# Utility/Privacy Trade-off through the lens of Optimal Transport

Etienne Boursier & Vianney Perchet

(AISTATS, 2020)

Reporter: Fengjiao Gong Date: 2021-11-11/18

# Outline

## Utility/Privacy Trade-off through the lens of Optimal Transport

- Online Repeated auction Example
- Privacy Regularized Policy(PRP) Model
- Sinkhorn Loss with PRP
- DC program with PRP
- Experiments

# Example

## Online repeated auctions

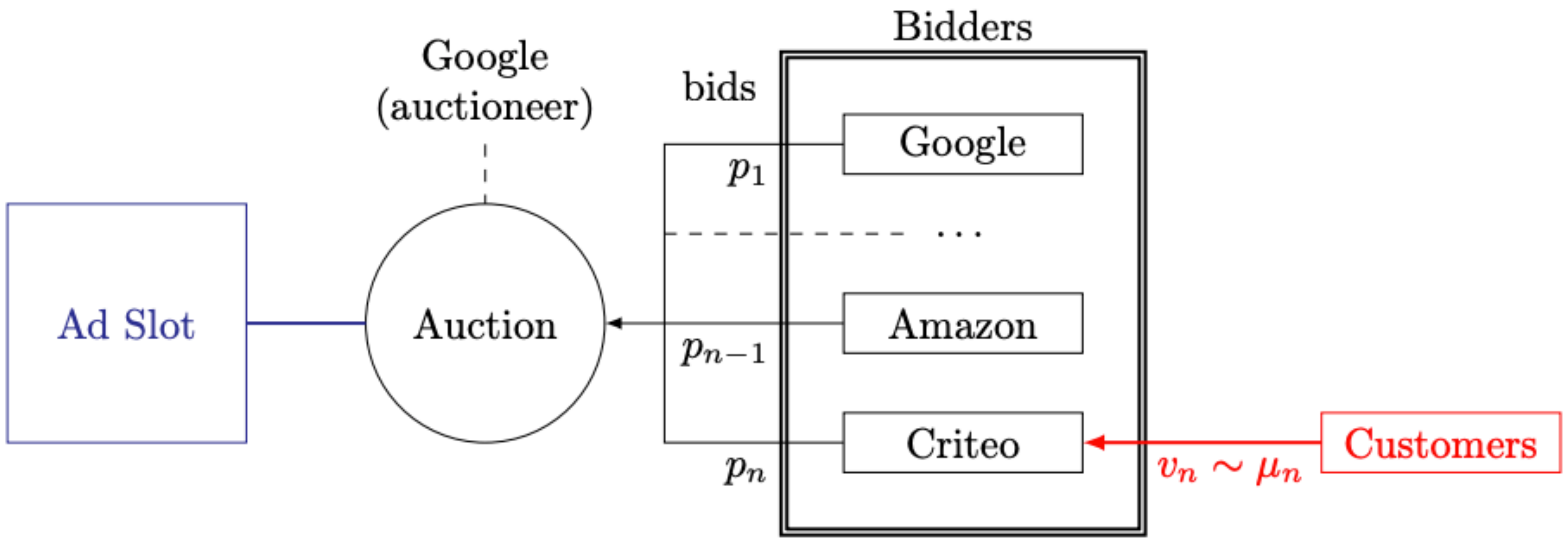



Figure 1: Online advertisement auction system.

## Scene:

auctioneer might bid the slot for its customer(so auctioneer is a direct concurrent bidder)

auctioneer is the only one who knows the true value

competitors observe a distribution of bids

If auctioneer doesn't consider the privacy, competitors can infer the true value.

# Example

## Online repeated auctions

**Suppose**:

$x$ is the agent's public action, $x \in \mathcal{X}$

type $k$ is the hidden information, $k \in [K]$

$c_k$ is the loss vector

## **Utility Loss** — without privacy concern

the **utility loss**

$$x^\top c_k$$

the optimal problem is

$$\min_{x \in \mathcal{X}} x^\top c_k$$

the **optimal solution** is

$$x_k^*$$

for each given $k$.

# Example

## Online repeated auctions

**Denote**:

$p_0$ — the prior of type $k$

$p_x$ — posterior distribution of the hidden type $k$

$\mu_k$ — the agent's strategy's probability distribution

**Privacy Loss:**

If the agent plays deterministically $x_k^*$ when type is k, then the adversary could infer the true value of $k$ based on the played action $x$ (**Bayes rule**).

So agent should hide the true type for the long-term utility, that is to **take a strategy** to control the amount of information given to the adversary.

# Example

## Online repeated auctions

**Measure information loss** — **KL divergence between prior and posterior**

$$\mathrm{KL}\left(p_x, p_0\right) = \sum_{k=1}^{K} \log\left(\frac{p_x(k)}{p_0(k)}\right) p_x(k)$$

where

$$p_x(k) = \frac{p_0(k)\mu_k(x)}{\sum_{l=1}^{K} p_0(l)\mu_l(x)}$$

classical cost of information in **economics**.

# Toy Example

**Total loss** = **utility loss** + **information loss**:

$$x^\top c_k + \lambda \mathrm{KL}\left(p_x, p_0\right) \quad (\lambda > 0)$$

**Global objective**:

$$\min_{\mu_1,\ldots,\mu_K} \sum_{k=1}^{K} p_0(k) \mathbb{E}_{x \sim \mu_k} \left[x^\top c_k + \lambda \mathrm{KL}\left(p_x, p_0\right)\right]$$

# Toy Example

**global objective**

$$\min_{\mu_1,\ldots,\mu_K} \sum_{k=1}^{K} p_0(k)\mathbb{E}_{x\sim\mu_k}\left[x^\top c_k + \lambda \mathrm{KL}\left(p_x, p_0\right)\right]$$

- if $\lambda = 0$ **totally revealing strategy**, the best strategy is to deterministically play $x_k^*$ given each $k$
- if $\lambda = \infty$ so called non-revealing strategy in game theory, the best strategy is to play

$$\arg\min_x x^\top c\left[p_0\right]$$

where

$$c\left[p_0\right] = \sum_{k=1}^{K} p_0(k)c_k$$

- if $0 < \lambda < \infty$ partially revealing strategy the behavior interpolates between two extreme strategies

# General Model

**Denote:**

the agent's strategy $\mathscr{Y} \to \mathscr{P}(\mathscr{X}) = X \mid Y \in \mathscr{P}(\mathscr{X})^{\mathscr{Y}}$

**utility loss** for playing $x$ with type $y$ $c(x, y)$

**information cost** $c_{priv}(X, Y) = \mathbb{E}_{x \sim X} D\left(p_x, p_0\right)$

where

$y \in \mathscr{Y}$ is the private type

$\mathscr{P}(\mathscr{X})$ is a set of distributions over $\mathscr{X}$

$(X, Y)$ is the joint distribution of action and type

$D\left(p_x, p_0\right)$ is the measurement of information cost

**Objective:**

$$\inf_{X \mid Y \in \mathscr{P}(\mathscr{X})^{\mathscr{Y}}} \mathbb{E}_{(x,y) \sim (X,Y)}[c(x, y)] + \lambda c_{priv}(X, Y) \tag{1}$$

# General Model

**Suppose:**

$\gamma$ is a **joint distribution** in $\mathscr{P}(\mathscr{X} \times \mathscr{Y})$

$\pi_{1\#}\gamma$ is marginal distribution of $X$,

$$\pi_{1\#}\gamma(A) = \gamma(A \times \mathscr{Y})$$

$\pi_{2\#}\gamma$ is marginal distribution of $Y$,

$$\pi_{2\#}\gamma(B) = \gamma(\mathscr{X} \times B)$$

here

$$\pi_{2\#}\gamma = p_0$$

**Privacy Regularized Policy**:

$$\inf_{\gamma \in \mathscr{P}(\mathscr{X} \times \mathscr{Y}), \pi_{2\#}\gamma = p_0} \int_{\mathscr{X} \times \mathscr{Y}} \left[ c(x, y) + \lambda D\left(p_x, p_0\right) \right] \mathrm{d}\gamma(x, y) \qquad (\text{PRP})$$

# PRP's theoretical properties

**Definition 1.**

$D$ is a ***f*-divergence** if for all distributions $P, Q$ such that $P$ is absolutely continuous w.r.t. $Q$,

$$D(P,Q) = \int_{\mathcal{Y}} f\left(\frac{\mathrm{d}P(y)}{\mathrm{d}Q(y)}\right) \mathrm{d}Q(y)$$

where $f$ **is a convex function** defined on $\mathbb{R}_+^*$ with $f(1) = 0$.

**common f-divergence:**

KL/ reverse KL /total variation distance

# PRP's theoretical properties

**Definition 1.**

$D$ is a ***f*-divergence** if for all distributions $P, Q$ such that $P$ is absolutely continuous w.r.t. $Q$,

$$D(P, Q) = \int_{\mathcal{Y}} f\left(\frac{\mathrm{d}P(y)}{\mathrm{d}Q(y)}\right) \mathrm{d}Q(y)$$

where $f$ **is a convex function** defined on $\mathbb{R}_+^*$ with $f(1) = 0$.

**Why f-divergence costs?**

1.considered in **non-Bayesian** cases

2.good properties of convexity, composition and post-processing invariance

# PRP's theoretical properties

**Theorem 1.**

If D is a f-divergence, PRP $\rightarrow$ a **convex minimization problem** in $\gamma \in \mathscr{P}(\mathscr{X} \times \mathscr{Y})$

Here,

suppose $D$ is always a ***f*-divergence** in the remaining part

minimum can be found by classical optimization methods such as **gradient descent**

# PRP's theoretical properties

**Theorem 1.**

If D is a f-divergence, PRP $\rightarrow$ a **convex minimization problem** in $\gamma \in \mathscr{P}(\mathscr{X} \times \mathscr{Y})$

Analyze

$\mathscr{P}(\mathscr{X} \times \mathscr{Y})$has generally an infinite dimension

$\mathscr{P}(\mathscr{X} \times \mathscr{Y})$ is **dimensionally finite** $\leftarrow$ if both sets $\mathscr{X}$ and $\mathscr{Y}$ are discrete

# PRP theoretical properties

## Discrete type

**Suppose** $p_0$ is a **discrete** prior of size $K$

$$p_0 = \sum_{k=1}^{K} p_0^k \delta_{y_k}$$

**Define** — (here $\mathscr{X}$ is an infinite space)

$$\mu_k(A) = \gamma\left(A \times \{y_k\}\right), \textbf{ for any } A \subset \mathscr{X}$$

$$\mu = \sum_{k=1}^{K} \mu_k = \pi_{1\#}\gamma$$

$$p^k(x) = \frac{\mathrm{d}\mu_k(x)}{\mathrm{d}\mu(x)}$$

**PRP** is equivalent to

$$\inf_{\substack{\mu, \left(p^k(\cdot)\right)_{1 \le k \le K} \\ p^k \ge 0, \sum_{l=1}^{K} p^l(\cdot) = 1}} \int_{\mathscr{X}} \left[ p^k(x) c\left(x, y_k\right) + \lambda p_0^k f\left(\frac{p^k(x)}{p_0^k}\right)\right] \mathrm{d}\mu(x)$$

such that $\forall k \le K$,

$$\int_{\mathscr{X}} p^k(x) \mathrm{d}\mu(x) = p_0^k$$

# PRP theoretical properties

Discrete type - **optimal solutions existence**

**Theorem 2.**

**If the prior is dicrete of size K**, for all $\epsilon > 0$, **(PRP) has an $\epsilon$-optimal solution** such that $\pi_1\#\gamma = \mu$ has a finite support of at most $K + 2$ points.

Furthermore, if X is compact and $c(\cdot, y_k)$ is lower semicontinuous for every k, then it also holds for $\varepsilon = 0$.

# PRP theoretical properties

Discrete type - optimal solutions existence

**Corollary 1.**

In the case of a discrete prior, (PRP) is equivalent to:

$$\inf_{(\gamma,x)\in\mathbb{R}_+^{(K+2)\times K}\times\mathcal{X}^{K+2}} \sum_{i,k} \gamma_{i,k} c\left(x_i, y_k\right) + \lambda \sum_{i,k} \gamma_{i,k} D\left(p_{x_i}, p_0\right)$$

such that $\forall k \leq K$,

$$\sum_{i} \gamma_{i,k} = p_0^k$$

where

$$\gamma_{i,k} := \gamma\left(\left\{\left(x_i, y_k\right)\right\}\right), \text{if } \gamma \in \left\{\left(x_i, y_k\right) \mid 1 \leq i \leq K+2, 1 \leq k \leq K\right\}.$$

PRP in Experiments

It is not jointly convex in $(\gamma, x)$.

# Sinkhorn Loss minimization

Discrete type - optimal solutions form

**Sinkhorn loss:**

$$\mathrm{OT}_{c,\lambda}(\mu,\nu) := \min_{\gamma\in\Pi(\mu,\nu)} \int c(x,y)\mathrm{d}\gamma(x,y) + \lambda \int \log\left(\frac{\mathrm{d}\gamma(x,y)}{\mathrm{d}\mu(x)\mathrm{d}\nu(y)}\right)\mathrm{d}\gamma(x,y) \tag{2}$$

where

$$\Pi(\mu,\nu) = \left\{\gamma \in \mathscr{P}(\mathscr{X}\times\mathscr{Y}) \mid \pi_{1\#}\gamma = \mu, \pi_{2\#}\gamma = \nu\right\}$$

given distributions

$$(\mu,\nu) \in \mathscr{P}(\mathscr{X})\times\mathscr{P}(\mathscr{Y})$$

here, the last part is **the regularization term** added to speed up computations.

# Sinkhorn Loss minimization

## Discrete type - optimal solutions form

### Sinkhorn Algorithm

Sinkhorn algorithm has a **linear convergence rate** to compute $OT_{c,\lambda}(\mu, \nu)$ for distributions

$$\mu = \sum_{i=1}^{n} \alpha_i \delta_{x_i}$$

$$\nu = \sum_{j=1}^{m} \beta_j \delta_{y_j}$$

the unique matrix $\gamma$ solution of the Problem (2) has the form $\mathrm{diag}(u) K \, \mathrm{diag}(v)$ in the discrete case, where

$$K_{i,j} = e^{-\frac{c\left(x_i, y_j\right)}{\lambda}}$$

it updates

$$(u, v) \leftarrow \left(\alpha / Kv, \beta / K^{\top} u\right)$$

for $n$ iterations or until convergence.

[Reference] *M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In Advances in Neural Information Processing Systems, pages* 2292 – 2300, 2013.

# Sinkhorn Loss minimization

## Discrete type - optimal solutions form

**PRP**:

$$\inf_{\mu \in \mathscr{P}(\mathscr{X})} \mathrm{OT}_{c,\lambda}\left(\mu, p_0\right)$$

posterior probability (Bayes rule)

$$\mathrm{d}p_x(y) = \frac{\mathrm{d}\gamma(x,y)}{\mathrm{d}\mu(x)}$$

where

$$\mathrm{D} = \mathrm{KL}$$
$$\nu = p_0$$

**with additional constraint**

$$\pi_{1\#}\gamma = \mu$$

Minimizing without this constraint is thus equivalent to minimizing the Sinkhorn loss over all action distributions $\mu$.

It is a new interpretation of Sinkhorn loss:

regularization term → privacy loss

# Sinkhorn Loss minimization

## Discrete type - optimal solutions form

**PRP**:

$$\inf_{\mu \in \mathscr{P}(\mathscr{X})} \mathrm{OT}_{c,\lambda}\left(\mu, p_0\right)$$

where

$$\mathrm{D} = \mathrm{KL}$$
$$\nu = p_0$$

**with additional constraint**

$$\pi_{1\#}\gamma = \mu$$

When $\mu$ and $\nu$ are both fixed, the optimal transport plan $\gamma^*$ remains the same.

But, here ***$\mu$* is varying**, and it is much more complex.

# Sinkhorn Loss minimization

## Discrete type - optimal solutions form

Consider discrete support, we can look for a distribution

$$\mu = \sum_{j=1}^{K+2} \alpha_j \delta_{x_j}$$

**minimization problem over tuple** $(\alpha, x)$

$$\inf_{(\alpha,x)\in\Delta_{K+2}\times\mathcal{X}^{K+2}} \mathrm{OT}_{c,\lambda}\left(\sum_{i=1}^{K+2} \alpha_i \delta_{x_i}, p_0\right) \tag{3}$$

Sinkhorn in Experiments

Given $\alpha$, Sinkhorn algorithm can get the unique optimal solution $\gamma^*$.

# Sinkhorn Loss minimization

## Discrete type - optimal solutions form

Consider discrete support, we can look for a distribution

$$\mu = \sum_{j=1}^{K+2} \alpha_j \delta_{x_j}$$

**minimization problem over tuple** $(\alpha, x)$

$$\inf_{(\alpha,x)\in\Delta_{K+2}\times\mathcal{X}^{K+2}} \mathrm{OT}_{c,\lambda}\left(\sum_{i=1}^{K+2} \alpha_i \delta_{x_i}, p_0\right) \tag{3}$$

Given $\gamma^*$, compute $\nabla \mathrm{OT}_{c,\lambda}$ to get optimal $\alpha^*$.

# Sinkhorn Loss minimization

## discrete minimization algorithem

### Gradient computation

Computing $\nabla \mathrm{OT}_{c,\lambda}$ is a known difficult task!

Solution - **the dual solution of Sinkhorn loss Problem**

It is **fast** as it does not need to store all the Sinkhorn iterations in memory and backpropagate through them afterwards.

**Convergence** of Sinkhorn algorithm is guaranteed to provide **an accurate approximation of the gradient.**

[*reference*] *G. Peyre and M. Cuturi. Computational optimal transport. Foundations and Trends in Machine Learning*, 11(5 – 6) : 355 – 607, 2019.

# DC program with PRP

linear utility cost over huperrectangle

**Definition**: standard DC program is of the form

$$min_{x \in \mathcal{X}} f(x) - g(x)$$

where both $f$ and $g$ are convex functions.

Methods:

- DCA (a local minimum)

$[reference]$ $P.Tao$ $and$ $L.An$ $Convex$ $analysis$ $approach$ $to$ $DC$ $programming$ : $Theory, algorithms$ $and$ $applications$ . $Acta$ $mathematatica$ $vietnamica$, 22(1) : 289 – 355, 1997.

# DC program with PRP

linear utility cost over huperrectangle

**Theorem 3.**

If $\mathscr{X} = \prod_{l=1}^{d} \left[a_l, b_l\right]$ and $c(x, y) = x^\top y$ then (PRP) is equivalent to the following *DC* program:

$$\min_{\gamma \in \mathbb{R}_+^{(K+2)\times K}} \lambda \sum_{i,k} p_0^k h_k\left(\gamma_i\right) - \sum_{i=1}^{K+2} \left\| \sum_{k=1}^{K} \gamma_{i,k} \phi\left(y_k\right) \right\|_1$$

such that $\forall k \leq K$,

$$\sum_{i=1}^{K+2} \gamma_{i,k} = p_0^k$$

with

$$\phi(y)^l := \left(b_l - a_l\right) y^l / 2$$

$$h_k\left(\gamma_i\right) := \left(\sum_{m=1}^{K} \gamma_{i,m}\right) f\left(\frac{\gamma_{i,k}}{p_0^k \sum_{m=1}^{K} \gamma_{i,m}}\right)$$

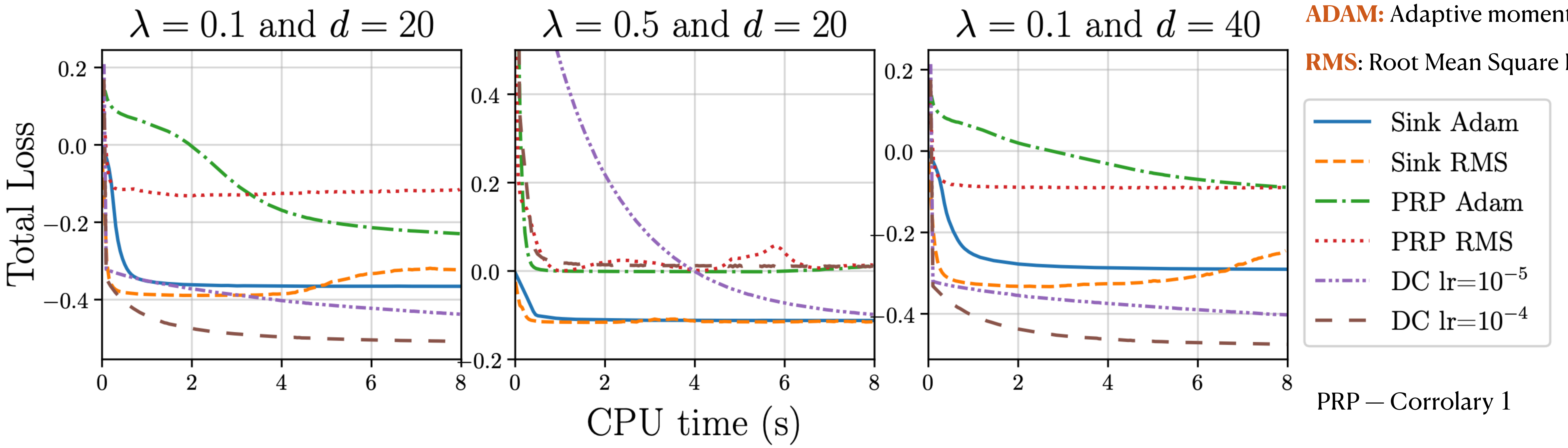
# Experiments

## convergence rates of usual non-convex optimization

gradient descent method

**ADAM:** Adaptive moment estimation

**RMS**: Root Mean Square Propagation




PRP — Corrolary 1

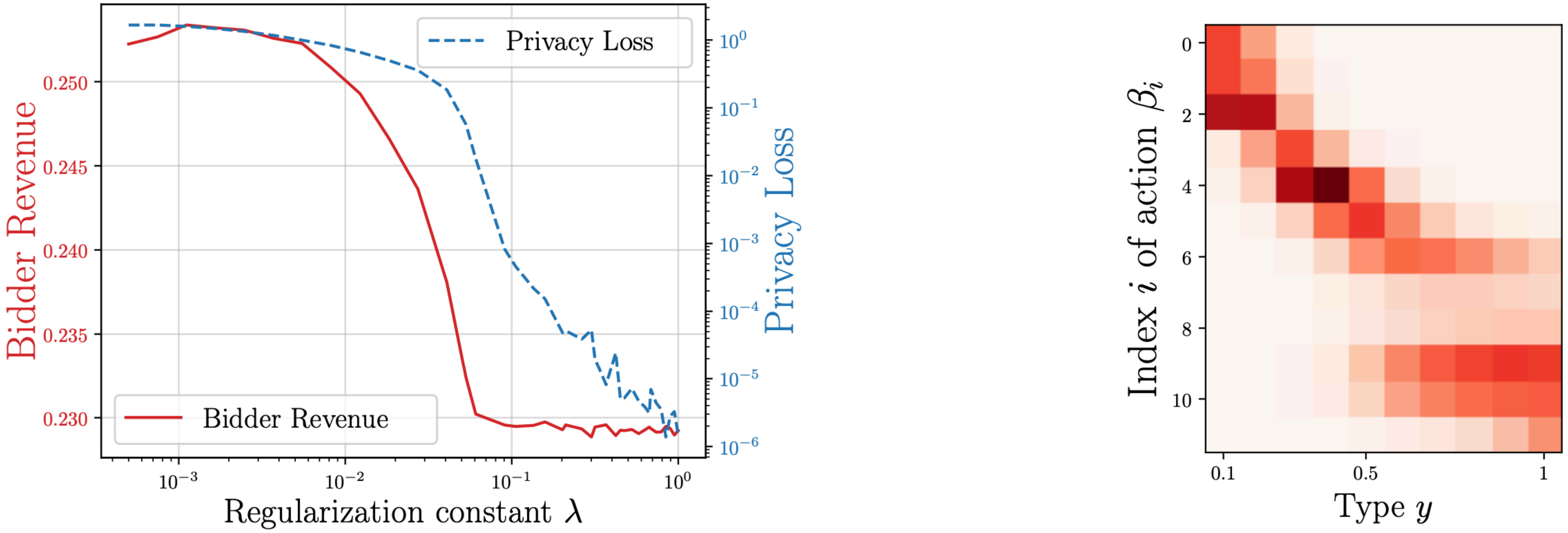
Sink — Equation (3)

Figure 2: Comparison of optimization schemes.

- PRP is **more sensitive to problem parameters**
- for **larger values of $\lambda$**, Sinkhorn performs well when the privacy cost is predominant
- for **larger values of $d$**, PRP converges to **worse spurious local minima**
- DC finds better local minima than the other ones

# Experiments

## utility-privacy in repeated auctions




(a) Evolution of privacy-utility with $\lambda$.

(b) Joint distribution map for $\lambda = 0.01$. The intensity of a point $(i, j)$ corresponds to the value of $\gamma(\beta_i, y_j)$.

Figure 3: Privacy-utility trade-off in online repeated auctions.

- Figure 3$a$ both **decrease with $\lambda$**, significantly drop at a critical point near 0.05, which can be seen as the cost of information here.
- Figure 3$b$ **partially revealing strategy** that randomizes the type over neighboring types and reveals more information when the revenue is sensible to action

# Experiments

## utility-privacy in repeated auctions

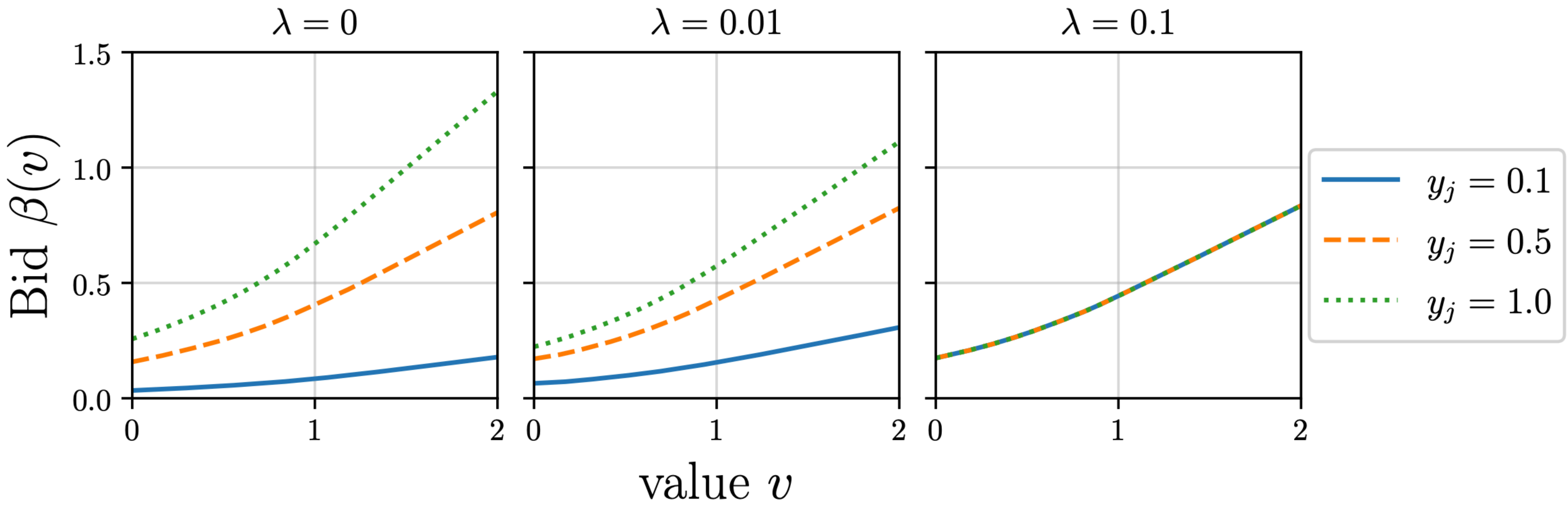



Figure 4: Evolution of the bidding strategy with the type and the regularization constant.

- revealing strategy — action significantly scales with type
- partially revealing strategy — action scales less with type

# Thanks!