{a} + \Delta a) - \int w(\widehat{f} + \Delta a \widehat{f}_a)$$

subject to

$$\int v(w) \widehat{f}_a = \left(\frac{\widehat{a} + \Delta a}{\widehat{a}}\right)^{1/\epsilon} \int v(w_A) \widehat{f}_a$$

$$\int v(w) (\widehat{f} + \Delta a \widehat{f}_a) \geq \int v(w_A) (\widehat{f} + \Delta a \widehat{f}_a)$$

Optimization program is convex as long as $\widehat{f} + \Delta a \widehat{f}_a > 0$ for all $x$.

**Stage 2:** Solve

$$\widehat{\Pi}^* = \max_{\Delta a} \widehat{\Pi}(\Delta a)$$

**Info. requirements:** Must know $\widehat{f}$, $\widehat{f}_a$, and $v'(\cdot)$ (using $\int \widehat{f}_a = 0$)

**Alternative:** Can approximate $v(w) \simeq v(w_A) + (w - w_A) v'(w_A)$ to make constraints linear in $w$—then stage 1 program is convex $\forall \Delta a$. 

*Solving for the Optimal non-local Perturbation*
Extensions

1. **Bounded payments.** Assume that $w_A(x) + t(x) \in [\underline{w}, \overline{w}]$
   - New constraints are linear, so principal’s problem remains convex.

2. **Heterogeneous abilities.** Assume that the principal offers a common contract to multiple agents who have heterogeneous effort costs.
   - Principal must classify the agents into types ($\phi$), and estimate $\Pr\{\phi\}$, $\widehat{f}_\phi$, $\widehat{f}_a^\phi$, and $\mathcal{D}a^\phi(\widehat{w}, \widehat{t})$ for each $\phi$.
   - Can induce selection by imposing participation for subset of types.

3. **Multidimensional effort.** Assume agent’s effort $a \in \mathbb{R}^N$ at cost $c(a)$
   - e.g., effort towards quantity & quality, or selling different products.
   - Principal must have output data for $K \geq (N + 3)/2$ contracts.
Extensions

4. **Parametric contract classes.** Assume the principal restricts attention to contracts of the form $w_\alpha$, where $\alpha$ is a vector of parameters.
   - Find optimal perturbation direction $z$. *(New contract: $w_{\alpha+\theta z}$)*
   - Same informational requirements as general case.

5. **Other sources of incentives.** *(Promotion, firing threat, prestige, etc)*
   - Results hold verbatim if the agent’s IC constraint can be written as
     \[
     \int v(w)f_\alpha dx + I(a(w)) = c'(a(w)),
     \]
   - where $I(a)$ denotes marginal benefit of effort due to *indirect incentives*.
     - *Key:* Additive separability and $I(\cdot)$ not directly dependent on $w$.

6. **Multiplicatively separable utility.** Agent’s payoff $u(\omega, a) = v(\omega)c(a)$
   - *Example:* Agent’s utility satisfies CARA.
   - Principal must take a stance on $v$ (instead of $v'$).
Empirical Validation

Dataset

- **Goal**: Illustrate application & evaluate methodology
- Dataset from DellaVigna and Pope (2017)
- Real-effort experiment on M-Turk: Subjects press a-b keys for 10 min
- 7 treatments with different monetary incentives:

<table>
<thead>
<tr>
<th>Contract (in ¢)</th>
<th>Mean effort</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>( w_1(x) = 100 )</td>
<td>1521</td>
<td>540</td>
</tr>
<tr>
<td>( w_2(x) = 100 + 0.001x )</td>
<td>1883</td>
<td>538</td>
</tr>
<tr>
<td>( w_3(x) = 100 + 0.01x )</td>
<td>2029</td>
<td>558</td>
</tr>
<tr>
<td>( w_4(x) = 100 + 0.04x )</td>
<td>2132</td>
<td>566</td>
</tr>
<tr>
<td>( w_5(x) = 100 + 0.10x )</td>
<td>2175</td>
<td>538</td>
</tr>
<tr>
<td>( w_6(x) = 100 + 40I_{{x\geq2000}} )</td>
<td>2136</td>
<td>545</td>
</tr>
<tr>
<td>( w_7(x) = 100 + 80I_{{x\geq2000}} )</td>
<td>2188</td>
<td>532</td>
</tr>
</tbody>
</table>

- Each subject participates in a single treatment, once.
Two Exercises

I. Assume subjects are identical, and make assumptions about $v'$ and $m$

1. Given data for any two treatments, predict effort & profits for others.
   - Test predictions of two models:
     \[
     \log a(w) = \beta + \epsilon \log M(w) \\
     a(w) = \beta_0 + \beta_1 M(w)
     \]
     where $M(w) = \int v(w) \hat{f}_a$, and constants are estimated using A-B test.
   - *Sensitivity analysis*: Prediction accuracy vs. assumptions about $v'$

II. Counterfactuals:

1. Use all seven treatments to estimate the parameters of the model
2. Characterize optimally perturbed contract
3. Compare projected profits to those of $w_A$ and optimal contract
Step 1

1. Assume subjects have CRRA utility — specifically, $v'(\omega) = \omega^{-0.3}$
2. Normalize $a(w_i) = \text{(Mean effort)}_i$.
3. Given A-B test, estimate $f(\cdot|a(w_i))$ for $i \in \{A, B\}$, and compute

$$\hat{f}_a(x) = \frac{f(x|a(w_B)) - f(x|a(w_A))}{a(w_B) - a(w_A)}$$

![Graphs showing empirical CDF and estimated PDF for different conditions involving effort levels and normalized efforts.](image-url)
Exercise 1(a): Effort Predictions given Treatments 2 and 4

Predicted effort using the two models

- Actual effort
- $\hat{a}_{lin}$
- $\hat{a}_{log}$

Effort vs. Treatment graph
Exercise 1(b): Effort Prediction Accuracy

Empirical Validation

- Mean APE($\hat{a}_{\text{lin}}$) (avg = 6.06)
- Mean APE($\hat{a}_{\text{log}}$) (avg = 2.08)
- Max. APE($\hat{a}_{\text{lin}}$)
- Max. APE($\hat{a}_{\text{log}}$)

Absolute Percentage Error (APE)

Treatment Pair

Georgiadis and Powell

Optimal Incentives under Moral Hazard

Northwestern Kellogg
Exercise 1(c): Sensitivity Analysis

Effort Prediction Accuracy as a function of assumptions about $v'(\cdot)$

- Middle line: MAPE ($\hat{a}_{lin}$, CRRA utility)
- Blue line: MAPE ($\hat{a}_{log}$, CRRA utility)
- Green line: MAPE ($\hat{a}_{lin}$, Quadratic utility)
- Cyan line: MAPE ($\hat{a}_{log}$, Quadratic utility)

Coefficient of risk aversion: $\rho$ (CRRA utility) or $B$ (Quadratic utility)
Exercise 1(d): Profit Prediction Accuracy

![Profit Prediction Accuracy Graph]

- Mean APE($\hat{\pi}_{lin}$) (avg = 6.02)
- Mean APE($\hat{\pi}_{log}$) (avg = 2.06)
- Max. APE($\hat{\pi}_{lin}$)
- Max. APE($\hat{\pi}_{log}$)
**Estimate Model**

1. Use estimates of \( \{ f(\cdot|a(w_i)) \}_{i} \) to fit \( f(\cdot|a) \) for all \( a \) using linear interpolation (thus assuming \( f_a(x|a) \) is piece-wise linear in \( a \)).

2. Assume agent has CRRA utility and isoelastic costs; i.e.,

\[
v(\omega) = \frac{\omega^{1-\rho}}{1-\rho} \quad \text{and} \quad c(a) = \frac{c_0}{p+1} a^{p+1},
\]

and given \( w \), he chooses his effort \( a(w) \) such that

\[
\int v(w) f_a(\cdot|a(w)) \, dx + I = c^p(a(w)).
\]

Then, we estimate the unknown coefficients.

3. Assign value to principal’s marginal profit — specifically, \( m = 0.2 \).
Exercise 2(a): Optimal Perturbation

A-B test comprises treatments 4 and 7

- Optimal Perturbation (Profit = 252.82)
- Status Quo ($w_A$) (Profit = 241.98)
- Optimal Contract (Profit = 253.64)
Exercise 2(b): Profits relative to Optimal Contract

- Status quo contract (Avg = 94.72)
- Optimal perturbation (Avg = 97.01)
- Optimal perturbation (LP) (Avg = 97.25)
Summary & Future Work

- Framework for using agency theory to address an empirical question.
  - How to improve an existing performance pay plan?
  - What information do you need to do so?

- Other questions:
  - Optimal experimentation (ratchet effects, behavioral constraints)?
  - Extend to other settings (non-monetary instruments, dynamics)?