High-dimensional Linear Bandits with Knapsacks

Wanteng Ma, Dong Xia, and Jiashuo Jiang

Presented by Wanteng Ma

Department of Mathematics
The Hong Kong University of Science and Technology

Online Resource Allocation

Consider the online allocation problem that at each time $t \in [T]$, we have:

- a reward function $r_t(x)$, $x \in [K]$ for K types of products
- together with consumption of m types of resources $b_t(x) \in \mathbb{R}^m$.

Our goal is to:

- maximize the total reward $\sum_{t=1}^{T} r_t(x_t)$
- subject to the resource constraint $\sum_{t=1}^{T} b_t(x_t) \preceq \boldsymbol{C} \in \mathbb{R}^m$

Online allocation can represent tasks like:

- online advertising/revenue management/matching/bidding, etc

High-dimensional Bandit with Knapsacks (BwK)

However, in many problems[1] like:

- user-specific recommendations
- personalized treatments,

we are facing allocation with

- high-dimensional contexts
- unknown rewards

This problem can be formulated by the high-dimensional contextual bandit with knapsacks (CBwK).

- At each time t, observe a high-dimensional context/feature $Z_t \in \mathbb{R}^d$, for very large d
- K arms with hidden parameters $\mu_a^\star \in \mathbb{R}^d$, $a \in [K]$
- After pulling arm $x_t \in [K]$, observe linear reward r_t and cost b_t
- Decision can be accepted only when $\sum_{j=1}^{t} b_j(x_j) \preceq \boldsymbol{C}$

[1] Hamsa Bastani and Mohsen Bayati. “Online decision making with high-dimensional covariates”. In: *Operations Research* 68.1 (2020), pp. 276–294.

High-dimensional Contextual BwK

Linear contextual BwK with

- $r_t(x_t, Z_t) = \sum_{a \in [K]} \langle \mu_a^\star, Z_t \rangle \cdot x_{a,t} + \xi_t$
- $b_t(x_t, Z_t) = \sum_{a \in [K]} W_a^\star Z_t \cdot x_{a,t} + \omega_t$

Our goal is to maximize $\sum_{t=1}^{T} r_t(x_t, Z_t)$, with $\sum_{t=1}^{T} b_t(x_t, Z_t) \preceq \boldsymbol{C}$

- $\mu_a^\star \in \mathbb{R}^d$: unknown s_0-sparse parameters of arms, $a \in [K]$
- $W_a^\star \in \mathbb{R}^{m \times d}$: unknown row-wise s_0-sparse parameters for cost
- $Z_t \in \mathbb{R}^d$: contexts (features) that we can observe. $\Sigma = \mathbb{E} Z_t Z_t^\top$
- $x_{a,t} \in \{0, 1\}$: decision variables. $\sum_{a \in [K]} x_{a,t} = 1$
- ξ_t, $\omega_t \in \mathbb{R}^m$: 0-mean sub-Gaussian noises

High-dimensional Contextual BwK

Our contributions:

- A new sparse estimation algorithm, Online HT, that is comparable with LASSO (statistically optimal) but runs **fully online** (computationally fast).
- A primal-dual framework based on Online HT to solve high-dimensional BwK, with **only** $\log d$**-dependent** regret.
- Application to high-dimensional bandit problems and reaches **optimal regrets** $\widetilde{O}(s_0^{2/3}T^{2/3})$ and $\widetilde{O}(\sqrt{s_0 T})$, in both data-poor and data-rich regimes respectively, which satisfies the so-called "**the best of two worlds**"[2]

[2] Botao Hao, Tor Lattimore, and Mengdi Wang. "High-dimensional sparse linear bandits". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 10753–10763.

Online Hard Thresholding (Online HT)

Consider optimal estimation of $\mu_a^\star$ for one a. It amounts to solving:

$$\min_{\|\mu\|_0 \le s_0} f(\mu) := \mathbb{E}(r_t - \mu^\top Z_t)^2 = \|\mu - \mu_a^\star\|_\Sigma^2 + \sigma_\xi^2.$$

To solve this stochastic programming, we have:

- Sample Average Approximation (SAA)
  - For batch setting
  - LASSO is massively used in the literature
  - Heavily relies on resolving
- Stochastic Approximation (SA)
  - For online setting
  - Computationally fast (online gradient descent)
  - Largely underexplored for high-dimensional sparse estimation

Online Hard Thresholding for a Fixed Arm

Online hard thresholding iteration:

- Single stochastic gradient:

$$\nabla f_t(\mu_{a,t}) = 2Z_t Z_t^\top (\mu_{a,t} - \mu_a^\star) - 2Z_t \xi_t$$

- Averaged gradient from 1 to t:

$$g_t = \frac{1}{t}\sum_{j=1}^{t} \nabla f_j(\mu_{a,t})$$

- Choose a slightly larger $s > s_0$ and perform hard thresholding after gradient descent:

$$\mu_{a,t} = \mathcal{H}_s(\mu_{a,t-1} - \eta_t g_t)$$

- Project back to the exact s_0 sparse parameter for estimation

$$\mu_{a,t}^{\mathsf{s}} = \mathcal{H}_{s_0}(\mu_{a,t})$$

Hard Thresholding Operator

Q: Why do we need the gradient averaging?

A: Because the hard thresholding operator shares poor smoothness property

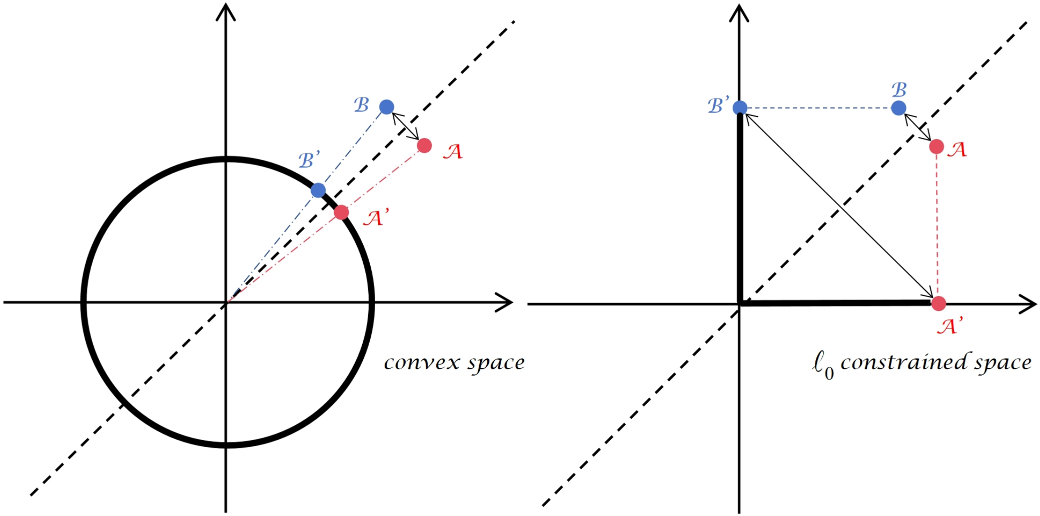

Figure: Poor smoothness property of ℓ_0-constrained projection

Online Hard Thresholding (Online HT)

Q: Why do we choose a slightly larger $s > s_0$

A: Because we need to preserve enough information against the hard thresholding operator

$$\mu - \eta_t g_t = \begin{bmatrix} \mu_1 - \eta_t g_{t,1} \\ \mu_2 - \eta_t g_{t,2} \\ \vdots \\ \mu_d - \eta_t g_{t,d} \end{bmatrix} \Longrightarrow \mathcal{H}_s(\mu - \eta_t g_t) = \left.\begin{bmatrix} 0 \\ \mu_2 - \eta_t g_{t,2} \\ 0 \\ \vdots \\ 0 \end{bmatrix}\right\} \text{only } s\text{-sparse}$$

- The iteration suffers from massive gradient information loss
- But larger s allows us to keep enough information for online learning

Online Hard Thresholding (Online HT)

- If we consider **simultaneously** estimating K arms, ε-greedy methods with importance sampling are required.

Theorem

Suppose ε_j is the lower bound of sampling each arm at time j, then choosing proper s, we have

$$\mathbb{E} \max_{a \in [K]} \left\| \mu^{\mathsf{s}}_{a,t} - \mu^{\star}_{a} \right\|_2^2 \lesssim \frac{\sigma^2 s_0}{\phi^2_{\min}(s)} \frac{\log(dK)}{t^2} \left(\sum_{j=1}^{t} \frac{1}{\varepsilon_j} \right)$$

- Online HT needs **linear** computations and is statistically **optimal**.

Algorithm Framework to Solve High-dimensional CBwK

Require: Dual variable $\eta_0 = \mathbf{1}/m$, ratio R, initial estimate $\mu^{\mathsf{s}}_{a,0}$, $\widehat{W}_{a,0}$

for $t = 1, ..., T$ **do**

Observe the feature Z_t

Compute $\mathrm{EstCost}(a) = Z_t^\top \widehat{W}_{a,t-1}^\top \eta_t$ for each arm $a \in [K]$

Sample a r.v. $\nu_t \sim \mathrm{Ber}(K\epsilon_t)$, and pull the arm x_t as:

$$x_t = \begin{cases} \mathrm{argmax}_{a\in[K]}\{\left\langle \mu^{\mathsf{s}}_{a,t-1}, Z_t \right\rangle - R \cdot \mathrm{EstCost}(a)\}, & \text{if } \nu_t = 0 \\ a, \quad \text{w.p. } 1/K \text{ for each arm } a \in [K] & \text{if } \nu_t = 1. \end{cases}$$

If one of the constraints is violated, then EXIT

Update η_t following Hedge algorithm

For each arm $a \in [K]$, obtain the estimate $\mu^{\mathsf{s}}_{a,t}$, $\widehat{W}_{a,t}$ from Online HT

end for

Solve High-dimensional CBwK

Theorem

if R satisfies $\frac{\mathsf{OPT}}{C_{\min}} \le Z \le O\left(\frac{\mathsf{OPT}}{C_{\min}} + 1\right)$, then the regret can be upper bounded by

$$\begin{aligned}\mathrm{Regret}(\pi) \le & O\left(\frac{\mathsf{OPT}}{C_{\min}} + 1\right) \cdot \sqrt{T \cdot \log m} \\ & + O\left(\phi_{\min}^{-\frac{2}{3}}(s) \cdot \left(\frac{\mathsf{OPT}}{C_{\min}} + 1\right)^{\frac{1}{3}} K^{\frac{1}{3}} \sigma^{\frac{2}{3}} s_0^{\frac{2}{3}} T^{\frac{2}{3}} (\log(dmK))^{\frac{1}{3}}\right)\end{aligned}$$

- Two-phase: $\mathrm{Regret}(\pi) = \tilde{O}\left(\frac{\mathsf{OPT}}{C_{\min}}\sqrt{T} + \left(\frac{\mathsf{OPT}}{C_{\min}}\right)^{\frac{1}{3}} T^{\frac{2}{3}}\right)$, with $\frac{\mathsf{OPT}}{C_{\min}} = T^{\frac{1}{4}}$ as the transition point.
- If $\frac{\mathsf{OPT}}{C_{\min}} \precsim T^{\frac{1}{4}}$, resource-abundant, primal information will be the barrier, which leads to $\mathrm{Regret}(\pi) = \tilde{O}\left(\left(\frac{\mathsf{OPT}}{C_{\min}}\right)^{\frac{1}{3}} T^{\frac{2}{3}}\right)$
- If $\frac{\mathsf{OPT}}{C_{\min}} \succsim T^{\frac{1}{4}}$, resource-deficient, dual information will be the barrier, $\mathrm{Regret}(\pi) = \tilde{O}\left(\frac{\mathsf{OPT}}{C_{\min}}\sqrt{T}\right)$

Diverse Covariate

However, the primal barrier can be breached if the covariates are diverse enough. Given the following Assumption, we will have

$$\text{Regret} \leq O\left(\frac{\mathsf{OPT}}{C_{\min}} \cdot \sqrt{T \cdot \log m}\right)$$

Assumption (Diverse covariate)

There are positive constants $\gamma(K)$ and $\zeta(K)$, such that for any unit vector $v \in \mathbb{R}^d$, $\|v\|_2 = 1$ and any $a \in [K]$, conditional on the history $\mathcal{H}_{t-1}$, there is

$$\mathbb{P}\left(v^\top Z_t Z_t^\top v \cdot \mathbb{I}\left\{x_t = a\right\} \geq \gamma(K) \middle| \mathcal{H}_{t-1}\right) \geq \zeta(K),$$

where $x_t = \text{argmax}_{a\in[K]}\{(\mu^{\mathsf{s}}_{a,t-1})^\top Z_t - R \cdot EstCost(a)\}$ is primal-dual choice.

- A primal-dual version of the diverse covariate condition for greedy algorithms[3].
- Typically met in the online allocation problem where the optimal strategy is often a distribution within arms, rather than a single arm[4]

[3] Zhimei Ren and Zhengyuan Zhou. "Dynamic batch learning in high-dimensional sparse linear contextual bandits". In: *Management Science* (2023).

[4] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. "Bandits with knapsacks". In: *Journal of the ACM (JACM)* 65.3 (2018), pp. 1–55.

Application to High-dimensional Bandit

We can also apply our Online HT to high-dimensional bandit problems.

Algorithm High Dimensional Bandit by Online HT

Require: ϵ-greedy sampling probability ϵ_t for each t

for all $t = 1, ..., T$ **do**

Observe the feature Z_t.

Sample a random variable $\nu_t \sim \text{Ber}(K\epsilon_t)$.

Pull the arm x_t with ϵ_t-greedy strategy defined as follows:

$$x_t = \begin{cases} \arg\max_{a \in [K]} \left\langle Z_t, \mu^{\mathsf{s}}_{a,t-1} \right\rangle, & \text{if } \nu_t = 0 \\ a, \quad \text{w.p. } 1/K \text{ for each arm } a \in [K] & \text{if } \nu_t = 1, \end{cases}$$

and receive a reward r_t.

For each arm $a \in [K]$, update the sparse estimate $\mu^{\mathsf{s}}_{a,t}$ by Online HT with each $p_{a,t} = (1 - K\epsilon_t)x_{a,t} + \epsilon_t$

end for

We can achieve "**the best of two worlds**" under our unified framework.

- For general conditions, choose $\epsilon_t = \Theta(t^{-\frac{1}{3}})$. Regret can be controlled by $\widetilde{O}\left(s_0^{\frac{2}{3}} T^{\frac{2}{3}}\right)$, which is optimal in the data-poor regime
- Given the diverse covariate condition, choose $\epsilon_t = 0$. Regret can be controlled by $\widetilde{O}\left(\sqrt{s_0 T}\right)$, which is optimal in the data-rich regime

Summary

A brief summary:

- Online HT, a powerful and efficient online sparse estimation approach
  - Computational cost: $O(d^2T)$ for Online HT ($O(d^3T + d^2T^2)$ for LASSO resolving)
  - Statistically optimal
- Apply Online HT to solve high-dimensional BwK, with $\log d$ dependent regret
- Application to high-dimensional bandit problem, with optimal regret in both data-poor and data-rich regimes

Thank you for listening!