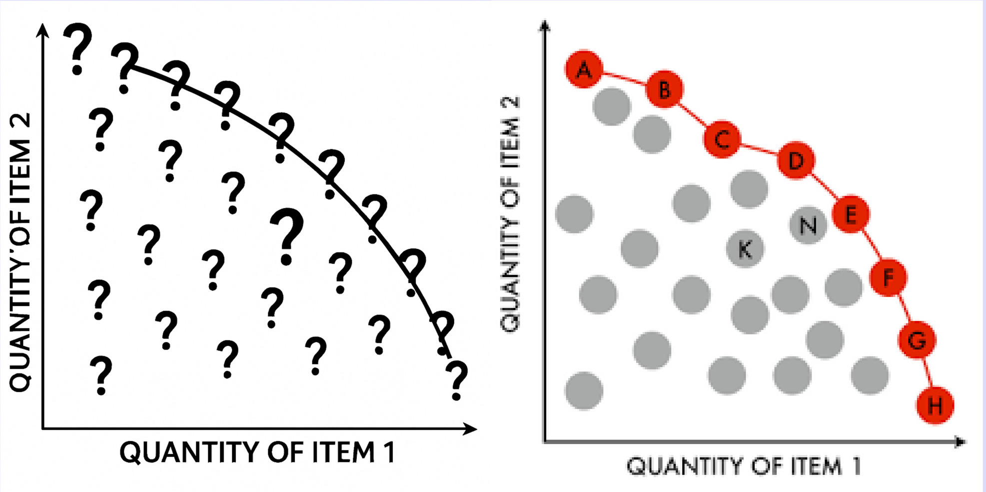
Learning the Pareto Front Using Bootstrapped Observation Samples

Wonyoung Kim / Garud Iyengar / Assaf Zeevi

Chung-Ang University / Columbia University / Columbia University

wyk7@cau.ac.kr / garud@ieor.columbia.edu / assaf@gsb.columbia.edu

1. Pareto Front Identification for Linear Bandits



- At each decision point, the algorithm chooses one of offers (contexts) and receives the corresponding **vector feedback** (rewards) from the customer. The goal is to find the best offer whose rewards are on the Pareto Front.
- **Applications:**
  - Prescribing a drug to some patients (contexts) yields multiple responses (rewards) including efficacy, toxicity, and potentially all its side.
  - Recommender systems must find some offers (contexts) that yields good feedback (rewards) in terms of price, design, and practicality.
- **Challenging point:**
  - Tradeoff between **exploration** (inference for vector rewards) and **exploitation** (maximizing the current reward)
  - Unlike the best arm identification, the number of vectors on the Pareto Front is unknown.

2. Main Contributions

- We propose an algorithm that achieves nearly **optimal sample complexity** and **optimal regret bound** among all algorithms that identifies all Pareto Fronts.
- We introduce a novel estimation procedure for linear bandit feedback that ensures **fast convergence rate for the reward vectors of all arms while largely exploiting low regret arms.**
- Experiment results show that our estimator converges on the rewards of all contexts while exploiting low-regret arms and our algorithm achieves both Pareto Front Identification and regret minimization.

3. Problem Formulation

- An action $k \in \{1, \dots, K\} := [K]$ is associated with a known d-dimensional context vector $x_k \in \mathbb{R}^d$. In period t, the decision-maker chooses an $a_t \in [K]$, and observes a sample of the random reward vector $\boldsymbol{y}_{a_t,t} = \Theta_\star^\top x_{a_t} + \eta_t$, where $\Theta_\star := (\theta_\star^{(1)}, \dots, \theta_\star^{(L)}) \in \mathbb{R}^{d\times L}$ is the unknown (but fixed) parameters and $\eta_t \in \mathbb{R}^L$ is a mean-zero, σ-sub-Gaussian random error vector.
- Let $y_k := \Theta_\star^\top x_k$ denote the true mean reward vector for arm $k \in [K]$. We want to identify the Pareto front $\mathcal{P}_\star := \{k \in [K] | \nexists k' : y_k \prec y_{k'}\}$ where $a \prec b$ represents the domination, i.e., $a_\ell \le b_\ell$ for all $\ell \in [L]$. The Pareto front $\mathcal{P}_\star$ is the set of arms whose mean reward vector is not dominated by the reward of any other arm.
- Let $\Delta_k^\star := \max_{k_\star \in \mathcal{P}_\star} \max\{0, \min_{\ell\in[L]}(y_{k_\star}^{\langle \ell \rangle} - y_k^{\langle \ell \rangle})\}$ denote the amount by which each component of the reward vector y_k must be increased to ensure that action k is not dominated by any Pareto optimal action $k_\star \in \mathcal{P}_\star$.
- (PFI success condition) For precision $\epsilon > 0$ and confidence $\delta \in (0,1)$, an algorithm must output a set of arms $\mathcal{P} \subseteq [K]$ such that, with probability at least $1-\delta$,

$$\mathcal{P}_\star \subseteq \mathcal{P} \;\text{ and }\; \Delta_k^\star \le \epsilon, \text{ for all } k \in \mathcal{P}_\star \setminus \mathcal{P}$$

3.1 Theorem 1

For any algorithm, the PFI condition requires at least $(\sigma^2/3)\sum_{k=1}^d \Delta_{(k),\epsilon}^{-2} \log(3L/4\delta)$ number of samples.

4. A Context Basis

- Let $X = [x_1, \dots, x_K] \in \mathbb{R}^{d\times K}$ denote the matrix of contexts vectors. Using the (reduced) singular value decomposition (SVD), one can compute $X = \sum_{i=1}^d \lambda_i u_i v_i^\top$ and it follows that $v_i^\top X^\top \theta_\star^{(\ell)} = \sqrt{\lambda_i} u_i^\top \theta_\star^{(\ell)}$ for $\ell \in [L]$ and $i \in [d]$.
- For $i \in [d]$, define a probability mass function, $\pi_k^{(i)} = |v_{ik}|/\|v_i\|_1$ over actions $k \in [K]$. Then, for a randomized action $a \sim \pi^{(i)}$, we have

$$\mathbb{E}\left[\|v_i\|_1 \text{sign}(v_{ia}) Y_{a,s}^{(\ell)}\right] = \mathbb{E}\Big[\sum_{k=1}^K v_{ik} Y_{k,s}^{(\ell)}\Big] = \sum_{k=1}^K v_{ik} x_k^\top \theta_\star^{(\ell)} = v_i^\top X^\top \theta_\star^{(\ell)} = (\sqrt{\lambda_i} u_i)^\top \theta_\star^{(\ell)},$$

- Thus, $\|v_i\|_1 \text{sign}(v_{ia}) Y_{a,s}^{(\ell)}$ can be viewed as the random reward corresponding to the **context basis** $\sqrt{\lambda_i} u_i$. Sampling $a_i \sim \pi^{(i)}$ for $i \sim \text{unif}([d])$ yields the expected design matrix $d^{-1}\sum_{i=1}^d \lambda_i u_i u_i^\top = d^{-1}\sum_{k=1}^K x_k x_k^\top$ that satisfies $\max_{k\in[K]} \|x_k\|^2_{(d^{-1}\sum_{k'=1}^K x_{k'} x_{k'}^\top)^{-1}} \le d$. This design yields a tighter bound than the G-optimal design that is widely used in BAI problems.

Paper

5. Recycling Reward Samples in the Exploration Phase

- We construct our exploration set:

$$\mathcal{E}_t := \begin{cases} \mathcal{E}_{t-1} & \sum_{u\in\mathcal{E}_t} \mathbb{I}(\check{a}_u = \check{a}_t) > \frac{\gamma_t}{t}\sum_{s=1}^t \mathbb{I}(\check{a}_s = \check{a}_t) \\ \mathcal{E}_{t-1} \cup \{t\} & \text{otherwise} \end{cases}$$

- We **recycle** the reward sample observed in a previous exploration round by bootstrapping. Let $\mathcal{E}_t(\check{a}_t) = \{s \in \mathcal{E}_t : a_s = \check{a}_t\}$ denote the set of previous exploration rounds where the action $\check{a}_t$ was chosen. For the exploitation rounds $\tau \in [t-1] \setminus \mathcal{E}_{t-1}$, let $\check{n}_\tau$ denote time index of the exploration sample **recycled** at exploitation round τ and "mixed" with the chosen action a_τ.
- We "mix" the action a_t with the exploration sample **recycled** from round $\check{n}_t := \arg\min_{n\in\mathcal{E}_t(\check{a}_t)} \sum_{\tau\in[t-1]\setminus\mathcal{E}_{t-1}} \mathbb{I}\big(\check{n}_\tau = n\big)$, i.e. we want to balance the reuse choice over the set $\mathcal{E}_t(\check{a}_t)$.
- We define the exploration-mixed contexts and rewards as follows: for all $\ell \in [L]$, and $w_t, \check{w}_t \sim \text{unif}[-\sqrt{3}, \sqrt{3}]$ sampled independently,

$$\check{X}_{a_t,t} := w_t x_{a_t} + \check{w}_t \sqrt{\lambda_{i_t}} u_{i_t}, \quad \check{Y}_{a_t,t}^{\langle \ell \rangle} := w_t Y_{a_t,t}^{(\ell)} + \check{w}_t \|v_{i_t}\|_1 \text{sign}(v_{i_t, a_{\check{n}_t}}) Y_{a_{\check{n}_t},\check{n}_t}^{(\ell)}.$$

- Then we define the **exploration-mixed estimator**,

$$\check{\theta}_t^{\langle \ell \rangle} := \Big(\sum_{s\in\mathcal{E}_t} x_{a_s} x_{a_s} + \sum_{s\in[t]\setminus\mathcal{E}_t} \tilde{X}_{a_s,s}\tilde{X}_{a_s,s}^\top + \frac{1}{2}I_d\Big)^{-1}\Big(\sum_{s\in\mathcal{E}_t} x_{a_s} Y_{a_s,s}^{(l)} + \sum_{s\in[t]\setminus\mathcal{E}_t} \tilde{X}_{a_s,s}\tilde{Y}_{a_s,s}\Big)$$

.

- The dependency caused by the recycling rewards are controlled by the **doubly robust estimation**.

6. Doubly Robust Estimation

- We first reduce K rewards into $d+1$ rewards $\tilde{Y}_{i,t}^{\langle \ell \rangle} := \sum_{k=1}^K v_{i,k} Y_{k,t}^{(\ell)}$ corresponding to the d context basis $\sqrt{\lambda_i} u_i$, $i = 1, \dots, d$, and $\tilde{Y}_{d+1,t}^{\langle \ell \rangle} := Y_{a_t,t}^{(\ell)}$.
- Then $\{\tilde{Y}_{i,t}^{\langle \ell \rangle} : i \in [d]\}$ is missing and $\tilde{Y}_{d+1,t}$ is observable. We induce the probability mass function $\tilde{\pi}_i = 1/(2d),\ \forall i = 1, \dots, d$ and $\tilde{\pi}_{d+1} = 1/2$. To couple the observed reward $Y_{a_t,t}^{(\ell)}$ and the randomly selected reward $\tilde{Y}_{\tilde{a}_t,t}^{\langle \ell \rangle}$, we resample both action a_t and pseudo-action $\tilde{a}_t$ until the matching event $\{Y_{a_t,t}^{(\ell)} = \tilde{Y}_{\tilde{a}_t,t}^{\langle \ell \rangle}\} = \{\tilde{a}_t = d+1\}$ happens.
- For given $\delta' \in (0,1)$, let $\mathcal{M}_t$ denote the event of obtaining the matching $\{Y_{a_t,t}^{(\ell)} = \tilde{Y}_{\tilde{a}_t,t}^{\langle \ell \rangle}\}$ within $\rho_t := \log((t+1)^2/\delta')/\log(2)$ number of resampling so that the event $\mathcal{M}_t$ happens with probability at least $1 - \delta'/(t+1)^2$.
- Define new contexts $\tilde{x}_{i,t} := \sqrt{\lambda_i} u_i$ for $i = 1, \dots, d$ and $\tilde{x}_{d+1,t} := x_{a_t,t}$. Then we construct the pseudo-rewards for the missing rewards as:

$$\widehat{Y}_{i,t}^{\langle \ell \rangle} := \tilde{x}_{i,t}^\top \check{\theta}_t \ell + \frac{\mathbb{I}(\tilde{a}_t = i)}{\tilde{\pi}_i}(\tilde{Y}_{i,t}^{\langle \ell \rangle} - \tilde{x}_{i,t}^\top \check{\theta}_t^{\langle \ell \rangle}).$$

- We define our **DR-mix estimator** as a ridge estimator using $\{(\widehat{Y}_{i,s}^{\langle \ell \rangle}, \tilde{x}_{i,s}) : s = 1, \dots, t, i = 1, \dots, d+1\}$:

$$\widehat{\theta}_t^{(\ell)} = \Big(\sum_{s:\mathbb{I}(\mathcal{M}_s)=1}^t \sum_{i=1}^{d+1} \tilde{x}_{i,s}\tilde{x}_{i,s}^\top + I_d\Big)^{-1}\Big(\sum_{s:\mathbb{I}(\mathcal{M}_s)=1}^t \sum_{i=1}^{d+1} \tilde{x}_{i,s}\widehat{Y}_{i,t}^{\langle \ell \rangle}\Big).$$

7. Pareto Front Identification with Regret Minimization

1: **INPUT:** context matrix $X = [x_1, \dots, x_K]$, accuracy parameter $\epsilon > 0$, confidence $\delta > 0$.
2: Set $\mathcal{A}_0 = [K]$, $\mathcal{P}_0 = \mathcal{E}_0 = \emptyset$ and $\widehat{\theta}_0^{\langle \ell \rangle} = \mathbf{0}_d$, for all $\ell \in [L]$ and apply reduced SVD on $X = \sum_{i=1}^d \lambda_i u_i v_i^\top$.
3: **while** $\mathcal{A}_t \neq \emptyset$ **do**
4: Sample $i_t \sim \text{unif}([d])$ and $\check{a}_{i_t} \sim \pi^{(i_t)}$ and update $\mathcal{E}_t$
5: **If** $t \in \mathcal{E}_t$ **then** set $a_t = \check{a}_{i_t}$ **else** randomly sample a_t over $\{k \in \mathcal{A}_{t-1} : \nexists k' \in \mathcal{A}_{t-1}, \widehat{y}_{k,t} \prec \widehat{y}_{k',t}\}$
6: Compute the DR-mix estimator $\widehat{\theta}_t^{\langle \ell \rangle}$ and $\widehat{y}_{k,t}^{\langle \ell \rangle} := x_k^\top \widehat{\theta}_t^{\langle \ell \rangle}$ and the estimated distances:

$$\widehat{m}_t(k,k') := \min\{\alpha \ge 0 | \exists \ell \in [L] : \widehat{y}_{k,t}^{\langle \ell \rangle} + \alpha \ge \widehat{y}_{k',t}^{\langle \ell \rangle}, \widehat{M}_t^{2\epsilon}(k,k') := \min\{\alpha \ge 0 \big| \forall \ell \in [L] : \widehat{y}_{k,t}^{\langle \ell \rangle} + 2\epsilon \le \widehat{y}_{k',t}^{\langle \ell \rangle} + \alpha\}.$$

7: Compute the confidence intervals:

$$\beta_{k,t} := 3\|x_k\|_{F_t^{-1}}\Big(\theta_{\max} + \sigma\sqrt{d\log\frac{7Lt}{\delta}}\Big)\ |\mathcal{A}_t| > d, \quad 3\|x_k\|_{F_t^{-1}}\Big(\theta_{\max} + 3\sigma\sqrt{\log\frac{56Ldt^2}{\delta}}\Big)\ |\mathcal{A}_t| \le d.$$

8: Estimate Pareto front

$$\mathcal{C}_t := \{k \in \mathcal{A}_{t-1} | \forall k' \in \mathcal{A}_{t-1}\cup\mathcal{P}_{t-1} : \widehat{m}_t(k,k') \le \beta_{k,t} + \beta_{k',t}\}, \quad \mathcal{P}_t^{(1)} := \{k \in \mathcal{C}_t | \forall k' \in \mathcal{C}_t \cup \mathcal{P}_{t-1} \setminus \{k\} : \widehat{M}_t^{2\epsilon}(k,k') \ge \beta_k$$

9: Update $\mathcal{P}_t \leftarrow \mathcal{P}_{t-1} \cup \mathcal{P}_t^{(1)}$ and $\mathcal{A}_t \leftarrow \mathcal{C}_t \setminus \mathcal{P}_t^{(1)}$.

8. Theoretical and Experimental Results

8.1 Theorem 1

The sample complexity of our proposed method is $O\Big(\sum_{k=1}^d \frac{(\theta_{\max}+\sigma)^2}{\Delta_{(k),\epsilon}^2}\log\frac{(\theta_{\max}+\sigma)dL}{\Delta_{(k),\epsilon}\delta}\Big)$, where $\Delta_{(k),\epsilon}$ is the problem-dependent gap.

8.2 Theorem 2

The cumulative regret of our proposed method is $\bar{O}\Big(\theta_{\max} d^3 \log\frac{\theta_{\max} d}{\delta\Delta_{(1),\epsilon}} + \frac{\theta_{\max} d\sigma}{\Delta_\epsilon^\star}\log\frac{\theta_{\max} d\sigma}{\Delta_\epsilon^\star\delta}\Big)$. This rate is **optimal** among all algorithms that satisfies PFI success condition.

Comparison of PFIwR (proposed) and MultiPFI (Auer et al.,2016) on the SW-LLVM dataset.
Both algorithms satisfies PFI success condition on all 500 independent experiments.