# Selling Training Data (Preliminary)

Jingmin Huang, Wei Zhao and Renjie Zhong*

Renmin University of China

Berkeley Theory Lunch

# Selling Input Data/Within Consumer Data

Bergemann, Bonatti, Smolin (2018 AER) "The Design and Price of Information"

- monopolist screening: data broker, buyer with private information (interim belief)

  ex. lenders with knowledge of a borrower; doctors with access to patients' family history...

- private information + input data $\rightarrow$ optimize their decision under uncertainty

  input data: update prediction algorithms $\Rightarrow$ cost and/or quality of offerings (Joshua,2024)

- key attributes: position & quality

- private signal $s \iff$ interim belief $\mu_s$

  $\Longrightarrow$ certain Type of statistical error induced by action selection $a_s$ and $u(a_s, \omega)$ ✓

# Selling Training Data/Across Consumer Data

- monopolist screening: data broker, data buyer with private ~~information~~ baseline dataset

- ~~private information + input data → optimize their decision under uncertainty~~

  baseline data + supplemental data → train its predictive model.

  training data: develop AI prediction algorithms ⇒ market entry (Joshua,2024)

- key attributes: multi-dimension & combinatorial nature & allocation rigidity

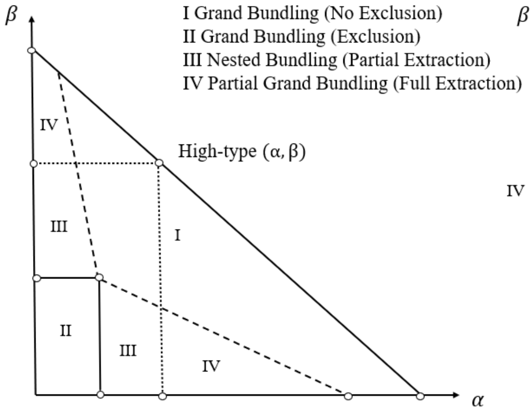
- private experiment $\Pr(s|\omega) \Longleftrightarrow$ distribution of posteriors $F(\mu)$ ✗

  $\Longrightarrow$ a bundle of statistical error induced by action scheme $\Pr(a, \omega)$ and $u(a, \omega)$ ✓

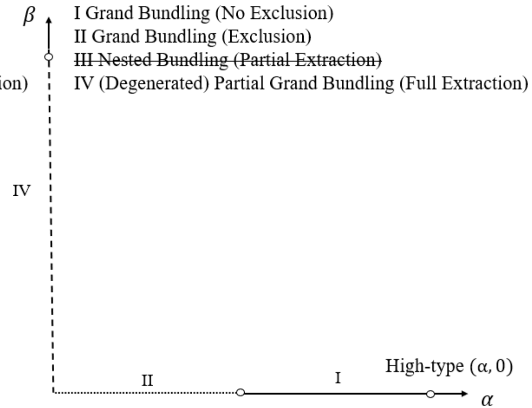
# Training Data v.s. Input Data

Training data reduces error of baseline data $(\alpha, \beta) \Rightarrow$ constrained multi-dimensional goods

Input data reduces certain error of private information $(\alpha, 0)$ or $(0, \beta) \Rightarrow$ separate multi-goods




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




The Low-type Buyer $(\alpha', 0)$ or $(0, \beta')$ located inside $\alpha' \leq \alpha$ or $\beta' \leq \alpha$

## A simplified model for this talk (Hypothesis Testing)

- two states $\{\omega_1, \omega_2\}$, prior: $\mu = (\frac{1}{2}, \frac{1}{2})$, binary action $\{a_1, a_2\}$, payoff $u(a_i, \omega_j) = 1_{i=j}$
- private type: $(\alpha, \beta)$, $\alpha + \beta \leq \frac{1}{2}$

Private Experiment

| | $E'$ | $s_1'$ | $s_2'$ | Statistical Errors |
|---|---|---|---|---|
| Null hypothesis → | $\omega_1$ | $\Pr(s_1'\|\omega_1)$ | $\boxed{\Pr(s_2'\|\omega_1)}$ | → Type I Error $\alpha = \Pr(s_2'\|\omega_1)\,\mu(\omega_1)$ |
| Alternative hypothesis → | $\omega_2$ | $\boxed{\Pr(s_1'\|\omega_2)}$ | $\Pr(s_2'\|\omega_2)$ | → Type II Error $\beta = \Pr(s_1'\|\omega_2)\,\mu(\omega_2)$ |

$s_1'$: accept $\omega_1$ $\quad$ $s_2'$: reject $\omega_1$ $\quad$ $\Pr(s_1'|\omega_1) + \Pr(s_2'|\omega_2) \geq 1$

Figure: Baseline Dataset $\Longrightarrow$ Statistical Error

# Supplemental Data

Data broker recommends action profiles for different private signal realizations.

| $E$ | $(a_1, a_1)$ | $(a_1, a_2)$ | $(a_2, a_1)$ | $(a_2, a_2)$ |
|---|---|---|---|---|
| $\omega_1$ | $\pi_{11}$ | $\pi_{12}$ | $\pi_{13}$ | $\pi_{14}$ |
| $\omega_2$ | $\pi_{21}$ | $\pi_{22}$ | $\pi_{23}$ | $\pi_{24}$ |

Table: Straight Experiment

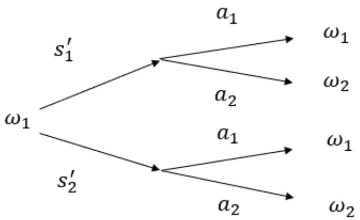



Figure: Data Merging

In the reduced-form, data broker allocates the reduction ratio of Type I and II error

Obedience: $\pi_1\alpha + \pi_2\beta \leq \min\{\frac{1}{2}\pi_1, \frac{1}{2}\pi_2\}$

| $E$ | $(a_1, a_1)$ | $(a_1, a_2)$ | $(a_2, a_1)$ | $(a_2, a_2)$ |
|---|---|---|---|---|
| $\omega_1$ | $1-\pi_1$ | $\pi_1$ | 0 | 0 |
| $\omega_2$ | 0 | $\pi_2$ | 0 | $1-\pi_2$ |

Table: Statistical Error Allocation

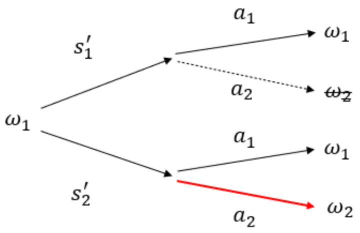



Figure: Reducing Type I error

Value of experiment $(\pi_1, \pi_2)$ for $(\alpha, \beta)$: incremental probability of correct identification

$$V(E,\theta) = \underbrace{\alpha + \beta}_{\text{initial overall error}} - \min \underbrace{\left\{\alpha\pi_1 + \beta\pi_2, \frac{1}{2}\pi_1, \frac{1}{2}\pi_2\right\}}_{\text{new overall error}}$$

type $\theta = (\alpha, \beta) \in \Theta$, mechanism $\mathcal{M} = \{\pi_1(\theta), \pi_2(\theta), t_\theta\}_{\theta \in \Theta}$

Designer's Problem:

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

$$\alpha\pi_1(\theta) + \beta\pi_2(\theta) \leq \min\{\tfrac{1}{2}\pi_1(\theta), \tfrac{1}{2}\pi_2(\theta)\}, \ \forall \theta \in \Theta$$

$$\alpha + \beta - \alpha\pi_1(\theta) - \beta\pi_2(\theta) - t_\theta \geq 0, \ \forall \theta \in \Theta$$

$$\alpha + \beta - \alpha\pi_1(\theta) - \beta\pi_2(\theta) - t_\theta \geq \alpha + \beta - \underbrace{\min\{\alpha\pi_1(\theta') + \beta\pi_2(\theta'), \frac{1}{2}\pi_1(\theta'), \frac{1}{2}\pi_2(\theta')\}}_{\text{two-step deviation}} - t_{\theta'}, \forall \theta, \theta' \in \Theta$$

Rigidity:

1. allocation rigidity: $\alpha\pi_1 + \beta\pi_2 \leq \min\{\frac{1}{2}\pi_1, \frac{1}{2}\pi_2\}$
2. differentiation rigidity: $\min\{\alpha\pi_1 + \beta\pi_2, \frac{1}{2}\pi_1, \frac{1}{2}\pi_2\}$

Flexibility : exploit the horizontal difference to neutralize the vertical difference

# Assumption: Perfectly Correlated

For the private type $(\alpha, \beta)$, it holds that $k\alpha + \beta = m$, with $m \in [0, \frac{1}{2})$, $k \in [0, 2m]$, and $\alpha \in [\underline{\alpha}, \overline{\alpha}] = [0, \frac{\frac{1}{2}-m}{1-k}]$ draws from absolutely continuous distribution $F$

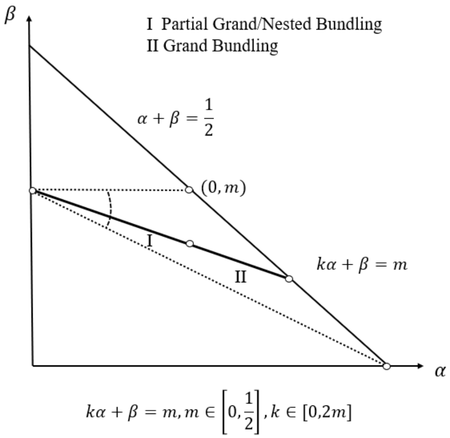



Figure: Optimal Menu

Two-tiered pricing is optimal:

1. $(E_\alpha, t_\alpha) = (\overline{E}, \overline{t})$ for $\alpha \in [\alpha^*, \overline{\alpha}]$
2. $(E_\alpha, t_\alpha) = (E^*, t^*)$ for $\alpha \in [\underline{\alpha}, \alpha^*)$

| $E^*$ | $(a_1, a_1)$ | $(a_1, a_2)$ | $(a_2, a_2)$ |
|---|---|---|---|
| $\omega_1$ | $k(1 - \frac{\alpha^*}{\overline{\alpha}})$ | $1 - k(1 - \frac{\alpha^*}{\overline{\alpha}})$ | $0$ |
| $\omega_2$ | $0$ | $\frac{\alpha^*}{\overline{\alpha}}$ | $1 - \frac{\alpha^*}{\overline{\alpha}}$ |

3. $\alpha^* \in \arg\max_\alpha \alpha \left((1-k)\overline{\alpha} - \frac{1}{2}F(\alpha)\right)$

Define $\lambda(\cdot) : [\underline{\alpha}, \overline{\alpha}] \to [\underline{\alpha}, \overline{\alpha}]$ such that $\mathrm{IC}[x \to \alpha]$ is one-step deviation for $x \in [\alpha, \lambda(\alpha)]$

Define $\gamma(\cdot) : [\underline{\alpha}, \overline{\alpha}] \to [\underline{\alpha}, \overline{\alpha}]$ such that $\mathrm{IC}[\gamma(\alpha) \to \alpha]$ is binding

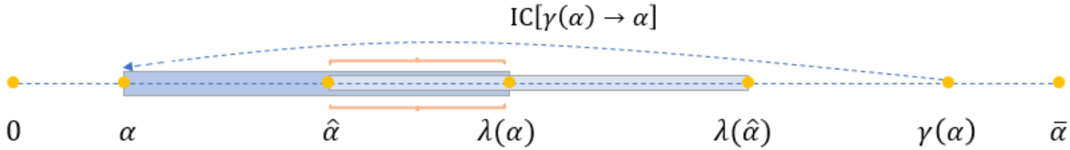



Figure: Optimal Structure of Menu

Key Conclusions:

1. **"FOC" property is transitive**: $\lambda(\cdot)$ is increasing
2. **price gap = informativeness gap**: $t_\alpha - t_{\alpha'} = \frac{1}{2}\pi_2(\alpha) - \frac{1}{2}\pi_2(\alpha')$ for all $\alpha, \alpha' \in [\underline{\alpha}, \bar{\alpha}]$

exploit horizontal differences to neutralize vertical difference

$\Rightarrow$ nullify the impact of private dataset and include of the low type

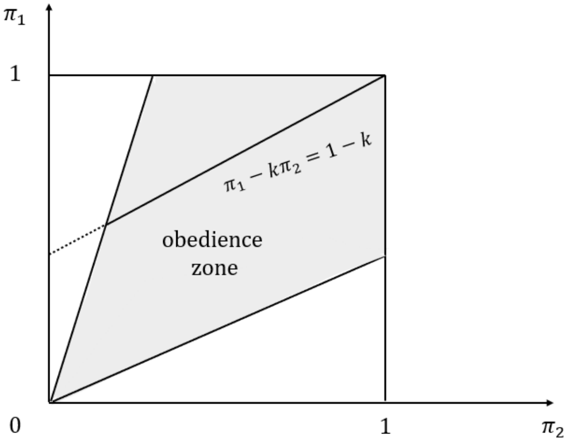



Figure: Neutralization Line and Allocation Rigidity

# Future Work

- a solid statistical decision foundation for two states
- extend the distributional assumption to generalize the result in two states
  1. robust optimal mechanism
  2. distributional assumptions sufficient for a simple optimal menu (?)
- trade-off in many signals, many actions and many states (?)
- concrete applications

# Literature Review

1. Information Design as Screening Tools: Admati and Pfleiderer (1986), Admati and Pfleiderer (1990), Babaioff et al. (2012), Bergemann et al. (2018), Yang (2022), Segura-Rodriguez (2022), Bonatti et al. (2023), Bonatti et al. (2024), Rodriguez Olivera (2024)
2. Multi-dimensional Screening: Adams and Yellen (1976), McAfee et al. (1989), Armstrong and Rochet (1999), Manelli and Vincent (2007), Hart and Reny (2015), Daskalakis et al. (2017), Carroll (2017), Haghpanah and Hartline (2021); Yang (2022), Deb and Roesler (2023)

# Thank You!