Dynamic Batch Learning in High-Dimensional Sparse Linear Contextual Bandits

Zhimei Ren

MOILS, November 30th 2020

Collaborator

Zhengyuan Zhou (NYU stern)

Background: Linear Contexual Bandits

- Sequential decision making problem.

Background: Linear Contexual Bandits

- Sequential decision making problem.
- Time horizon: T.

Background: Linear Contexual Bandits

- Sequential decision making problem.
- Time horizon: T.
- Action space: K arms.

Background: Linear Contexual Bandits

- Sequential decision making problem.
- Time horizon: T.
- Action space: K arms.
- Each action is associated with a covariate vector (in $\mathbb{R}^d$).

Background: Linear Contexual Bandits

- Sequential decision making problem.
- Time horizon: T.
- Action space: K arms.
- Each action is associated with a covariate vector (in $\mathbb{R}^d$).
- A random reward is generated based on the chosen action.

Background: Linear Contexual Bandits

- Sequential decision making problem.
- Time horizon: T.
- Action space: K arms.
- Each action is associated with a covariate vector (in $\mathbb{R}^d$).
- A random reward is generated based on the chosen action.
- The expectation of the reward is a linear function of the covariate.

Background: Linear Contexual Bandits

- Sequential decision making problem.
- Time horizon: T.
- Action space: K arms.
- Each action is associated with a covariate vector (in $\mathbb{R}^d$).
- A random reward is generated based on the chosen action.
- The expectation of the reward is a linear function of the covariate.
- Target: maximize the cumulative rewards.

Background: Linear Contexual Bandits

- Sequential decision making problem.
- Time horizon: T.
- Action space: K arms.
- Each action is associated with a covariate vector (in $\mathbb{R}^d$).
- A random reward is generated based on the chosen action.
- The expectation of the reward is a linear function of the covariate.
- Target: maximize the cumulative rewards.

Clinical trial

Recommendation system

Bandit feedback: online case

- The reward is immediately observed after an arm is pulled.

Bandit feedback: online case

- The reward is immediately observed after an arm is pulled.

Arm \ Time	1	2	3	4	5	6	7	...	T
1									
2									
3									
4									
5									
$\vdots$									
K									

Bandit feedback: online case

- The reward is immediately observed after an arm is pulled.

Arm \ Time	1	2	3	4	5	6	7	...	T
1									
2									
3	✓								
4									
5									
⋮									
K									

Bandit feedback: online case

- The reward is immediately observed after an arm is pulled.

Time / Arm	1	2	3	4	5	6	7	...	T
1									
2		✓							
3	✓								
4									
5									
⋮									
K									

Bandit feedback: online case

- The reward is immediately observed after an arm is pulled.

Arm \ Time	1	2	3	4	5	6	7	...	T
1									
2		✓							
3	✓								
4									
5									
⋮									
K			✓						

Bandit feedback: online case

- The reward is immediately observed after an arm is pulled.

Arm \ Time	1	2	3	4	5	6	7	...	T
1									
2		✓							
3	✓								
4									
5				✓					
⋮									
K			✓						

Bandit feedback: online case

- The reward is immediately observed after an arm is pulled.

Arm \ Time	1	2	3	4	5	6	7	...	T
1									
2		✓							
3	✓				✓				
4									
5				✓					
⋮									
K			✓						

Bandit feedback: online case

- The reward is immediately observed after an arm is pulled.

Arm \ Time	1	2	3	4	5	6	7	...	T
1						✓			
2		✓							
3	✓				✓				
4									
5				✓					
$\vdots$									
K			✓						

Bandit feedback: online case

- The reward is immediately observed after an arm is pulled.

Arm \ Time	1	2	3	4	5	6	7	...	T
1						✓			
2		✓							
3	✓				✓				
4									
5				✓					
$\vdots$									
K			✓				✓		

Bandit feedback: online case

- The reward is immediately observed after an arm is pulled.

Arm \ Time	1	2	3	4	5	6	7	...	T
1						✓			
2		✓							
3	✓				✓				
4									
5				✓					
$\vdots$								✓	
K			✓				✓		

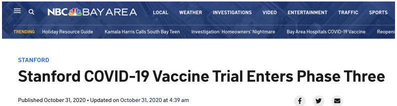
Bandit feedback: online case

- The reward is immediately observed after an arm is pulled.

Arm \ Time	1	2	3	4	5	6	7	...	T
1						$\checkmark$			
2		$\checkmark$							
3	$\checkmark$				$\checkmark$				
4									$\checkmark$
5				$\checkmark$					
$\vdots$								$\checkmark$	
K			$\checkmark$				$\checkmark$		

Limitations of online learning

It can be not feasible/practical to conduct fully online learning.



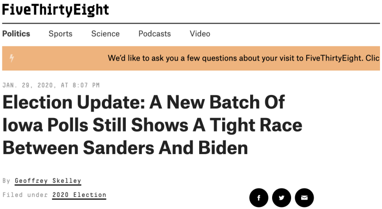
NBC BAY AREA LOCAL WEATHER INVESTIGATIONS VIDEO ENTERTAINMENT TRAFFIC SPORTS

TRENDING Holiday Resource Guide Kamala Harris Calls South Bay Teen Investigation: Homeowners' Nightmare Bay Area Hospitals COVID-19 Vaccine Reopeni

STANFORD

Stanford COVID-19 Vaccine Trial Enters Phase Three

Published October 31, 2020 • Updated on October 31, 2020 at 4:39 am



FiveThirtyEight

Politics Sports Science Podcasts Video

We'd like to ask you a few questions about your visit to FiveThirtyEight. Clic

JAN. 29, 2020, AT 8:07 PM

Election Update: A New Batch Of Iowa Polls Still Shows A Tight Race Between Sanders And Biden

By Geoffrey Skelley

Filed under 2020 Election

Bandit Feedback: Batched Case

- The time horizon is split into M batches;
- The rewards can only be observed simultaneously at the end of each batch.

- The time horizon is split into M batches;
- The rewards can only be observed simultaneously at the end of each batch.

Arm \ Time	1	2	3	4	5	6	7	$\dots$	T
1									
2									
3									
4									
5									
$\vdots$									
K									

Bandit Feedback: Batched Case

- The time horizon is split into M batches;
- The rewards can only be observed simultaneously at the end of each batch.

Arm \ Time	1	2	3	4	5	6	7	...	T
1									
2		✓							
3	✓								
4									
5									
$\vdots$									
K									

Bandit Feedback: Batched Case

- The time horizon is split into M batches;
- The rewards can only be observed simultaneously at the end of each batch.

Arm \ Time	1	2	3	4	5	6	7	...	T
1									
2		✓							
3	✓				✓				
4									
5				✓					
⋮									
K			✓						

Bandit Feedback: Batched Case

- The time horizon is split into M batches;
- The rewards can only be observed simultaneously at the end of each batch.

Arm \ Time	1	2	3	4	5	6	7	...	T
1						✓			
2		✓							
3	✓				✓				
4									✓
5				✓					
⋮								✓	
K			✓				✓		

Our Setting

Sequential decision making problem in

- Linear contexual bandits
- High-dimensional regime with sparse parameters
- Batched observations

Our Setting

Sequential decision making problem in

- Linear contexual bandits
- High-dimensional regime with sparse parameters
- Batched observations

Clinical trial

Recommendation system

Mathematical Formulation: Linear Contexual Bandits

- Time horizon T; number of arms K;

Mathematical Formulation: Linear Contexual Bandits

- Time horizon T; number of arms K;
- Each arm $a \in [K]$ is associated with a d-dimensional feature context $x_{t,a}$;

Mathematical Formulation: Linear Contexual Bandits

- Time horizon T; number of arms K;
- Each arm $a \in [K]$ is associated with a d-dimensional feature context $x_{t,a}$;
- The contexts $\{x_{t,a}\}_{a \in [K]}$ are i.i.d. drawn from a Kd-dimensional joint distribution.

Mathematical Formulation: Linear Contexual Bandits

- Time horizon T; number of arms K;
- Each arm $a \in [K]$ is associated with a d-dimensional feature context $x_{t,a}$;
- The contexts $\{x_{t,a}\}_{a\in[K]}$ are i.i.d. drawn from a Kd-dimensional joint distribution.
- If a decision maker selects action $a \in [K]$, a reward $r_{t,a} \in \mathbb{R}$ is incurred:

$$r_{t,a} = x_{t,a}^{\top}\theta^{\star} + \xi_t.$$

Mathematical Formulation: Linear Contexual Bandits

- Time horizon T; number of arms K;
- Each arm $a \in [K]$ is associated with a d-dimensional feature context $x_{t,a}$;
- The contexts $\{x_{t,a}\}_{a\in[K]}$ are i.i.d. drawn from a Kd-dimensional joint distribution.
- If a decision maker selects action $a \in [K]$, a reward $r_{t,a} \in \mathbb{R}$ is incurred:

$$r_{t,a} = x_{t,a}^\top \theta^\star + \xi_t.$$

- $\theta^\star \in \mathbb{R}^d$ is the underlying unknown parameter vector; $\{\xi_t\}_{t\geq 1}$ is a sequence of i.i.d. zero-mean 1-sub-Gaussian random variables.

Mathematical Formulation: Linear Contexual Bandits

- Time horizon T; number of arms K;
- Each arm $a \in [K]$ is associated with a d-dimensional feature context $x_{t,a}$;
- The contexts $\{x_{t,a}\}_{a\in[K]}$ are i.i.d. drawn from a Kd-dimensional joint distribution.
- If a decision maker selects action $a \in [K]$, a reward $r_{t,a} \in \mathbb{R}$ is incurred:
$$r_{t,a} = x_{t,a}^{\top}\theta^{\star} + \xi_t.$$
- $\theta^{\star} \in \mathbb{R}^d$ is the underlying unknown parameter vector; $\{\xi_t\}_{t\geq 1}$ is a sequence of i.i.d. zero-mean 1-sub-Gaussian random variables.
- Policy $\pi = (\pi_1, \pi_2, \ldots, \pi_T)$. π_t is determined by the observed rewards before the current batch.

Mathematical Formulation: Batch Constraint

- Number of batches M

Mathematical Formulation: Batch Constraint

- Number of batches M
- Batch constraint represented by a grid $t_1 < t_2 < \cdots < t_M = T$

Mathematical Formulation: Batch Constraint

- Number of batches M
- Batch constraint represented by a grid $t_1 < t_2 < \cdots < t_M = T$

Mathematical Formulation: Batch Constraint

- Number of batches M
- Batch constraint represented by a grid $t_1 < t_2 < \cdots < t_M = T$

Types of grids

- Static grid: $\mathcal{T} = \{t_1, \cdots, t_M\}$ fixed in advance

Mathematical Formulation: Batch Constraint

- Number of batches M
- Batch constraint represented by a grid $t_1 < t_2 < \cdots < t_M = T$

Types of grids

- Static grid: $\mathcal{T} = \{t_1, \cdots, t_M\}$ fixed in advance
- Adaptive grid: the next grid point determined by historic data

Mathematical Formulation: Batch Constraint

- Number of batches M
- Batch constraint represented by a grid $t_1 < t_2 < \cdots < t_M = T$

Types of grids

- Static grid: $\mathcal{T} = \{t_1, \cdots, t_M\}$ fixed in advance
- Adaptive grid: the next grid point determined by historic data

Mathematical Formulation: Batch Constraint

- Number of batches M
- Batch constraint represented by a grid $t_1 < t_2 < \cdots < t_M = T$

Types of grids

- Static grid: $\mathcal{T} = \{t_1, \cdots, t_M\}$ fixed in advance
- Adaptive grid: the next grid point determined by historic data

Task

Design policy + grid

Mathematical Formulation: Metric

Regret

$$R_T(\pi, \mathcal{T}) \triangleq \sum_{t=1}^{T} \left(\max_{a \in [K]} x_{t,a}^\top \theta^\star - x_{t,a_t}^\top \theta^\star \right)$$

Mathematical Formulation: Metric

Regret

$$R_T(\pi, \mathcal{T}) \triangleq \sum_{t=1}^{T} \left(\max_{a \in [K]} x_{t,a}^\top \theta^\star - x_{t,a_t}^\top \theta^\star \right)$$

Minimax Regret

$$R_{\mathsf{maxmin}}(K, M, T, s_0) = \inf_{\pi, \mathcal{T}} \sup_{\|\theta^\star\|_2 \leq 1, \|\theta^\star\|_0 \leq s_0} \mathbb{E}\left[R_T(\pi, \mathcal{T})\right]$$

Previous results: batched bandits in low dimensions

Two-arm batched bandits with static grids [PRCS'16]:

$$R_{\mathsf{maxmin}}(2, M, T, 1) = \tilde{\Theta}(T^{\frac{1}{2-2^{1-M}}})$$

Multi-arm batched bandits with adaptive grids [GHRZ'19]

$$R_{\mathsf{maxmin}}(K, M, T, 1) = \tilde{\Theta}(\sqrt{K}T^{\frac{1}{2-2^{1-M}}})$$

Batched contexual bandits in low dimensions [HZZBGY'20]

$$R_{\mathsf{maxmin}}(M, T, d) = \tilde{\Theta}\left(\sqrt{dT}\left(T/d^2\right)^{\frac{1}{2(2^M-1)}}\right)$$

Previous results: Online bandits in high dimensions

Online contexual bandits in high dimensions (with margin conditions)

[BB'20]

$$R_{\mathsf{maxmin}}(T, T, s_0) = O\left({s_0}^2(\log d + \log T)^2\right)$$

[WWY'18]

$$R_{\mathsf{maxmin}}(T, T, s_0) = O\left({s_0}^2(\log d + s_0)\log T\right)$$

Our Contributions

- Study the batched contexual bandits in the high-dimensional setting
- Allow the grids to be designed adaptively

Our Contributions

- Study the batched contexual bandits in the high-dimensional setting
- Allow the grids to be designed adaptively

Theorem (R. and Zhou '20, informally)

Under some assumptions (to be specified later), when $M = O(\log\log(T/s_0))$

$$R_{\mathsf{maxmin}}(M, T, s_0) = \tilde{\Theta}\left(\sqrt{Ts_0}\,(T/s_0)^{\frac{1}{2(2^M-1)}}\right);$$

When $M = \Omega(\log\log(T/s_0))$,

$$R_{\mathsf{maxmin}}(T, T, s_0) = \tilde{\Theta}(\sqrt{Ts_0}).$$

Assumption 1

Assumption (Sub-Gaussianity)

The marginal distribution of $x_{t,a}$ is 1-sub-Gaussian, $\forall a \in [k]$.

Assumption 2

Assumption (Restricted Bounded Density)

There $\exists$ *a constant* $\gamma > 0$, *s.t., for each* $a \in [K]$, *any subset* $S \subset [d]$ *with* $|S| = s_0$, *and any unit vector* $v \in R^{s_0}$, *the probability density function of* $v\top x_{t,a}(S)$ *exists and is bounded above by* $\gamma/2$.

- A wide range of distributions satisfies this assumption, e.g., (non-degenerate) Gaussians, uniform distribution.

Assumption 3 amd 4

Assumption (Sparsity in high-dimension)

The linear contextual bandits have:

- *high-dimensional contexts:* $d = \mathrm{Poly}(T)$;
- *sparse parameters:* $\|\theta^\star\|_0 \leq s_0 = O(T^{1-\varepsilon})$, *for some* $\varepsilon > 0$.

Assumption (Not too many arms)

The number of actions K *satisfies* $K^2 \log K = O(d/s_0)$.

Lower Bound

Theorem (R. and Zhou '20)

Consider the two-action setting where $x_{t,1} \overset{iid}{\sim} \mathcal{N}(0, I_d)$, $x_{t,2} \overset{iid}{\sim} \mathcal{N}(0, I_d)$ and $x_{t,1}$ is independent of $x_{t,2}$. Then for any $M \leq T$, any policy π and adaptive grid $\mathcal{T}$, we have:

$$\sup_{\substack{\theta^\star : \|\theta^\star\|_0 \leq s_0, \\ \|\theta^\star\|_2 \leq 1}} \mathbb{E}_{\theta^\star}\left[R_T(\pi, \mathcal{T})\right] \geq c \cdot \max\left(M^{-2}\sqrt{Ts_0}\left(\frac{T}{s_0}\right)^{\frac{1}{2(2^M-1)}}, \sqrt{Ts_0}\right)$$

where $c > 0$ is a numerical constant independent of (T, M, d, s_0).

Lower Bound

Theorem (R. and Zhou '20)

Consider the two-action setting where $x_{t,1} \overset{iid}{\sim} \mathcal{N}(0, I_d)$, $x_{t,2} \overset{iid}{\sim} \mathcal{N}(0, I_d)$ *and* $x_{t,1}$ *is independent of* $x_{t,2}$. *Then for any* $M \le T$, *any policy* π *and adaptive grid* $\mathcal{T}$, *we have:*

$$\sup_{\substack{\theta^\star : \|\theta^\star\|_0 \le s_0, \\ \|\theta^\star\|_2 \le 1}} \mathbb{E}_{\theta^\star}\left[R_T(\pi, \mathcal{T})\right] \ge c \cdot \max\left(M^{-2}\sqrt{Ts_0}\left(\frac{T}{s_0}\right)^{\frac{1}{2(2^M-1)}}, \sqrt{Ts_0} \right)$$

where $c > 0$ *is a numerical constant independent of* (T, M, d, s_0).

- When $M = O(\log\log T)$, the term $M^{-2}\sqrt{Ts_0}\left(\frac{T}{s_0}\right)^{\frac{1}{2(2^M-1)}}$ dominates;

Lower Bound

Theorem (R. and Zhou '20)

Consider the two-action setting where $x_{t,1} \overset{iid}{\sim} \mathcal{N}(0, I_d)$, $x_{t,2} \overset{iid}{\sim} \mathcal{N}(0, I_d)$ and $x_{t,1}$ is independent of $x_{t,2}$. Then for any $M \leq T$, any policy π and adaptive grid $\mathcal{T}$, we have:

$$\sup_{\substack{\theta^\star : \|\theta^\star\|_0 \leq s_0, \\ \|\theta^\star\|_2 \leq 1}} \mathbb{E}_{\theta^\star}\left[R_T(\pi, \mathcal{T})\right] \geq c \cdot \max\left(M^{-2}\sqrt{Ts_0}\left(\frac{T}{s_0}\right)^{\frac{1}{2(2^M-1)}}, \sqrt{Ts_0}\right)$$

where $c > 0$ is a numerical constant independent of (T, M, d, s_0).

- When $M = O(\log\log T)$, the term $M^{-2}\sqrt{Ts_0}\left(\frac{T}{s_0}\right)^{\frac{1}{2(2^M-1)}}$ dominates;
- When $M = \Omega(\log\log T)$, the term $\sqrt{Ts_0}$ dominates.

Proof Idea: Fixed Hypothesis Testing

Construct several reward distributions such that:

- Large separation: if a policy performs well under one distribution, it will perform badly under others
- Indistinguishability: these reward distributions are information theoretically hard to distinguish given observed rewards

Proof Sketch

- Construct a sequence of prior distribution of $\theta^{\star}$: $\{Q_m\}_{m\in[m]}$

Proof Sketch

- Construct a sequence of prior distribution of $\theta^\star$: $\{Q_m\}_{m\in[m]}$
- Define the fixed grids: $T_m = \left\lfloor s_0 (T/s_0)^{\frac{1-2^{-m}}{1-2^{-M}}} \right\rfloor$, for $m \in [M]$

Proof Sketch

- Construct a sequence of prior distribution of $\theta^\star$: $\{Q_m\}_{m\in[m]}$
- Define the fixed grids: $T_m = \left\lfloor s_0(T/s_0)^{\frac{1-2^{-m}}{1-2^{-M}}} \right\rfloor$, for $m \in [M]$
- Given a policy π and a grid design $\mathcal{T} = \{t_1, \ldots, t_m, \ldots, t_M\}$, we now define for each $m \in [M]$ the "bad" event $A_m = \{t_{m-1} \leq T_{m-1} < T_m \leq t_m\}$ (why?)

Proof Sketch

- Construct a sequence of prior distribution of $\theta^\star$: $\{Q_m\}_{m\in[m]}$
- Define the fixed grids: $T_m = \left\lfloor s_0(T/s_0)^{\frac{1-2^{-m}}{1-2^{-M}}} \right\rfloor$, for $m \in [M]$
- Given a policy π and a grid design $\mathcal{T} = \{t_1, \ldots, t_m, \ldots, t_M\}$, we now define for each $m \in [M]$ the "bad" event $A_m = \{t_{m-1} \le T_{m-1} < T_m \le t_m\}$ (why?)
- Show that at least one A_m occurs with a large enough probability under the corresponding prior Q_m

Upper Bound

Theorem (R. and Zhou '20)

Under the assumptions and when $M = O(\log\log(T/s_0))$, we have

$$\sup_{\substack{\theta^\star:\|\theta^\star\|_2\le 1,\\ \|\theta^\star\|_0\le s_0}} \mathbb{E}_{\theta^\star}[R_T(\mathsf{Alg})] = \tilde{O}\left(\sqrt{Ts_0}(T/s_0)^{\frac{1}{2(2^M-1)}}\right)$$

- $M = \log\log T$ batches sufficient for achieving the online minimax regret $\tilde{O}(\sqrt{Ts_0})$ (up to logarithmic terms)
- The upper bound matches the lower bound (up to logarithmic terms)

Optimal Grid Design

- It suffices to use a static grid to achive the optimal regret under adaptive grids.

$\mathcal{T} = \{t_1, \ldots, t_M\}$ with

$$t_1 = a, t_m = \left\lfloor a\sqrt{t_{m-1}} \right\rfloor,$$

where a is chosen such that $t_M = T$.

Algorithm

Lasso Batched Greedy Learning

Input Time horizon T; context dimension d; number of batches M; sparsity bound s_0.

Initialize $b = \Theta\left(\sqrt{T} \cdot \left(\frac{T}{s_0}\right)^{\frac{1}{2(2^M-1)}}\right)$; $\hat{\theta}_0 = \mathbf{0} \in \mathbb{R}^d$;

Static grid $\mathcal{T} = \{t_1, \ldots, t_M\}$, with $t_1 = b\sqrt{s_0}$ and $t_m = b\sqrt{t_{m-1}}$ for $t \in \{2, \ldots, M\}$;

Partition each batch into M intervals evenly, i.e., $(t_{m-1}, t_m] = \cup_{j=1}^M T_m^{(j)}$, for $m \in [M]$.

Algorithm

Lasso Batched Greedy Learning

---

for $m = 1$ to M **do**

  for $t = t_{m-1} + 1$ to t_m **do**

    (a) Choose $a_t = \underset{a \in [K]}{\operatorname{argmax}}\ x_{t,a}^\top \hat{\theta}_{m-1}$ (break ties with lower action index).

    (b) Incur reward r_{t,a_t}.

  end for

  $T^{(m)} \leftarrow \cup_{m'=1}^{m} T_{m'}^{(m)}$; $\lambda_m \leftarrow 5\sqrt{\dfrac{2\log K(\log d + 2\log T)}{|T^{(m)}|}}$;

  Update $\hat{\theta}_m \leftarrow \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}}\ \dfrac{1}{2|T^{(m)}|} \sum_{t \in T^{(m)}} (r_{t,a_t} - x_{t,a_t}^\top \theta)^2 + \lambda_m \|\theta\|_1$.

end for

Conclusion

- Study the batched learning problem in high-dimensional linear contexual bandit setting
- Develop a lower bound that characterizes the fundamental learning limits
- Provide a algorithm that yields a matching upper bound

Future work

- Beyond linearity
- Develop an algorithm that does not require the knowledge of the sparsity parameter s_0
- Tighten the bound (remove the factor of M^{-2})

Dynamic Batch Learning in High-Dimensional Sparse linear Contextual Bandits

(https://arxiv.org/abs/2008.11918)