Optimal Statistical Estimation with Strategic Data Sources

Yang Cai, Constantinos Daskalakis, Christos Papadimitriou
presented by
Dean Grosbard
The setup:

• We wish to get an estimate $\hat{f}(\bar{x}, \bar{y})(x^*)$ of some function $f(x^*)$ for some $x^* \sim F$.

• To estimate $f$ we pay worker $w_i \in W' \subseteq W$, $p_i(\bar{x}, \bar{y})$ dollars to estimate $f(x_i)$

• The worker decides to invest $e_i$ dollars worth of effort to return an estimate

$$y_i = f(x_i) + \epsilon_i(e_i), \text{ where } E(\epsilon_i(e_i)) = 0, V(\epsilon_i(e_i)) = \sigma_i^2(e_i)$$

where $\sigma_i(\cdot)$ is a known convex strictly decreasing positive function
The Objective Function:

• We want to minimize

\[ \mathbb{E}_{x^*, y(\bar{e}^*)} \left[ \left( \frac{\hat{f}(\bar{x}, y(\bar{e}^*))}{x^*} - f(x^*) \right)^2 + \eta \cdot \sum_{i \in W'} p_i \left( (x_j, y_j(e_j^*))_{j \in W'} \right) \right] \]

by choice of \( x_i, p_i \) subject to individual rationality:

\[ \mathbb{E} \left[ p_i \left( (x_j, y_j(e_j^*))_{j \in W'} \right) \right] - e_i^* \geq 0. \text{ for all } i \in W' \]

• A worker will invest \( e^* \) dollars worth of effort where:

\[ e^* \in \arg \max_{e \in \mathcal{E}} \mathbb{E}[p(x, y(e))] - e \]
The Fair Dictator Optimization Problem

• If we could force worker $i \in W$ to invest to $e_i$ dollars (dictator) and pay him $e_i$ for his effort (fair) the problem becomes:

$$\min_{\mathcal{W}', (x_i, e_i)_{i \in \mathcal{W}'}} \left( \mathbb{E}_{x^*, \tilde{y}} \left[ \left( \hat{f}(\tilde{x}, \tilde{y})(x^*) - f(x^*) \right)^2 \right] + \eta \cdot \sum_i e_i \right)$$

where $\tilde{y} = \tilde{y}(\tilde{e})$.

• Assume we can solve this and the optimal solution is $\tilde{e}$, than this optimal solution is a lower bound for:

$$\mathbb{E}_{x^*, \tilde{y}(\tilde{e}^*)} \left( \left( \hat{f}(\tilde{x}, \tilde{y}(\tilde{e})))(x^*) - f(x^*) \right)^2 + \eta \cdot \sum_{i \in \mathcal{W}'} p_i \left( (x_j, y_j(e^*_j))_{j \in \mathcal{W}'} \right) \right)$$
The Fair Dictator Optimization Problem

- If we could force worker $i \in W$ to invest to $e_i$ dollars (dictator) and pay him $e_i$ for his effort (fair) the problem becomes:

$$
\min_{W', (x_i, e_i)_{i \in W'}} \left( \mathbb{E}_{x^*, \tilde{y}} \left[ \left( \hat{f}(\tilde{x}, \tilde{y}) (x^*) - f(x^*) \right)^2 \right] + \eta \cdot \sum_i e_i \right)
$$

where $\tilde{y} = \tilde{y}(\tilde{e})$.

- Assume we can solve this and the optimal solution is $\tilde{e}$, than the fair dictator solution is a lower bound for:

$$
\mathbb{E}_{x^*, \tilde{y}(\tilde{e}^*)} \left[ \left( \hat{f}(\tilde{x}, \tilde{y}(\tilde{e}^*)) (x^*) - f(x^*) \right)^2 + \eta \cdot \sum_{i \in W'} p_i \left( (x_j, y_j(e_j^*))_{j \in W'} \right) \right]
$$
What Is An Optimal Design

• The **fair dictator solution** is a lower because of **individual rationality**:

\[
\mathbb{E} \left[ p_i \left( (x_j, y_j(e^*_j))_{j \in W'} \right) \right] - e_i^* \geq 0. \quad \text{for all } i \in W'
\]

implies

\[
\mathbb{E} \left[ \sum_{i \in W'} p_i \left( (x_j, y_j(e^*_j))_{j \in W'} \right) \right] \geq \sum_i e_i^*
\]

• Let \( e_i^* \), \( \forall i \in W' \) be the **fair dictator solution**

\[
\forall i \in W', e_i(p_i(\hat{x}, \hat{y})) = e_i^* \land p_i(\hat{x}, \hat{y}) = e_i^* \Rightarrow p_i(\hat{x}, \hat{y}), i \in W' \text{ is optimal}
\]
**Desired Property of** \( P_i(\bar{x}, \bar{y}) \)

- We wish to design \( P_i(\bar{x}, \bar{y}) \) in such a way that investing an effort of \( e_i^* \) is a dominant strategy for worker \( i \in W' \).

\[
\mathbb{E} \left[ p_i \left( (x_i, y_i(e_i^*)), (x_j, y_j(e_j))_{j \in W \setminus \{i\}} \right) \right] - e_i^* \geq \mathbb{E} \left[ p_i \left( (x_j, y_j(e_j))_{j \in W} \right) \right] - e_i
\]

This would guarantee that \( e^* \) which we know to be optimal is a unique strong Nash equilibrium.
Design of $P_i(\bar{x}, \bar{y})$

• We define the structure of $P_i(\bar{x}, \bar{y})$ as:

$$p_i((\bar{x}, \bar{y})) = c_i - d_i \cdot \left(y_i - \hat{f}(\bar{x}, \bar{y})_{-i}(x_i)\right)^2$$

• Where $\hat{f}_{(\bar{x}, \bar{y})_{-i}}(x_i)$ is the prediction for $f(x_i)$ that we would get from estimating $f$ using only $\{(x_j, y_j)\}_{j \neq i}^{x_i} \in W'$

• Can we find $\{(c_i, d_i)\}_{i \in W}$ such that:

$$\forall i \in W', e_i(p_i(\bar{x}, \bar{y})) = e_i^* \land p_i(\bar{x}, \bar{y}) = e_i^*$$
Finding the Optimal $P_i(\bar{x}, \bar{y})$

$$p_i((\bar{x}, \bar{y})) = c_i - d_i \cdot \left(y_i - \hat{f}(\bar{x}, \bar{y})_{-i}(x_i)\right)^2 \Rightarrow$$

$$E_{\bar{y}(\bar{e})} [p_i ((\bar{x}, \bar{y}(\bar{e}')))] = c_i - d_i E_{\bar{y}(\bar{e})} \left[\left(y_i - \hat{f}(\bar{x}, \bar{y})_{-i}(x_i)\right)^2\right] =$$

$$c_i - d_i E_{\bar{y}(\bar{e})} \left[\left((y_i - f(x_i)) - (\hat{f}(\bar{x}, \bar{y})_{-i}(x_i) - f(x_i))\right)^2\right] =$$

$$= c_i - d_i \cdot E_{\bar{y}(\bar{e})} \left[(y_i - f(x_i))^2 + (f(x_i) - \hat{f}(\bar{x}, \bar{y})_{-i}(x_i))^2 - 2(y_i - f(x_i))(f(x_i) - \hat{f}(\bar{x}, \bar{y})_{-i}(x_i))\right]$$

$$= c_i - d_i \cdot \left(\sigma_i(e'_i)^2 + g(\bar{x}_{-i}, 1_{x_i}, \bar{\sigma}_{-i}(e'_{-i}))\right),$$

Known, **Bounded** and independent of $e'_i$
Finding the Optimal $P_i(\bar{x}, \bar{y})$

- $E_{\tilde{y}(e_{\epsilon})} \left[ p_i \left( (\tilde{x}, \tilde{y}(\tilde{e}')) \right) \right] = c_i - d_i \cdot (\sigma_i(e_i')^2 + g(\bar{x}_i, \mathbf{1}_{x_i}, \bar{\sigma}_i(\bar{e}_{-i})))$

- And so each worker $i \in W'$ will solve the following convex problem

\[
\max_{e_i'} \left( c_i - d_i \cdot (\sigma_i(e_i')^2 + g(\bar{x}_i, \mathbf{1}_{x_i}, \bar{\sigma}_i(\bar{e}_{-i}))) - e_i' \right)
\]

- By taking the derivative and equating to 0 we get the optimality condition:

\[2d_i \sigma_i(e_i^*) \cdot \sigma'_i(e_i^*) + 1 = 0.\]

where now $e_i' = e_i^*$ is the solution of the above problem
Finding the Optimal $P_i(\bar{x}, \bar{y})$

• If **now $e_i$ is the fair dictator solution** then by setting:

$$d_i = \frac{-1}{2\sigma_i(e_i)\sigma'_i(e_i)}.$$

We guarantee that $e_i^* = e_i$ (optimality condition $2d_i\sigma_i(e_i^*)\cdot\sigma'_i(e_i^*) + 1 = 0$)

• To guarantee **individual rationality** we set

$$c_i = d_i \cdot \left( \sigma_i(e_i)^2 + g(\bar{x} - i, 1_{x_i}, \bar{\sigma}_i(\bar{e} - i)) \right) + e_i.$$

• And we can verify that

$$p_i((\bar{x}, \bar{y})) = c_i - d_i \cdot \left( y_i - \hat{f}_{(\bar{x}, \bar{y}) - i}(x_i) \right)^2 = e_i$$
Finding the Optimal $P_i(\bar{x}, \bar{y})$

- If **now $e_i$ is the fair dictator solution** then by setting:

$$d_i = \frac{-1}{2\sigma_i(e_i)e_i}$$

We guarantee that $e_i^* = e_i$ (optimality condition $2d_i\sigma_i(e_i^*)\cdot\sigma'_i(e_i^*) + 1 = 0$)

- To guarantee **individual rationality** we set

$$c_i = d_i \cdot (e_i - e_i)^2 + g(\bar{x}_{-i}, 1x_i, \bar{\sigma}_{-i}(\bar{e}_{-i})) + e_i.$$ 

- And we can verify that

$$p_i((\bar{x}, \bar{y})) = c_i - d_i \cdot (y_i - \hat{f}(\bar{x}, \bar{y})_{-i}(x_i))^2 = e_i$$
What is the catch? *Complexity*

- If we can solve

\[
\min_{W', (x_i, e_i) \in W'} \left( \mathbb{E}_{x^*, y} \left[ \left( \hat{f}(\tilde{x}, \tilde{y}) - f(x^*) \right)^2 \right] + \eta \cdot \sum_{i} e_i \right)
\]

we can design \( p_i \left( (\tilde{x}, \tilde{y}(\tilde{e}')) \right), i \in W' \) to achieve the same optimal value but this is in general *NP-Hard*
What can we still do?

• For any approximation we can reach the same objective since we can decide $e'_i$ by designing $p_i \left( (\tilde{x}, \tilde{y}(\tilde{e}')) \right), i \in W'$

• If the $\{x_i\}$ are predetermined we can solve the optimal allocation of $x_i'$s to workers $w_j \in W$ by max-weight matching in polynomial time for a separable loss function
Conclusions

• Impressive and surprising result however.....

• The NP-hard fair dictator problem is very NP hard

• How would you know $\sigma_i(e_i)$ the authors give no example...