Asymptotically Optimal Exploration in Contextual Linear Bandits

Andrea Tirinzoni[1], Matteo Pirotta[2], Marcello Restelli[1], Alessandro Lazaric[2]

[1] Politecnico di Milano

[2] Facebook AI Research

To appear at NeurIPS 2020

POLITECNICO
MILANO 1863

FACEBOOK AI

Contextual Linear Bandits

At each time t, the learner

- Observes *context* $X_t \in \mathcal{X}$, $X_t \sim \rho$
- Plays *arm* $A_t \in \mathcal{A}$
- Receives *reward* $Y_t = \underbrace{\phi(X_t, A_t)^T \theta^\star}_{\mu_{\theta^\star}(X_t, A_t)} + \mathcal{N}(0, \sigma^2)$ with $\theta^\star \in \mathbb{R}^d$ *unknown*

Contextual Linear Bandits

At each time t, the learner

- Observes *context* $X_t \in \mathcal{X}$, $X_t \sim \rho$
- Plays *arm* $A_t \in \mathcal{A}$
- Receives *reward* $Y_t = \underbrace{\phi(X_t, A_t)^T \theta^\star}_{\mu_{\theta^\star}(X_t, A_t)} + \mathcal{N}(0, \sigma^2)$ with $\theta^\star \in \mathbb{R}^d$ *unknown*

Goal: minimize cumulative regret

$$\mathbb{E}[R_n(\theta^\star)] := \mathbb{E}\left[\sum_{t=1}^{n} \left(\max_{a \in \mathcal{A}} \mu_{\theta^\star}(X_t, a) - \mu_{\theta^\star}(X_t, A_t)\right)\right]$$

Contextual Linear Bandits

At each time t, the learner

- Observes *context* $X_t \in \mathcal{X}$, $X_t \sim \rho$
- Plays *arm* $A_t \in \mathcal{A}$
- Receives *reward* $Y_t = \underbrace{\phi(X_t, A_t)^T \theta^\star}_{\mu_{\theta^\star}(X_t, A_t)} + \mathcal{N}(0, \sigma^2)$ with $\theta^\star \in \mathbb{R}^d$ *unknown*

Goal: minimize cumulative regret

$$\mathbb{E}\left[R_n(\theta^\star)\right] := \mathbb{E}\left[\sum_{t=1}^{n} \left(\max_{a \in \mathcal{A}} \mu_{\theta^\star}(X_t, a) - \mu_{\theta^\star}(X_t, A_t)\right)\right]$$

Assumptions: $\mathcal{X}, \mathcal{A}$ finite, $\rho(x) > 0$ for all $x \in \mathcal{X}$, $\theta^\star \in \Theta := \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq B\}$, unique optimal arm $a^\star_{\theta^\star}(x)$ for all $x \in \mathcal{X}$

State of the Art

- Algorithms based on **optimism** [Abbasi-Yadkori et al., 2011] or **Thompson sampling** [Agrawal and Goyal, 2013] are not asymptotically optimal [Lattimore and Szepesvári, 2017]

State of the Art

- Algorithms based on **optimism** [Abbasi-Yadkori et al., 2011] or **Thompson sampling** [Agrawal and Goyal, 2013] are not asymptotically optimal [Lattimore and Szepesvári, 2017]
- Algorithms derived from **asymptotic problem-dependent lower bounds**

	OSSB [Combes et al., 2017]	CROP [Jun and Zhang, 2020]	SPL [Degenne et al., 2020]	OAM [Hao et al., 2020]	SOLID [ours]

State of the Art

- Algorithms based on **optimism** [Abbasi-Yadkori et al., 2011] or **Thompson sampling** [Agrawal and Goyal, 2013] are not asymptotically optimal [Lattimore and Szepesvári, 2017]
- Algorithms derived from **asymptotic problem-dependent lower bounds**

	OSSB [Combes et al., 2017]	CROP [Jun and Zhang, 2020]	SPL [Degenne et al., 2020]	OAM [Hao et al., 2020]	SOLID [ours]
Linear contextual	✗	✗	✗	✓	✓

State of the Art

- Algorithms based on **optimism** [Abbasi-Yadkori et al., 2011] or **Thompson sampling** [Agrawal and Goyal, 2013] are not asymptotically optimal [Lattimore and Szepesvári, 2017]
- Algorithms derived from **asymptotic problem-dependent lower bounds**

	OSSB [Combes et al., 2017]	CROP [Jun and Zhang, 2020]	SPL [Degenne et al., 2020]	OAM [Hao et al., 2020]	SOLID [ours]
Linear contextual	✗	✗	✗	✓	✓
Asympt. optimal	✗	✗	✓	✓	✓

State of the Art

- Algorithms based on **optimism** [Abbasi-Yadkori et al., 2011] or **Thompson sampling** [Agrawal and Goyal, 2013] are not asymptotically optimal [Lattimore and Szepesvári, 2017]
- Algorithms derived from **asymptotic problem-dependent lower bounds**

	OSSB [Combes et al., 2017]	CROP [Jun and Zhang, 2020]	SPL [Degenne et al., 2020]	OAM [Hao et al., 2020]	SOLID [ours]
Linear contextual	✗	✗	✗	✓	✓
Asympt. optimal	✗	✗	✓	✓	✓
No forced explore	✗	✓	✓	✗	✓

State of the Art

- Algorithms based on **optimism** [Abbasi-Yadkori et al., 2011] or **Thompson sampling** [Agrawal and Goyal, 2013] are not asymptotically optimal [Lattimore and Szepesvári, 2017]
- Algorithms derived from **asymptotic problem-dependent lower bounds**

	OSSB [Combes et al., 2017]	CROP [Jun and Zhang, 2020]	SPL [Degenne et al., 2020]	OAM [Hao et al., 2020]	SOLID [ours]
Linear contextual	✗	✗	✗	✓	✓
Asympt. optimal	✗	✗	✓	✓	✓
No forced explore	✗	✓	✓	✗	✓
Efficient/scalable	✗	✗	✓	✗	✓

State of the Art

- Algorithms based on **optimism** [Abbasi-Yadkori et al., 2011] or **Thompson sampling** [Agrawal and Goyal, 2013] are not asymptotically optimal [Lattimore and Szepesvári, 2017]
- Algorithms derived from **asymptotic problem-dependent lower bounds**

	OSSB [Combes et al., 2017]	CROP [Jun and Zhang, 2020]	SPL [Degenne et al., 2020]	OAM [Hao et al., 2020]	SOLID [ours]
Linear contextual	✗	✗	✗	✓	✓
Asympt. optimal	✗	✗	✓	✓	✓
No forced explore	✗	✓	✓	✗	✓
Efficient/scalable	✗	✗	✓	✗	✓
Dep. $\log(\lvert\mathcal{A}\rvert)$	✗	✗	✗	✗	✓

State of the Art

- Algorithms based on **optimism** [Abbasi-Yadkori et al., 2011] or **Thompson sampling** [Agrawal and Goyal, 2013] are not asymptotically optimal [Lattimore and Szepesvári, 2017]
- Algorithms derived from **asymptotic problem-dependent lower bounds**

	OSSB [Combes et al., 2017]	CROP [Jun and Zhang, 2020]	SPL [Degenne et al., 2020]	OAM [Hao et al., 2020]	SOLID [ours]
Linear contextual	✗	✗	✗	✓	✓
Asympt. optimal	✗	✗	✓	✓	✓
No forced explore	✗	✓	✓	✗	✓
Efficient/scalable	✗	✗	✓	✗	✓
Dep. $\log(\lvert\mathcal{A}\rvert)$	✗	✗	✗	✗	✓
Minimax optimal	✗	✗	✗	✗	✓*

* Only for linear non-contextual problems

Lower Bound

Any **uniformly consistent** bandit strategy satisfies

$$\liminf_{n\to\infty} \frac{\mathbb{E}\left[R_n(\theta^\star)\right]}{\log(n)} \geq v^\star(\theta^\star)$$

where $v^\star(\theta^\star)$ is the value of the **optimization problem**

$$\begin{aligned} &\inf_{\eta(x,a)\geq 0} \quad \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}} \eta(x,a)\ \underbrace{\Delta_{\theta^\star}(x,a)}_{\text{Sub-optimality gap}} \\ &\text{s.t.} \quad \inf_{\theta'\in\Theta_{\text{alt}}} \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}} \eta(x,a)\ \underbrace{d_{x,a}(\theta^\star,\theta')}_{\text{KL divergence}} \geq 1 \end{aligned} \qquad \text{(P)}$$

$$\boxed{\Theta_{\text{alt}} := \{\theta' \in \Theta \mid \exists x \in \mathcal{X},\ a^\star_{\theta^\star}(x) \neq a^\star_{\theta'}(x)\}}$$

Lower Bound Reformulation

(1) Constrain number of pulls for each context

$$
\begin{aligned}
\inf_{\eta(x,a)\geq 0} \quad & \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}} \eta(x,a)\Delta_{\theta^\star}(x,a) \\
\text{s.t.} \quad & \inf_{\theta'\in\Theta_{\mathrm{alt}}} \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}} \eta(x,a) d_{x,a}(\theta^\star,\theta') \geq 1 \\
& \sum_{a} \eta(x,a) = z\,\rho(x) \quad \forall x \in \mathcal{X}
\end{aligned}
$$

Sample budget z:
$$\sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}} \eta(x,a) = z$$

Lower Bound Reformulation

(2) Change of variables $\eta(x,a) \to z\rho(x)\omega(x,a)$

$$\min_{\omega(x,a)\geq 0} \quad z \cdot \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \omega(x,a)\Delta_{\theta^\star}(x,a)$$

Expectation under ρ

$$\text{s.t.} \quad \inf_{\theta'\in\Theta_{\text{alt}}} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \omega(x,a)\, d_{x,a}(\theta^\star,\theta') \geq 1/z \qquad (\mathsf{P}_z)$$

Probability simplex

$$\omega(x,\cdot) \in \Omega \quad \forall x \in \mathcal{X}$$

Conditional arm probabilities

Lower Bound Reformulation

(2) Change of variables $\eta(x,a) \to z\rho(x)\omega(x,a)$

$$\min_{\omega(x,a)\geq 0} \quad z \cdot \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \omega(x,a)\Delta_{\theta^\star}(x,a)$$

Expectation under ρ

$$\text{s.t.} \quad \inf_{\theta'\in\Theta_{\text{alt}}} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \omega(x,a)\, d_{x,a}(\theta^\star,\theta') \geq 1/z \qquad (\mathsf{P}_z)$$

Probability simplex

$$\omega(x,\cdot) \in \Omega \quad \forall x \in \mathcal{X}$$

Conditional arm probabilities

Lemma

- *(P_z) is* **feasible** *for* $1/z \leq \max_{\omega\in\Omega} \inf_{\theta'\in\Theta_{\text{alt}}} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \omega(x,a) d_{x,a}(\theta^\star,\theta')$
- *Let* $u_z^\star(\theta^\star)$ *be the optimal solution of (P_z), then* $u_z^\star(\theta^\star) \leq v^\star(\theta^\star) + \mathcal{O}(1/\sqrt{z})$

Lagrangian Relaxation

$$\boxed{\sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \omega(x,a)\mu_{\theta^\star}(x,a)}$$

$$\min_{\lambda\geq 0}\max_{\omega\in\Omega}\left\{ f(\omega;\theta^\star) + \lambda\, g(\omega; z, \theta^\star) \right\} \qquad (\mathsf{P}_\lambda)$$

$$\boxed{\inf_{\theta'\in\Theta_{\mathrm{alt}}} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \omega(x,a) d_{x,a}(\theta^\star,\theta') - \frac{1}{z}}$$

SOLID (aSymptotic Optimal Linear prImal Dual)

Initialize ω_1, λ_1, $\widehat{\theta}_0$, $\widehat{\rho}_0$, z_1
for $t = 1, 2, \dots, n$ **do**
  Receive context $X_t \sim \rho$
  if $\inf_{\theta' \in \overline{\Theta}_{t-1}} \|\widehat{\theta}_{t-1} - \theta'\|^2_{\overline{V}_{t-1}} > \beta_{t-1}$ **then**
    Exploitation: $A_t \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} \mu_{\widehat{\theta}_{t-1}}(X_t, a)$
  else
    Exploration: sample arm: $A_t \sim \omega_t(X_t, \cdot)$
    Optimization: Update $(\lambda_{t+1}, \omega_{t+1})$ by optimistic primal-dual sub-gradient
    Phases: Update z_{t+1} (increase after sufficient exploration steps)
  end if
  Pull A_t and observe outcome Y_t
  Estimation: update $\widehat{\theta}_t$, $\widehat{\rho}_t$
end for

Theoretical Results

Theorem (Problem-dependent regret bound)

For any finite n, the expected regret of SOLID is bounded as

$$\mathbb{E}\big[R_n(\theta^\star)\big] \leq v^\star(\theta^\star)\log n + \mathcal{O}\big((\log n)^{3/4}\big) + \mathcal{O}(1)$$

Theoretical Results

Theorem (Problem-dependent regret bound)

For any finite n, the expected regret of SOLID is bounded as

$$\mathbb{E}\big[R_n(\theta^\star)\big] \leq v^\star(\theta^\star)\log n + \mathcal{O}\big((\log n)^{3/4}\big) + \mathcal{O}(1)$$

- SOLID is **asymptotically optimal**

Theoretical Results

Theorem (Problem-dependent regret bound)

For any finite n, the expected regret of SOLID is bounded as

$$\mathbb{E}\big[R_n(\theta^\star)\big] \leq v^\star(\theta^\star)\log n + \mathcal{O}((\log n)^{3/4}) + \mathcal{O}(1)$$

- SOLID is **asymptotically optimal**
- $\mathcal{O}((\log n)^{3/4})$ regret due to *incremental* and *phased* nature (can be improved...)

Theoretical Results

Theorem (Problem-dependent regret bound)

For any finite n, the expected regret of SOLID is bounded as

$$\mathbb{E}\big[R_n(\theta^\star)\big] \leq v^\star(\theta^\star)\log n + \mathcal{O}\big((\log n)^{3/4}\big) + \mathcal{O}(1)$$

- SOLID is **asymptotically optimal**
- $\mathcal{O}\big((\log n)^{3/4}\big)$ regret due to *incremental* and *phased* nature (can be improved...)
- $\mathcal{O}(1)$ regret mostly due to the optimization problem initially being *infeasible*

Theoretical Results

Theorem (Problem-dependent regret bound)

For any finite n, the expected regret of SOLID is bounded as

$$\mathbb{E}\left[R_n(\theta^\star)\right] \leq v^\star(\theta^\star)\log n + \mathcal{O}\left((\log n)^{3/4}\right) + \mathcal{O}(1)$$

- SOLID is **asymptotically optimal**
- $\mathcal{O}\left((\log n)^{3/4}\right)$ regret due to *incremental* and *phased* nature (can be improved...)
- $\mathcal{O}(1)$ regret mostly due to the optimization problem initially being *infeasible*
- Regret bound scales only with $\log|\mathcal{A}|$ and does not depend on $1/\min_x \rho(x)$

Theoretical Results

Theorem (Worst-case regret bound)

For any finite n, the expected regret of SOLID is bounded as

$$\mathbb{E}\big[R_n(\theta^\star)\big] \leq \widetilde{\mathcal{O}}(|\mathcal{X}|\sqrt{nd})$$

Theoretical Results

Theorem (Worst-case regret bound)

For any finite n, the expected regret of SOLID is bounded as

$$\mathbb{E}\left[R_n(\theta^\star)\right] \leq \widetilde{\mathcal{O}}(|\mathcal{X}|\sqrt{nd})$$

- SOLID is **minimax optimal** in non-contextual linear bandits ($|\mathcal{X}| = 1$)

Theoretical Results

Theorem (Worst-case regret bound)

For any finite n, the expected regret of SOLID is bounded as

$$\mathbb{E}\big[R_n(\theta^\star)\big] \leq \widetilde{\mathcal{O}}(|\mathcal{X}|\sqrt{nd})$$

- SOLID is **minimax optimal** in non-contextual linear bandits ($|\mathcal{X}| = 1$)
- Open question whether the dependence on $|\mathcal{X}|$ could be reduced

SOLID - Estimation

At each step t, SOLID estimates

- $\theta^\star$ via RLS[1]: $\widehat{\theta}_t = \overline{V}_t^{-1} \sum_{s=1}^{t} \phi_s Y_s$, where $\overline{V}_t := \sum_{s=1}^{t} \phi_s \phi_s^T + \nu I$
- the context distribution: $\widehat{\rho}_t(x) = \frac{1}{t} \sum_{s=1}^{t} \mathbb{1}\left\{X_s = x\right\}$

[1] To carry out the analysis, we also need to project $\widehat{\theta}_t$ onto Θ

SOLID - Exploitation

At each step t, SOLID

- builds a confidence ellipsoid $\mathcal{C}_t := \{\theta \in \mathbb{R}^d : \|\theta - \widehat{\theta}_t\|^2_{V_t} \leq \beta_t\}$

SOLID - Exploitation

At each step t, SOLID

- builds a confidence ellipsoid $\mathcal{C}_t := \{\theta \in \mathbb{R}^d : \|\theta - \widehat{\theta}_t\|^2_{\overline{V}_t} \leq \beta_t\}$
- tests whether all alternative reward parameters (w.r.t. $\widehat{\theta}_t$) are outside $\mathcal{C}_t$

SOLID - Exploitation

At each step t, SOLID

- builds a confidence ellipsoid $\mathcal{C}_t := \{\theta \in \mathbb{R}^d : \|\theta - \widehat{\theta}_t\|^2_{V_t} \leq \beta_t\}$
- tests whether all alternative reward parameters (w.r.t. $\widehat{\theta}_t$) are outside $\mathcal{C}_t$
  - If true $\to$ the **empirical optimal arm** for the current context is pulled

SOLID - Exploitation

At each step t, SOLID

- builds a confidence ellipsoid $\mathcal{C}_t := \{\theta \in \mathbb{R}^d : \|\theta - \widehat{\theta}_t\|^2_{V_t} \leq \beta_t\}$
- tests whether all alternative reward parameters (w.r.t. $\widehat{\theta}_t$) are outside $\mathcal{C}_t$
  - If true $\rightarrow$ the **empirical optimal arm** for the current context is pulled

Theorem

For $c_{n,\delta}$ of order $\mathcal{O}(\log(1/\delta) + d \log\log n)$,[2]

$$\mathbb{P}\left\{\exists t \in [n] : \|\widehat{\theta}_t - \theta^\star\|^2_{V_t} \geq c_{n,\delta}\right\} \leq \delta,$$

[2] This improves the concentration bound of [Abbasi-Yadkori et al., 2011] which scales as $d\log(1/\delta)$

SOLID - Optimism

Confidence interval at x, a

SOLID builds an (almost) **optimistic Lagrangian**

$$f_t(\omega) := \sum_{x\in\mathcal{X}} \widehat{\rho}_{t-1}(x) \sum_{a\in\mathcal{A}} \omega(x,a) \left(\mu_{\widehat{\theta}_{t-1}}(x,a) + \sqrt{\gamma_t}\, \|\phi(x,a)\|_{\overline{V}_{t-1}^{-1}} \right)$$

$$g_t(\omega, z) := \inf_{\theta' \in \overline{\Theta}_{t-1}} \sum_{x\in\mathcal{X}} \widehat{\rho}_{t-1}(x) \sum_{a\in\mathcal{A}} \omega(x,a) \left(d_{x,a}(\widehat{\theta}_{t-1}, \theta') + c\sqrt{\gamma_t}\|\phi(x,a)\|_{\overline{V}_{t-1}^{-1}} \right) - \frac{1}{z}$$

Alternative parameters w.r.t. $\widehat{\theta}_{t-1}$

SOLID - Optimization

SOLID uses a **primal-dual sub-gradient method** [Beck and Teboulle, 2003]

SOLID - Optimization

SOLID uses a **primal-dual sub-gradient method** [Beck and Teboulle, 2003]

- Update rule for ω: **online mirror ascent** on the simplex

$$\omega_{t+1}(x,a) \leftarrow \frac{\omega_t(x,a)e^{\alpha_t \, q_t(x,a)}}{\sum_{a' \in \mathcal{A}} \omega_t(x,a') e^{\alpha_t q_t(x,a')}}$$

Sub-gradient
$q_t \in \partial\big(f_t(\omega_t) + \lambda_t g_t(\omega_t, z_t)\big)$

- Update rule for λ: **Projected sub-gradient descent**

$$\lambda_{t+1} \leftarrow \mathrm{clip}\Big(\lambda_t - \alpha_t g_t(\omega_t, z_t); 0, \lambda_{\max}\Big)$$

SOLID - Action Sampling

- SOLID does not use any explicit *tracking* procedure...

  [Garivier and Kaufmann, 2016, Combes et al., 2017, Degenne et al., 2019]

SOLID - Action Sampling

- SOLID does not use any explicit *tracking* procedure...
  [Garivier and Kaufmann, 2016, Combes et al., 2017, Degenne et al., 2019]
- ...but it directly **samples** from ω_t

SOLID - Action Sampling

- SOLID does not use any explicit *tracking* procedure...
  [Garivier and Kaufmann, 2016, Combes et al., 2017, Degenne et al., 2019]
- ...but it directly **samples** from ω_t
- Analysis of action sampling crucial for *removing* polynomial dependence on $|\mathcal{A}|$

SOLID - Action Sampling

- SOLID does not use any explicit *tracking* procedure...
  [Garivier and Kaufmann, 2016, Combes et al., 2017, Degenne et al., 2019]
- ...but it directly **samples** from ω_t
- Analysis of action sampling crucial for *removing* polynomial dependence on $|\mathcal{A}|$
- **Intuition**: we only need to concentrate *expectations* of the form

$$\left| \sum_{s \le t : E_s} \left(\varphi(X_s, A_s) - \sum_{x \in \mathcal{X}} \rho(x) \sum_{a \in \mathcal{A}} \omega_s(x, a) \varphi(x, a) \right) \right| \le ?$$

SOLID - Dealing with z

- SOLID uses an increasing schedule with **phases**: $\{z_k\}_{k\geq 1}$

SOLID - Dealing with z

- SOLID uses an increasing schedule with **phases**: $\{z_k\}_{k\geq 1}$
- Change phase when the number of exploration rounds exceeds given thresholds

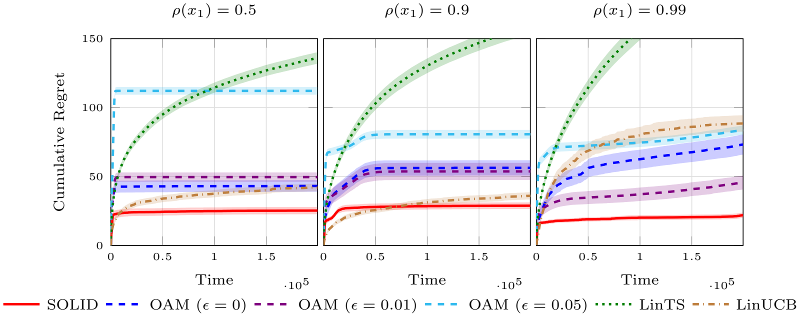
SOLID - Dealing with z

- SOLID uses an increasing schedule with **phases**: $\{z_k\}_{k\geq 1}$
- Change phase when the number of exploration rounds exceeds given thresholds
- Theoretical results for **exponential** schedule $z_k = z_0 e^k$

Numerical Results

Toy problem with $|\mathcal{X}| = 2$, $|\mathcal{A}| = 3$, $d = 3$

- Asymptotic lower bound: explore in x_2, go greedy in x_1



Open Questions

- Minimax optimality in contextual problems?

Open Questions

- Minimax optimality in contextual problems?
- How to deal with continuous contexts?

Open Questions

- Minimax optimality in contextual problems?
- How to deal with continuous contexts?
- Finite-time problem-dependent optimality?

Open Questions

- Minimax optimality in contextual problems?
- How to deal with continuous contexts?
- Finite-time problem-dependent optimality?
- How to handle misspecified linear models?

The End

Details are in the paper:

An Asymptotically Optimal Primal-Dual Incremental Algorithm for Contextual Linear Bandits
NeurIPS 2020
Andrea Tirinzoni, Matteo Pirotta, Marcello Restelli, Alessandro Lazaric

Thank you!

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011).
Improved algorithms for linear stochastic bandits.
In *Advances in Neural Information Processing Systems*, pages 2312–2320.

Agrawal, S. and Goyal, N. (2013).
Thompson sampling for contextual bandits with linear payoffs.
In *International Conference on Machine Learning*, pages 127–135.

Beck, A. and Teboulle, M. (2003).
Mirror descent and nonlinear projected subgradient methods for convex optimization.
Operations Research Letters, 31(3):167–175.

Combes, R., Magureanu, S., and Proutière, A. (2017).
Minimal exploration in structured stochastic bandits.
In *NIPS*, pages 1763–1771.

Degenne, R., Koolen, W. M., and Ménard, P. (2019).
Non-asymptotic pure exploration by solving games.
In *NeurIPS*, pages 14465–14474.

References (cont.)

Degenne, R., Shao, H., and Koolen, W. (2020).
Structure adaptive algorithms for stochastic bandits.
In *International Conference on Machine Learning*, Vienna, Austria.
Virtual conference.

Garivier, A. and Kaufmann, E. (2016).
Optimal best arm identification with fixed confidence.
In *Conference on Learning Theory*, pages 998–1027.

Hao, B., Lattimore, T., and Szepesvari, C. (2020).
Adaptive exploration in linear contextual bandit.
volume 108 of *Proceedings of Machine Learning Research*, pages 3536–3545, Online. PMLR.

Jun, K.-S. and Zhang, C. (2020).
Crush optimism with pessimism: Structured bandits beyond asymptotic optimality.
arXiv preprint arXiv:2006.08754.

Lattimore, T. and Szepesvári, C. (2017).
The end of optimism? an asymptotic analysis of finite-armed linear bandits.
In *AISTATS*, volume 54 of *Proceedings of Machine Learning Research*, pages 728–737. PMLR.