Selling Training Data

Jingmin Huang, Wei Zhao and **Renjie Zhong***

Renmin University of China

2024 CCER Summer Institute
June 30, 2024

This slide deck:

Selling Training Data

- data is trained to make **predictions** to improve the quality of **decision making under uncertainty**

Selling Training Data

- data is trained to make **predictions** to improve the quality of **decision making under uncertainty**
- data buyer, **endowed with private dataset**, seeks to augment the private data and refine the initial prediction by **purchasing more data**

Selling Training Data

- data is trained to make **predictions** to improve the quality of **decision making under uncertainty**
- data buyer, **endowed with private dataset**, seeks to augment the private data and refine the initial prediction by **purchasing more data**
- data seller versions the training data and designs the associated tariff to screen buyers with different private training data

Selling Training Data

- data is trained to make **predictions** to improve the quality of **decision making under uncertainty**
- data buyer, **endowed with private dataset**, seeks to augment the private data and refine the initial prediction by **purchasing more data**
- data seller versions the training data and designs the associated tariff to screen buyers with different private training data
- our question: what is the **optimal data selling mechanism**?

Selling Training Data

- data is trained to make **predictions** to improve the quality of **decision making under uncertainty**
- data buyer, **endowed with private dataset**, seeks to augment the private data and refine the initial prediction by **purchasing more data**
- data seller versions the training data and designs the associated tariff to screen buyers with different private training data
- our question: what is the **optimal data selling mechanism**?
- two screening toolkits: **prediction accuracy of data** and **price**

- two states : ω_1 (null hypothesis), ω_2 (alternative hypothesis), prior: $\mu_0 = (\frac{1}{2}, \frac{1}{2})$

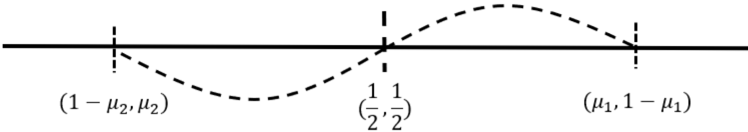
- two states : ω_1 (null hypothesis), ω_2 (alternative hypothesis), prior: $\mu_0 = (\frac{1}{2}, \frac{1}{2})$
- the predictions s'_1, s'_2 of private data, with data structure $\Pr(s'_i|\omega_i) = \pi'_i$, split the prior into two "interim" beliefs

$$(\tfrac{1}{2}, \tfrac{1}{2}) = \Pr(\mu_1)(\mu_1, 1-\mu_1) + \Pr(\mu_2)(1-\mu_2, \mu_2),\ \mu_1,\ \mu_2 \geq \tfrac{1}{2}$$

$(\mu_1, 1-\mu_1)$ identifies true state as ω_1 , $(1-\mu_2, \mu_2)$ identifies true state as ω_2

E'	s'_1	s'_2
ω_1	π'_1	$1-\pi'_1$
ω_2	$1-\pi'_2$	π'_2

Private Data $(1 \geq \pi'_1, \pi'_2 \geq 0.5)$



- two states : ω_1 (null hypothesis), ω_2 (alternative hypothesis), prior: $\mu_0 = (\frac{1}{2}, \frac{1}{2})$
- the predictions s_1', s_2' of private data, with data structure $\Pr(s_j'|\omega_i) = \pi_{ij}'$, split the prior to two posteriors

$$(\tfrac{1}{2}, \tfrac{1}{2}) = \Pr(\mu_1)(\mu_1, 1-\mu_1) + \Pr(\mu_2)(1-\mu_2, \mu_2), \ \mu_1, \ \mu_2 \geq \tfrac{1}{2}$$

$(1-\mu_2, \mu_2)$ identifies true state as $\omega_2 \Rightarrow$ Type I error $\alpha = \Pr(\mu_2)(1-\mu_2)$

$(\mu_1, 1-\mu_1)$ identifies true state as $\omega_1 \Rightarrow$ Type II error $\beta = \Pr(\mu_1)(1-\mu_1)$

- data $= (\alpha, \beta) =$ a bundle of statistical error

	E'	s_1'	s_2'	Statistical Errors
Null hypothesis $\rightarrow$	ω_1	π_1'	$\boxed{1-\pi_1'}$	$\rightarrow$ Type I Error $\alpha = \Pr(\mu_2)(1-\mu_2)$
Alternative hypothesis $\rightarrow$	ω_2	$\boxed{1-\pi_2'}$	π_2'	$\rightarrow$ Type II Error $\beta = \Pr(\mu_1)(1-\mu_1)$

Private Data $(1 \geq \pi_1', \pi_2' \geq 0.5)$

purchased data refines the initial predictions (s_1', s_2'), if combined with the private data purchased data refines the initial predictions (s_1', s_2'), if combined with the private data

- the predictions is in the form of (s_i, s_j),

  s_i is the refinement of prediction s_1', s_j is the refinement of prediction s_2'
- a specific data structure (π_1, π_2): reduces Type i error by π_i ratio

  π_i: probability inducing **Type i** error from identifying ω_{-i} in ω_i

Refinement of Original Predictions s_1' s_2'

E	(s_1, s_1)	$(\boxed{s_1}, \boxed{s_2})$	(s_2, s_1)	(s_2, s_2)
ω_1	π_{11}	π_{12}	π_{13}	π_{14}
ω_2	π_{21}	π_{22}	π_{23}	π_{24}

Supplement Training Data

E	(s_1, s_1)	(s_1, s_2)	(s_2, s_1)	(s_2, s_2)
ω_1	$1-\pi_1$	π_1	0	0
ω_2	0	π_2	0	$1-\pi_2$

A Subclass of Supplement Data

- purchased data $+$ private data $=$?

- purchased data + private data = ?
- the data buyer chooses to minimize the overall error by
  1. either combining the two datasets $\Rightarrow$ error $= \alpha\pi_1 + \beta\pi_2$ (**Typewise Reduction**)
  2. or only using the purchased one $\Rightarrow$ error $= \frac{1}{2}\pi_1$ or $\frac{1}{2}\pi_2$

	(a_1, a_1)	(a_1, a_2)	(a_2, a_2)
ω_1	$1 - \pi_1$	π_1	0
ω_2	0	π_2	$1 - \pi_2$

combine E with E′

E_θ	(a_1, a_1)	(a_2, a_2)
ω_1	1	0
ω_2	π_2	$1 - \pi_2$

E_θ	(a_1, a_1)	(a_2, a_2)
ω_1	$1 - \pi_1$	π_1
ω_2	0	1

only use E

- purchased data + private data = ?
- the data buyer chooses to minimize the overall error by
  1. either combining the two datasets $\Rightarrow$ error $= \alpha\pi_1 + \beta\pi_2$ (**Typewise Reduction**)
  2. or only using the purchased one $\Rightarrow$ error $= \frac{1}{2}\pi_1$ or $\frac{1}{2}\pi_2$

	(a_1, a_1)	(a_1, a_2)	(a_2, a_2)
ω_1	$1-\pi_1$	π_1	0
ω_2	0	π_2	$1-\pi_2$

combine E with E′

E_θ	(a_1, a_1)	(a_2, a_2)
ω_1	1	0
ω_2	π_2	$1-\pi_2$

E_θ	(a_1, a_1)	(a_2, a_2)
ω_1	$1-\pi_1$	π_1
ω_2	0	1

only use E

- combine data $\Longleftrightarrow \alpha\pi_1 + \beta\pi_2 \leq \min\{\frac{1}{2}\pi_1, \frac{1}{2}\pi_2\}$

  combination is **type-dependent**

  key observations: not combine low quality data + combination incongruence

Data Goods

- data = a **bundle of predictions** about different states

Data Goods

- data = a **bundle of predictions** about different states

  $\Rightarrow$ private data = a bundle of multiple Types of **statistical error**

Data Goods

- data = a **bundle of predictions** about different states

  $\Rightarrow$ private data = a bundle of multiple Types of **statistical error**

  $\Rightarrow$ private data generates **multi-dimensional preference** to reduce multiple Types of statistical error

Data Goods

- data = a **bundle of predictions** about different states

  ⇒ private data = a bundle of multiple Types of **statistical error**

  ⇒ private data generates **multi-dimensional preference** to reduce multiple Types of statistical error

- supplement data = **refinement of original predictions**

  ⇒ purchased data = **constrained re-allocation** of statistical error (damage goods)

- combination of datasets = **re-minimization of statistical error**

  ⇒ **correlations in predictions** determines the sub-additive or super-additive valuations for combined data

Data, Mechanism Design and Information Design

- to summarize, three attributes of data shapes the design of data selling
  1. data as goods is inherently **multi-dimensional** $\Rightarrow$ design multi-dimensional screening
  2. production of data is inherently **constrained** $\Rightarrow$ production possibility set of datasets
  3. the value data is intrinsically **combinatorial** $\Rightarrow$ design persuasion scheme in screening
- basic trade-off: **information rent extraction** versus **efficient value extraction**
- toolkits for designer:

$$\text{Type incongruence} = \underbrace{\text{preference incongruence}}_{\text{multi-dimensional mechanism}} + \underbrace{\text{combination incongruence}}_{\text{information design}}$$

different vectors of statistical error induce:

1. horizontal differences in preferences $\Rightarrow$ products differentiation
2. horizontal correlations in allocations $\Rightarrow$ avoid data distortion and destruction

Main Result

we extend the traditional revelation principle in information design (Kamenica and Gentzkow,2011; Bergemann and Morris,2016; Taneva,2019)

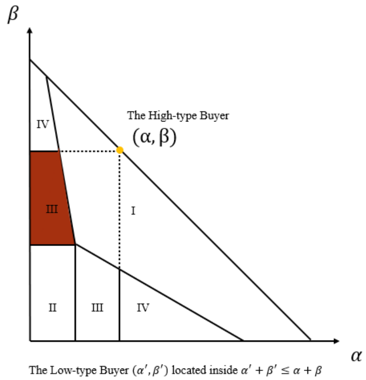
we derive some structural properties of the optimal data sale menu

1. we characterize of the **implementable set of predictions**
2. we show that the distortion of purchased data is constrained ("non-dispersed")
3. ,in binary state, we can restrict attention to purchased data with the specific form

   data $= (\pi_1, \pi_2)$ reducing original error (α, β) **Typewise**

   New Statistical Error $= \pi_1\alpha + \pi_2\beta$

we explicitly construct the four selling scheme in binary type



1. $\overline{E}$: eliminating all statistical error
2. ϕ: no reduction
3. E^*: **only** reduce some Type error by a **constant ratio**
4. $E_{(\alpha,\beta)}$: reduce both Types by **type-contingent** ratio

Zone	Data Menu	Selling Policy	Related Literature
I	$(\overline{E}, \overline{E})$	Full Disclosure	Riley & Zeckhauser(1983)
II	$(\overline{E}, \phi)$	Exclusive	Riley & Zeckhauser(1983)
III	$(\overline{E}, E^*)$	Partial Discrimination	**NEW!**
IV	$(\overline{E}, E_{(\alpha,\beta)})$	Full Discrimination	Bergemann et al.(2018)

- we then generalize the model to an economic significant situation

  Type II error $\beta = \mathit{constant}$, Type I error/significance level $\alpha \sim F_\alpha$with p.d.f f_α

- we then generalize the model to an economic significant situation

  Type II error $\beta = \mathit{constant}$, Type I error/significance level $\alpha \sim F_\alpha$ with p.d.f f_α
- optimal mechanism: **two-tiered** pricing
  1. in the first tier, a partially informative data, only reducing Type II error β by a fixed ratio
  2. in the second tier, data eliminating all statistical error
  3. the threshold is determined similar to the monopolist pricing
- similar to our new prediction in binary situation (zone III)

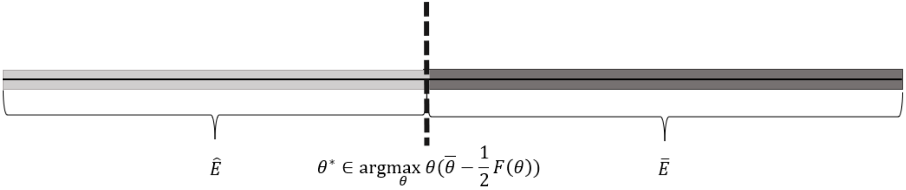


Figure: Two-tierd Pricing Mechanism

Literature Review

1. the design and price of information: Admati and Pfleiderer (1986); Admati and Pfleiderer (1990); Babaioff et al. (2012); Bergemann and Bonatti (2015); **Bergemann et al. (2018)**, Li (2022)
2. multi-dimensional screening: Adams and Yellen (1976); McAfee et al. (1989); Armstrong and Rochet (1999); Carroll (2017); Haghpanah and Hartline (2021); Yang (2021); Yang (2022)
3. applications of the infinite-dimensional extension of Carathéodory's theorem in economic design: Fuchs and Skrzypacz (2015); Bergemann et al. (2018); Kang (2023); Loertscher and Muir (2023); Dworczak and Muir (2024); Le Treust and Tomala (2019); Doval and Skreta (2024)

"The Design and Price of Information" (Bergemann et al.,18*AER*) v.s. Ours

"The Design and Price of Information" (Bergemann et al.,18*AER*) v.s. Ours

1. applications and interpretations: **input** data (cookie/history) v.s. **training** data

"The Design and Price of Information" (Bergemann et al.,18*AER*) v.s. Ours

1. applications and interpretations: **input** data (cookie/history) v.s. **training** data

   Bergemann et al.(2018): health care providers **with access to a patient's family** make predictions to enhances health care delivery

“The Design and Price of Information” (Bergemann et al.,18*AER*) v.s. Ours

1. applications and interpretations: **input** data (cookie/history) v.s. **training** data

   Bergemann et al.(2018): health care providers **with access to a patient’s family** make predictions to enhances health care delivery

   ours: the purchased medical data is used for **training supervised learning models** to predict disease diagnosis or drug efficacy

"The Design and Price of Information" (Bergemann et al.,18*AER*) v.s. Ours

1. applications and interpretations: **input** data (cookie/history) v.s. **training** data

   Bergemann et al.(2018): health care providers **with access to a patient's family** make predictions to enhances health care delivery

   ours: the purchased medical data is used for **training supervised learning models** to predict disease diagnosis or drug efficacy

2. private type: **one** interim belief v.s. a **distribution** of "interim" belief

"The Design and Price of Information" (Bergemann et al.,18*AER*) v.s. Ours

1. applications and interpretations: **input** data (cookie/history) v.s. **training** data

   Bergemann et al.(2018): health care providers **with access to a patient's family** make predictions to enhances health care delivery

   ours: the purchased medical data is used for **training supervised learning models** to predict disease diagnosis or drug efficacy

2. private type: **one** interim belief v.s. a **distribution** of "interim" belief

   ⇒ only **one** Type of error exists v.s. **both** Types exist

"The Design and Price of Information" (Bergemann et al.,18*AER*) v.s. Ours

1. applications and interpretations: **input** data (cookie/history) v.s. **training** data

   Bergemann et al.(2018): health care providers **with access to a patient's family** make predictions to enhances health care delivery

   ours: the purchased medical data is used for **training supervised learning models** to predict disease diagnosis or drug efficacy

2. private type: **one** interim belief v.s. a **distribution** of "interim" belief

   $\Rightarrow$ only **one** Type of error exists v.s. **both** Types exist

   $\Rightarrow$ **one**-preference + multi-allocation v.s. **multi**-preference + multi-allocation

“The Design and Price of Information” (Bergemann et al.,18*AER*) v.s. Ours

1. applications and interpretations: **input** data (cookie/history) v.s. **training** data

   Bergemann et al.(2018): health care providers **with access to a patient’s family** make predictions to enhances health care delivery

   ours: the purchased medical data is used for **training supervised learning models** to predict disease diagnosis or drug efficacy

2. private type: **one** interim belief v.s. a **distribution** of “interim” belief

   $\Rightarrow$ only **one** Type of error exists v.s. **both** Types exist

   $\Rightarrow$ **one**-preference + multi-allocation v.s. **multi**-preference + multi-allocation

3. economic insights: preference incongruence v.s. preference & combination incongruence

   Type incongruence = preference incongruence + combination incongruence

   preference incongruence: horizontal **differentiation** in **information** goods

   combination incongruence: horizontal **correlations** in **data** goods allocations

"The Design and Price of Information" (Bergemann et al.,18*AER*) v.s. Ours

1. applications and interpretations: **input** data (cookie/history) v.s. **training** data

   Bergemann et al.(2018): health care providers **with access to a patient's family** make predictions to enhances health care delivery

   ours: the purchased medical data is used for **training supervised learning models** to predict disease diagnosis or drug efficacy

2. private type: **one** interim belief v.s. a **distribution** of "interim" belief

   $\Rightarrow$ only **one** Type of error exists v.s. **both** Types exist

   $\Rightarrow$ **one**-preference + multi-allocation v.s. **multi**-preference + multi-allocation

3. economic insights: preference incongruence v.s. preference & combination incongruence

   Type incongruence = preference incongruence + combination incongruence

   preference incongruence: horizontal **differentiation** in **information** goods

   combination incongruence: horizontal **correlations** in **data** goods allocations

4. predictions: menu of **at most two** informative input data v.s. **two-tiered** menu

This slide deck:

Timeline

1. the seller posts a mechanism $\mathcal{M} = \{\mathcal{E}, t\}$
   1. a collection of experiments $\mathcal{E}$
   2. associated tariff $t : \mathcal{E} \to \mathbb{R}_+$
2. the buyer (with private experiment) chooses an experiment $E \in \mathcal{E}$ and pays price $t(E)$
3. the true state ω is realized
4. the buyer receive **two signal realizations** to update his belief, one from her **private experiment**, another from **the experiment E he purchased**, and she chooses an action a to maximize her expected utility
5. payoffs are realized

Individual Decision Problem

- finite states $\Omega = \{\omega_1, ..., \omega_I\}$ and common prior $\mu \in \Delta(\Omega)$, $\mu_i \equiv \mu(\omega_i)$
- the buyer chooses an action to maximize his expected payoff based on his information, from the finite action set $A = \{a_1, ..., a_J\}$
- utility function $u : A \times \Omega \to \mathbb{R}$, $u_{ij} \equiv u(a_j, \omega_i)$

$$\begin{array}{c|ccc} u & a_1 & \cdots & a_J \\ \hline \omega_1 & u_{11} & \cdots & u_{1J} \\ \vdots & \vdots & & \vdots \\ \omega_I & u_{I1} & \cdots & u_{IJ} \end{array}$$

- hereafter we assume matching utility if $u(a_j, \omega_i) = 1_{i=j}$ and $I = J$ to simplify harmful algebra

Buyer's Private Data

- type θ buyer, with private data E'_θ, decides to purchase data E_θ
- $E'_\theta = \{S, \pi'_\theta\}$ consists of signals $S' = \{s'_1, ..., s'_K\}$, with $\pi : \omega \to \Delta S$, $\pi'_{\theta ik} \equiv Pr[s'_k|\omega_i]$

E'_θ	s'_1	$\cdots$	s'_K
ω_1	$\pi'_{\theta 11}$	$\cdots$	$\pi'_{\theta 1K}$
$\vdots$	$\vdots$		$\vdots$
ω_I	$\pi'_{\theta I1}$	$\cdots$	$\pi'_{\theta IK}$

- each signal induces a posterior $\Rightarrow$ the agent's type is the **distribution of posteriors**, i.e. $\{\mu_{\theta 1}, ..., \mu_{\theta K}\}$ with $\Pr(\mu_{\theta k}) = \sum_{i=1}^{I} \mu_i \pi'_{\theta ik}$ for all $k \in 1, ..., K$, where $\mu_{\theta k} \in \Delta(\Omega)$

Outside Option without Purchasing Data

- without seller's data, conditioning on the prediction (signal) from the private data he accepted, the buyer chooses an optimal action
- the payoff is his expected value (in signals), constituting the **outside option** in this mechanism
- optimal action and payoff conditional on accepting s'_k for agent θ:

$$a\left(s'_k \mid E_\theta\right) \in \underset{a_j \in A}{\arg\max}\{\textstyle\sum_{i=1}^{I} \mu_{\theta k i} u_{ij}\} \text{ and } u(s'_k \mid E_\theta) \triangleq \max_j\{\textstyle\sum_{i=1}^{I} \mu_{\theta k i} u_{ij}\}$$

- expected payoff for type θ:

$$u_\theta \triangleq \textstyle\sum_{k=1}^{K} \Pr(\mu_{\theta k}) u(s'_k \mid E_n) = \sum_{k=1}^{K} \max_j \left\{\sum_{i=1}^{I} \mu_i \pi'_{\theta ik} u_{ij}\right\}$$

Value of Data in Individual Decision Problem

- suppose that the buyer combine the prediction from the purchased data with the one from his private data w.l.o.g (order invariance of Bayesian Updating)
- the optimal action and payoff conditional on accepting s_r and s'_k for agent n:

$$a\left(s_r \mid s'_k\right) \in \arg\max_{a_j \in A} \left\{ \sum_{i=1}^{I} \left(\frac{\mu_{\theta k i} \pi_{ir}}{\sum_{i'=1}^{I} \mu_{\theta k i'} \pi_{i' r}} \right) u_{ij} \right\}$$

$$u\left(s_r \mid s'_k\right) \triangleq \max_{j} \left\{ \sum_{i=1}^{I} \left(\frac{\mu_{\theta k i} \pi_{ir}}{\sum_{i'=1}^{I} \mu_{\theta k i'} \pi_{i' r}} \right) u_{ij} \right\}$$

- expected payoff for type θ:

$$u(\mathrm{E}, \theta) \triangleq \sum_{r=1}^{R} \sum_{k=1}^{K} \max_{j} \left\{ \sum_{i=1}^{I} \mu_i \pi'_{nik} \pi'_{ir} u_{ij} \right\}$$

- the value of data: $V(\mathrm{E}, \theta) \triangleq u(\mathrm{E}, \theta) - u_\theta$

Designer's Problem

- the seller posts a menu $\mathcal{M} = \{\mathcal{E}, t\}$ to maximize his profits
- we can restrict to the direct menu $\mathcal{M} = \{E_\theta, t_\theta\}_{\theta \in \Theta}$ by the revelation principle

Designer's Problem

$$\max_{\mathcal{M}} \int_\Theta t_\theta dF(\theta)$$

$$V(E_\theta, \theta) - t_\theta \geq 0, \ \forall \theta \in \Theta \qquad \text{(IR)}$$

$$V(E_\theta, \theta) - t_\theta \geq V(E_{\theta'}, \theta) - t_{\theta'}, \ \forall \theta, \theta' \in \Theta \qquad \text{(IC)}$$

This slide deck:

Motivation

we need some structural properties of the experiments in the optimal menu to

1. drop the **maximizer operator** $\Rightarrow$ "revelation principle" (recommendation)
2. reduce the **dimensions of screening** $\Rightarrow$ structural properties
3. tackle the **interactions between obedience in information design and mutual IC analysis**

"revelation principle" in data selling

1. signal set $S \Rightarrow$ action profile $\times_{k=1}^{K} A$ for all possible posteriors
2. signal realization $s \Rightarrow$ recommendation profile $a^r = (a_{r1}, ..., a_{rK})$ for all possible posteriors

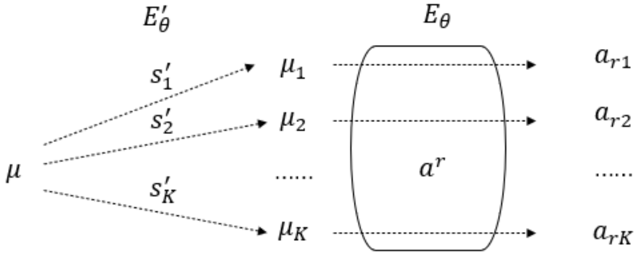


Figure: Direct Recommendation Mechanism

binary state $\{\omega_1, \omega_2\}$, binary action $\{a_1, a_2\}$

binary signal induces binary posterior $\{\mu_{\theta 1}, \mu_{\theta 2}\}$, $\frac{\mu_{\theta 1}(\omega_1)}{\mu_{\theta 1}(\omega_2)} > 1 > \frac{\mu_{\theta 2}(\omega_1)}{\mu_{\theta 2}(\omega_2)}$

initial prediction: ω_1 for $\mu_{\theta 1}$, ω_2 for $\mu_{\theta 2}$

only need to design recommendation schemes with a^1, a^2, a^4 as below:

E_θ	$a^1 = (a_1, a_1)$	$a^2 = (a_1, a_2)$	$a^3 = (a_2, a_1)$	$a^4 = (a_2, a_2)$
ω_1	$1 - \pi_1$	π_1	0	0
ω_2	0	π_2	0	$1 - \pi_2$

reversing the initial prediction $\times \Rightarrow a^3 = (a_2, a_1)$ cannot be implementabled

totally wrong predictions, (a_1, a_1) in ω_2, (a_2, a_2) in ω_1, are always undesirable

π_i: probability inducing (dis-utility from) **Type i** error from choosing a_{-i} in ω_i

π_i: probability inducing (dis-utility from) **Type i** error from choosing a_{-i} in ω_i

$a^1 = (a_1, a_1)$, $a^4 = (a_2, a_2)$ are always obeyed for any θ'

obedience of $a^2 = (a_1, a_2)$ for θ: whether to **combine the private dataset**

with binary type, $a^2 = (a_1, a_2)$ **should be obeyed for any θ'** in the **optimal** menu

	(a_1, a_1)	(a_1, a_2)	(a_2, a_2)
ω_1	$1-\pi_1$	π_1	0
ω_2	0	π_2	$1-\pi_2$

combine E with E′

E_θ	(a_1, a_1)	(a_2, a_2)
ω_1	1	0
ω_2	π_2	$1-\pi_2$

E_θ	(a_1, a_1)	(a_2, a_2)
ω_1	$1-\pi_1$	π_1
ω_2	0	1

only use E

This slide deck:

Binary Situation

- binary state $\{\omega_1, \omega_2\}$, action $\{a_1, a_2\}$, type $\{\theta, \theta'\}$, and posterior
- common prior $\mu = (\frac{1}{2}, \frac{1}{2})$, uniform type distribution $\Pr(\text{type } \theta) = \frac{1}{2}$
- to simplify the notation, denote the two type as:

$$
\begin{aligned}
\mu &= \left(\frac{1}{2}, \frac{1}{2}\right) \\
&= \frac{\mu_2 - \frac{1}{2}}{\mu_1 + \mu_2 - 1}(\mu_1, 1 - \mu_1) + \frac{\mu_1 - \frac{1}{2}}{\mu_1 + \mu_2 - 1}(1 - \mu_2, \mu_2) \\
&= \frac{\mu'_2 - \frac{1}{2}}{\mu'_1 + \mu'_2 - 1}(\mu'_1, 1 - \mu'_1) + \frac{\mu'_1 - \frac{1}{2}}{\mu'_1 + \mu'_2 - 1}(1 - \mu'_2, \mu'_2)
\end{aligned}
$$

- suppose $\mu_1, \mu'_1, \mu_2, \mu'_2 > \frac{1}{2}$ and $V(\bar{E}, \theta) \geqslant V(\bar{E}, \theta')$ w.l.o.g

  $\bar{E}$: fully informative experiment with $\pi(s_i|\omega_i) = 1$ for $i = 1, 2$

Economic Intepretation for (α, β)

- now the private type $\theta = (\alpha, \beta)$ and $\theta' = (\alpha', \beta')$

$$\alpha = \underbrace{\frac{(\mu_1 - \frac{1}{2})}{\mu_1 + \mu_2 - 1}}_{\text{market share}} \underbrace{(1 - \mu_2)}_{V(\bar{E},(\mu_2,1-\mu_2))} \quad \beta = \underbrace{\frac{\mu_2 - \frac{1}{2}}{\mu_1 + \mu_2 - 1}}_{\text{market share}} \underbrace{(1 - \mu_1)}_{V(\bar{E},(\mu_1,1-\mu_1))}$$

- α and β represent the (dis-utility) from Type I error and Type II error respectively, reflecting the **prediction accuracy** of private data
- the higher Type *i* error, the stronger **preference** for reducing that Type
- **vertical preference**: overall statistical error $\alpha + \beta$
- **horizontal preference**: different Types of statistical error α and β

Designer's Problem

	(a_1, a_1)	(a_1, a_2)	(a_2, a_2)
ω_1	$1-\pi_1$	π_1	0
ω_2	0	π_2	$1-\pi_2$

by the shared responsiveness, the obedience constraint is:

$$\underbrace{\max\left\{\alpha\pi_1+\beta\pi_2, \alpha'\pi_1+\beta'\pi_2\right\}}_{\textbf{error when obeying } (a_1,a_2)} \leqslant \underbrace{\min\left\{\frac{1}{2}\pi_1, \frac{1}{2}\pi_2\right\}}_{\textbf{error when choosing } (a_1,a_1) \textbf{ or } (a_2,a_2)}$$

$$k_1 \equiv \max\left\{\frac{\beta}{\frac{1}{2}-\alpha}\frac{\beta'}{\frac{1}{2}-\alpha'}\right\} \leqslant \frac{\pi_1}{\pi_2} \leqslant \min\left\{\frac{\frac{1}{2}-\beta}{\alpha}, \frac{\frac{1}{2}-\beta'}{\alpha'}\right\} \equiv k_2$$

valuation for this experiment

$$V(E,\theta) = \alpha+\beta-\alpha\pi_1-\beta\pi_2 \quad V(E,\theta') = \alpha'+\beta'-\alpha'\pi_1-\beta'\pi_2$$

Existence of Fully Informative Experiment

Lemma (No Distortion)

The fully informative experiment $\overline{E}$ always lies in the optimal menu

- if not, replace the experiment selling to the one charging the highest fee as and charge her a higher fee
- by the existence of the fully informative experiment, the designer only designs the one for another type (two parameters)

$E_{\theta'}$	(a_1, a_1)	(a_1, a_2)	(a_2, a_2)
ω_1	$1 - \pi_1$	π_1	0
ω_2	0	π_2	$1 - \pi_2$

- suppose allocate the fully informative one $\bar{E}$ to the high value type w.l.o.g

Designer's Problem

$$\max_{E,t_\theta,t_{\theta'}} \frac{1}{2}\left(t_\theta + t_{\theta'}\right)$$

s.t.

$$\begin{aligned}
&V(\overline{E},\theta) - t_\theta \geqslant 0 && \text{(IR-}\theta\text{)}\\
&V(E,\theta') - t_{\theta'} \geqslant 0 && \text{(IR-}\theta'\text{)}\\
&V(\overline{E},\theta) - t_\theta \geqslant V(E,\theta) - t_{\theta'} && \text{(IC-}\theta\text{)}\\
&V(E,\theta') - t_{\theta'} \geqslant V(\overline{E},\theta') - t_\theta && \text{(IC-}\theta'\text{)}\\
&k_1 \leqslant \frac{\pi_1}{\pi_2} \leqslant k_2 && \text{(Responsiveness)}
\end{aligned}$$

This slide deck:

Four Selling Schemes

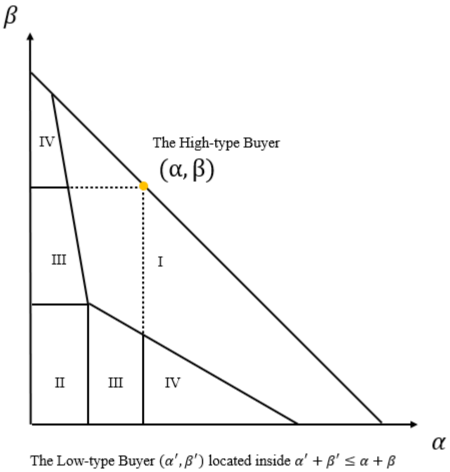


Figure: Optimal Selling Schemes

the high type always gets $\bar{E}$

the designer implements four selling schemes to low type

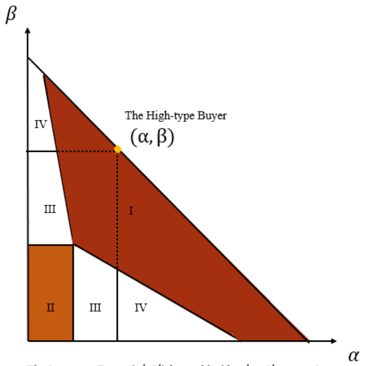
1. zone I: no discrimination
2. zone II: exclusive policy
3. zone III: partial discrimination
4. zone IV: perfect discrimination

Zone I and II: No-haggling

both Types of error are much smaller or not much smaller

$\Rightarrow$ approximately **one-dimensional** preference

Riley and Zeckhauser (1983)'s classic **no-haggling** result applies



The Low-type Buyer (α', β') located inside $\alpha' + \beta' \leq \alpha + \beta$

- zone I: no discrimination

  selling $\overline{E}$ to both types.

- zone II: exclusive policy

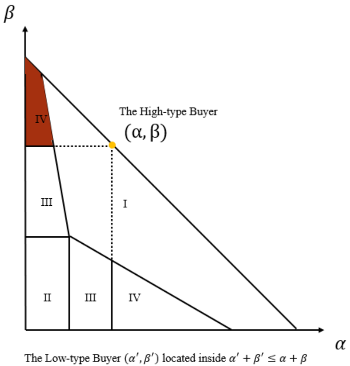
  only selling $\overline{E}$ to type-H

Zone IV: Type Incongruence

Type incongruence: some Type error of low type $>$ the one of high type (e.g. $\beta' > \beta$)

$\Rightarrow$ extract all the information rent v.s. huge loss in the extraction of low type valuation

horizontal differences in information products (Bergemann et al.,2018)



zone IV: perfect discrimination

selling $\bar{E}$ to type-H, and E to type-L, E smoothly changes in this zone

(IR-L),(IC-H),(IR-H),(Responsiveness) is binding

- zone IV (perfect discrimination)
- allocate (π_1^*, π_2^*) such that

  no information rent: $\alpha(1-\pi_1^*)+\beta(1-\pi_2^*)=\alpha'(1-\pi_1^*)+\beta'(1-\pi_2^*)$

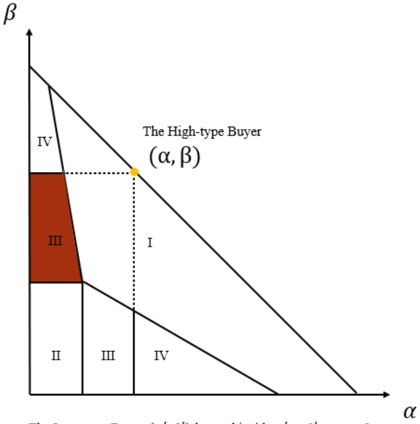
  exploitation of data structure: $\pi_1^*\alpha+\pi_2^*\beta=\frac{1}{2}\pi_i^*$

	(a_1, a_1)	(a_1, a_2)	(a_2, a_2)
ω_1	$1-\pi_1^*(\alpha', \beta')$	$\pi_1^*(\alpha', \beta')$	0
ω_2	0	$\pi_2^*(\alpha', \beta')$	$1-\pi_2^*(\alpha', \beta')$

- new statistical error $=\pi_1^*(\alpha', \beta')\cdot\alpha'+\pi_1^*(\alpha', \beta')\cdot\beta'$
- both Types of error of type-L is reduced by some **type-contingent ratio**

Zone III

when one much smaller while another not much smaller, trade-off emerges



The Low-type Buyer (α', β') located inside $\alpha' + \beta' \leq \alpha + \beta$

zone III: partial discrimination

selling $\bar{E}$ to type-H, and E^* to type-L

E^* is the same in this zone

(IR-L),(IC-H),(Responsiveness) is binding

- zone III (partial discrimination): $(\alpha, \beta) > (\alpha', \beta')$

  $\Rightarrow \pi_1\alpha + \pi_2\beta > \pi_1\alpha' + \pi_2\beta'$, given (π_1, π_2)

  $\Rightarrow$ information rent > 0 & higher tendency to not combine with private data

- design when $\frac{\beta}{2} < \beta' < \beta$ and $\alpha' < \frac{\alpha}{2}$

	(a_1, a_1)	(a_1, a_2)	(a_2, a_2)
ω_1	0	1	0
ω_2	0	π_2^*	$1 - \pi_2^*$

α'	a_1	a_2
ω_1	1	0
ω_2	1	0

β'	a_1	a_2
ω_1	1	0
ω_2	π_2^*	$1 - \pi_2^*$

- new statistical error $= 1 \cdot \alpha' + \pi_2^* \cdot \beta'$

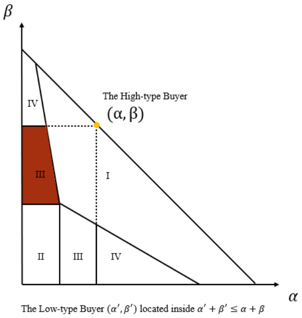
- Type II error is reduced by a **fixed** ratio while Type I error remains **unchanged**

- the determination of π_2^*: making the high type indifferent from combining or not

$$\alpha + \pi_2^*\beta = \frac{1}{2}\pi_2^*$$

- **pecking order**: Type II error $\rightarrow$ Type I error

Takeaway



1. $\overline{E}$: eliminating all statistical error
2. ϕ: no reduction
3. E^*: **only** reduce some Type error by a **constant ratio**
4. $E_{(\alpha,\beta)}$: reduce both Types by **type-contingent** ratio

Zone	Data Menu	Selling Policy	Related Literature
I	$(\overline{E}, \overline{E})$	Full Disclosure	Riley & Zeckhauser(1983)
II	$(\overline{E}, \phi)$	Exclusive	Riley & Zeckhauser(1983)
III	$(\overline{E}, E^*)$	Partial Discrimination	**NEW!**
IV	$(\overline{E}, E_{(\alpha,\beta)})$	Full Discrimination	Bergemann et al.(2018)

This slide deck:

Main Result

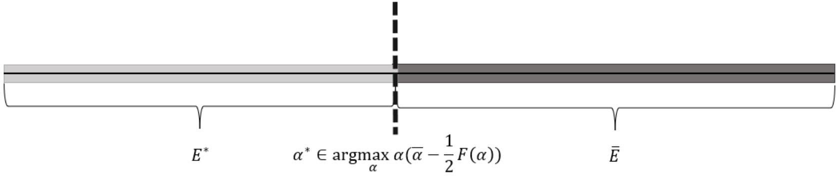
a generalized economic significant situation:

β=constant, $\alpha \sim F(\alpha)$ with p.d.f f, $\mathrm{Supp}(F) = [\underline{\alpha}, \overline{\alpha}] = [0, \frac{1}{2} - \beta]$

Theorem (The Optimality of the Cutoff Mechanism)

The optimal selling mechanism is

1. $(E_\alpha, t_\alpha) = (\bar{E}, \bar{t})$ *for all for* $\alpha \in [\alpha^*, \overline{\alpha}]$
2. $(E_\alpha, t_\alpha) = (E^*, t^*)$ *for* $\alpha \in [\underline{\alpha}, \alpha^*)$, *where* $\pi_1 = \frac{\alpha^*}{\overline{\alpha}}$, $\pi_2 = 1$
3. $\alpha^* \in \arg\max_\alpha \alpha \left(\overline{\alpha} - \frac{1}{2}F(\alpha)\right)$



- the optimal mechanism takes a simple and economically interpretable structure
  1. the types are partitioned into two tiers according to their predictive power
  2. the first tier: E^*, only reducing Type II error β by a constant ratio
  3. the second tier: $\overline{E}$ eliminating all error
  4. the threshold is determined similar to the monopolist pricing (prior-dependent)
- the threshold type α^* is indifferent between (i) the two menus $(\bar{E}, \bar{t})$ and (E^*, t^*) (ii) merging his private data or not when purchasing E^*

E^*	(a_1, a_1)	(a_1, a_2)	(a_2, a_2)
ω_1	0	1	0
ω_2	0	$\frac{\alpha^*}{\overline{\alpha}}$	$1 - \frac{\alpha^*}{\overline{\alpha}}$

$\overline{E}$	(a_1, a_1)	(a_1, a_2)	(a_2, a_2)
ω_1	1	0	0
ω_2	0	0	1

Combinations of Datasets

- $(\pi_1(\alpha), \pi_2(\alpha), t_\alpha)$: menu for type α

Combinations of Datasets

- $(\pi_1(\alpha), \pi_2(\alpha), t_\alpha)$: menu for type α
- β is fixed $\Rightarrow$ eliminating all Type II error π_2 for valuation extraction is desirable

Combinations of Datasets

- $(\pi_1(\alpha), \pi_2(\alpha), t_\alpha)$: menu for type α
- β is fixed $\Rightarrow$ eliminating all Type II error π_2 for valuation extraction is desirable
- (Lemma) statistical error induced by the combination of purchased data (π_1, π_2) with private data (α, β) in the optimal menu:

$$\min\{\pi_1\alpha + \pi_2\beta, \tfrac{1}{2}\pi_1, \tfrac{1}{2}\pi_2\} = \min\{\pi_1\alpha + \pi_2\beta, \tfrac{1}{2}\pi_2\}$$

Combinations of Datasets

- $(\pi_1(\alpha), \pi_2(\alpha), t_\alpha)$: menu for type α
- β is fixed $\Rightarrow$ eliminating all Type II error π_2 for valuation extraction is desirable
- (Lemma) statistical error induced by the combination of purchased data (π_1, π_2) with private data (α, β) in the optimal menu:

$$\min\{\pi_1\alpha + \pi_2\beta, \tfrac{1}{2}\pi_1, \tfrac{1}{2}\pi_2\} = \min\{\pi_1\alpha + \pi_2\beta, \tfrac{1}{2}\pi_2\}$$

- $(\alpha', \beta') \geq (\alpha, \beta) \iff \alpha' \geq \alpha$: monotone data quality

Combinations of Datasets

- $(\pi_1(\alpha), \pi_2(\alpha), t_\alpha)$: menu for type α
- β is fixed $\Rightarrow$ eliminating all Type II error π_2 for valuation extraction is desirable
- (Lemma) statistical error induced by the combination of purchased data (π_1, π_2) with private data (α, β) in the optimal menu:

$$\min\{\pi_1\alpha + \pi_2\beta, \tfrac{1}{2}\pi_1, \tfrac{1}{2}\pi_2\} = \min\{\pi_1\alpha + \pi_2\beta, \tfrac{1}{2}\pi_2\}$$

- $(\alpha', \beta') \geq (\alpha, \beta) \iff \alpha' \geq \alpha$: monotone data quality

  $\Rightarrow \exists \lambda : \text{Supp}(F) \to \text{Supp}(F)$, $\pi_1(\alpha)\lambda(\alpha) + \pi_2(\alpha)\beta = \frac{1}{2}\pi_2(\alpha)$

  $\Rightarrow [\lambda(\alpha), \overline{\alpha}]$ always do not combine their private dataset when purchasing E_α, while $[\underline{\alpha}, \lambda(\alpha)]$ combine

Combinations of Datasets

- $(\pi_1(\alpha), \pi_2(\alpha), t_\alpha)$: menu for type α
- β is fixed $\Rightarrow$ eliminating all Type II error π_2 for valuation extraction is desirable
- (Lemma) statistical error induced by the combination of purchased data (π_1, π_2) with private data (α, β) in the optimal menu:

$$\min\{\pi_1\alpha + \pi_2\beta, \tfrac{1}{2}\pi_1, \tfrac{1}{2}\pi_2\} = \min\{\pi_1\alpha + \pi_2\beta, \tfrac{1}{2}\pi_2\}$$

- $(\alpha', \beta') \geq (\alpha, \beta) \iff \alpha' \geq \alpha$: monotone data quality

$\Rightarrow \exists\lambda : \mathrm{Supp}(F) \to \mathrm{Supp}(F)$, $\pi_1(\alpha)\lambda(\alpha) + \pi_2(\alpha)\beta = \frac{1}{2}\pi_2(\alpha)$

$\Rightarrow [\lambda(\alpha), \overline{\alpha}]$ always do not combine their private dataset when purchasing E_α, while $[\underline{\alpha}, \lambda(\alpha)]$ combine

$\Rightarrow [\lambda(\alpha), \overline{\alpha}]$ makes **type-independent** error, thus sharing **the same incentives**

Basic Structure of Constraints

Basic Structure of Constraints

In the optimal mechanism, there always exists $\alpha^* \in \alpha$,

1. $E_\alpha = \overline{E}$ if and only if $\alpha \geq \alpha^*$
2. for all $\alpha < \alpha^*$,
    1. the responsiveness of α is not binding ($\lambda(\alpha) > \alpha$)
    2. there exists α', E_α is non-responsive for α', and $\mathrm{IC}[\alpha' \to \alpha]$ is binding

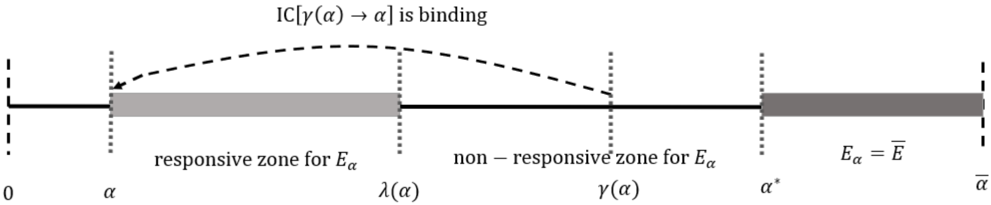


Figure: Constraint Structure

Basic Structure of Constraints

In the optimal mechanism, there always exists $\alpha^* \in \alpha$,

1. $E_\alpha = \overline{E}$ if and only if $\alpha \geq \alpha^*$
2. for all $\alpha < \alpha^*$,
   1. the responsiveness of α is not binding ($\lambda(\alpha) > \alpha$)
   2. there exists α', E_α is non-responsive for α', and $\mathrm{IC}[\alpha' \to \alpha]$ is binding

- optimality $\Rightarrow \exists\gamma : \mathrm{Supp}(F) \to \mathrm{Supp}(F)$, $\gamma \geq \lambda$, $\mathrm{IC}[\gamma(\alpha) \to \alpha]$ is binding

  define the single-valued correspondence $\gamma(\alpha)$ as below:

  1. $\gamma(\alpha) = \alpha$ if $\alpha = \lambda(\alpha)$
  2. $\gamma(\alpha) \in \{\alpha' \mid \mathrm{IC}[\alpha' \to \alpha] \text{ is binding}\}$ if $\alpha < \lambda(\alpha)$

- monotone data quality + endogenous minimization of error

  $\Rightarrow \lambda(\alpha) = \alpha \iff E_\alpha = \bar{E}$

Structure of Constraints

- structure of constraints $\Longleftrightarrow$ properties of $\lambda(\alpha)$ and $\gamma(\alpha)$

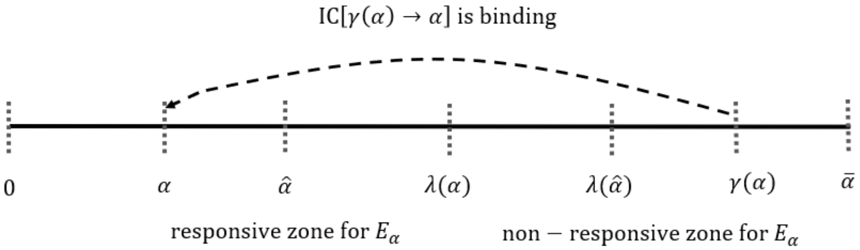
Structure of Constraints

- structure of constraints $\Longleftrightarrow$ properties of $\lambda(\alpha)$ and $\gamma(\alpha)$

Properties of $\lambda(\alpha)$ and $\gamma(\alpha)$

In the optimal menu,

1. $\lambda(\alpha) \leq \lambda(\hat{\alpha}) \leq \gamma(\alpha)$ for $\hat{\alpha} \in [\alpha, \lambda(\alpha)]$
2. $\lambda(\alpha)$: Supp$(F) \to \alpha$ is non-decreasing
   and characterizes the selling tiers, i.e. $(\pi_1(\alpha), \pi_2(\alpha), t_\alpha) = (\pi_1(\hat{\alpha}), \pi_2(\hat{\alpha}), t_{\hat{\alpha}})$ iff $\lambda(\alpha) = \lambda(\hat{\alpha})$
3. $\pi_1(\alpha)$: Supp$(F) \to [0, 1]$ is non-increasing



Price of Data in Screening Mechanism

combination rule: min {Statistical Error}

combine E with E′

only use E

	(s_1, s_1)	(s_1, s_2)	(s_2, s_2)
ω_1	$1 - \pi_1$	$\boldsymbol{\pi_1}$	0
ω_2	0	$\boldsymbol{\pi_2}$	$1 - \pi_2$

E_θ	(s_1, s_1)	(s_2, s_2)
ω_1	1	0
ω_2	$\boldsymbol{\pi_2}$	$1 - \pi_2$

E_θ	(s_1, s_1)	(s_2, s_2)
ω_1	$1 - \pi_1$	$\boldsymbol{\pi_1}$
ω_2	0	1

NEW ERROR!! $Type\ I + Type\ II = \alpha\pi_1 + \beta\pi_2$

$Type\ II = \frac{1}{2}\pi_2$

$Type\ I = \frac{1}{2}\pi_1$

Price of Data in Screening Mechanism

***combination rule*: *min* {*Statistical Error*}**

combine E with E′

	(s_1, s_1)	(s_1, s_2)	(s_2, s_2)
ω_1	$1 - \pi_1$	$\boldsymbol{\pi_1}$	0
ω_2	0	$\boldsymbol{\pi_2}$	$1 - \pi_2$

NEW ERROR!! $Type\ I + Type\ II = \alpha\pi_1 + \beta\pi_2$

only use E

E_θ	(s_1, s_1)	(s_2, s_2)
ω_1	1	0
ω_2	$\boldsymbol{\pi_2}$	$1 - \pi_2$

$Type\ II = \frac{1}{2}\pi_2$

E_θ	(s_1, s_1)	(s_2, s_2)
ω_1	$1 - \pi_1$	$\boldsymbol{\pi_1}$
ω_2	0	1

$Type\ I = \frac{1}{2}\pi_1$

- price of data for some type $=$ min statistical error $+$ monetary transfer
- price of data in screening menu $=$ max min statistical error $+$ monetary transfer

Price of Data in Screening Mechanism

combination rule: min {Statistical Error}

combine E with E' — *only use E*

	(s_1, s_1)	(s_1, s_2)	(s_2, s_2)
ω_1	$1-\pi_1$	π_1	0
ω_2	0	π_2	$1-\pi_2$

E_θ	(s_1, s_1)	(s_2, s_2)
ω_1	1	0
ω_2	π_2	$1-\pi_2$

E_θ	(s_1, s_1)	(s_2, s_2)
ω_1	$1-\pi_1$	π_1
ω_2	0	1

NEW ERROR!! $\text{Type I} + \text{Type II} = \alpha\pi_1 + \beta\pi_2$ — $\text{Type II} = \frac{1}{2}\pi_2$ — $\text{Type I} = \frac{1}{2}\pi_1$

- price of data for some type $=$ min statistical error $+$ monetary transfer
- price of data in screening menu $=$ max min statistical error $+$ monetary transfer
- responsiveness $\Longleftrightarrow$ price of data in screening menu $=$ constant

i.e. $\frac{1}{2}\pi_2(\alpha) + t_\alpha = C$

Denote $V(\alpha) = V(E_\alpha, \alpha) - t_\alpha$ as the net value of type α.

Equivalent Conditions of Constraints

In the optimal mechanism, IC,IR and Responsiveness conditions are equivalent to

1. $\frac{1}{2}\pi_2(\alpha) + t_\alpha = \bar{t}$ for all $\alpha \in \alpha$, $\bar{t}$ is the associated tariff for all $\alpha \in [\alpha^*, \overline{\alpha}]$
2. $V(\alpha) = \int_0^\alpha (1 - \pi_1(t))dt + V(\underline{\alpha})$
3. $\text{IR}[\underline{\alpha}]$ holds
4. $\pi_1(\alpha) : \text{Supp}(F) \to [0,1]$ is non-increasing

Designers' Problem of Choosing Optimal π_1

$$\max_{\pi_1(\alpha)} \int_{\underline{\alpha}}^{\overline{\alpha}} \Phi(\alpha) d\pi_1(\alpha) \quad \text{s.t.} \quad \pi_1(\alpha) : \text{Supp}(F) \to [0,1] \text{ is non-increasing}$$

where $\Phi(\alpha) : \alpha \to \mathbb{R}$.

designer's problem: maximizing a linear functional subject to monotonicity

Theorem (An Infinite-dimensional Extension of Carathéodory Theorem)

Let K be a convex, compact set in a locally convex Hausdorff space, and let $l : K \to \mathcal{R}^m$ be a continuous affine function such that $\sum \subseteq \operatorname{im} l$ is a closed and convex set. Suppose that $l^{-1}(\sum)$ is nonempty and and that $\Omega : K \to \mathcal{R}$ is a continuous convex function. Then there exists $z^* \in l^{-1}(\sum)$ such that $\Omega(z^*) = \max_{z \in l^{-1}(\sum)} \Omega(z)$ and

$$z^* = \sum_{i=1}^{m+1} \alpha_i z_i, \text{ where } \sum_{i=1}^{m+1} \alpha_i = 1, \text{ and for all } i,\ \alpha_i \geq 0,\ z_i \in \mathrm{ex} K$$

- Convexity, compact in the L_1 topology, and the existence of the optimalize, are satisfied in a mechanism design with transferable utility setting (Kang,2023)

Extreme Points of Monotone Allocation Functional Space

$\mathrm{ex}\Pi = \{\pi | \pi : \Theta \to [0,1],\ \pi \text{ is non-increasing}\} = \{\pi |\ \pi \in \Pi \text{ and } \operatorname{im} \pi \subseteq \{0,1\}\}$

- with the form of π_1 and the tiered-pricing structure, we can further deduce the optimality of two-tiered pricing of tiers
- we can transform the designer's problem as the choice of the optimal threshold α^*.

Optimal Threshold

$\alpha^* \in \arg\max_\alpha \alpha \left(\overline{\alpha} - \frac{1}{2}F(\alpha)\right)$ is the optimal threshold of the tiers.

Conclusion

- we propose a new framework to analyze selling training data
- we extend the traditional revelation principle and derive the specification of information structure in screening
- new prediction beyond conventional wisdom: the designer utilizes the combination incongruence to screen agents
- novel approach: we use functions to characterize the tightness and structure of constraints
- tight prediction in data selling mechanism: two-tiered pricing mechanism