Near-Optimal Regret in Linear MDPs with Aggregate Bandit Feedback

Asaf Cassel[* 1], Haipeng Luo[2], Aviv Rosenberg[* 4], Dmitry Sotnikov[3]

[1]Tel Aviv University, Israel [2]University of Southern California [3]Amazon Science [4]Google Research

The Blavatnik School of Computer Science
The Raymond and Beverly Sackler Faculty of Exact Sciences
Tel Aviv University

TLDR: We study regret minimization in linear MDPs under aggregate reward feedback, i.e., when only the sum of rewards is observed. We present the first Policy Optimization and Value Iteration based algorithms that are computationally efficient and achieve rate-optimal $\sqrt{K}$ regret. Our algorithms employ a novel ensemble randomization technique that may be of separate interest.

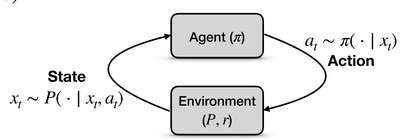
Motivation

Getting dense reward feedback can be expensive/impractical

Problem Setup

Finite horizon **MDP** $\mathcal{M} = (\mathcal{X}, \mathcal{A}, x_1, r, P, H)$

- $\mathcal{X}$ - states
- $\mathcal{A}$ - actions
- $x_1 \in \mathcal{X}$ initial state
- $r_h : \mathcal{X} \times \mathcal{A} \to [0,1]$ - reward
- $P_h : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})$ - transition kernel
- H - decision horizon



Linear MDP: Linear function approximation $P_h(\cdot \mid x, a) = \phi(x,a)^T \psi_h(\cdot)$, $r_h(x,a) = \phi(x,a)^T \theta_h$.

- $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ - **known** feature map
- $\psi_h : \mathcal{X} \to \mathbb{R}^d, \theta_h \in \mathbb{R}^d, h \in [H]$ - **unknown** parameters

Aggregate Bandit Feedback (ABF) Protocol and Performance

For** episode $k = 1,2,\ldots,K$ **do

1. *Agent chooses policy $\pi^k : \mathcal{X} \times [H] \to \Delta(\mathcal{X})$*
2. *Observes trajectory $\iota^k = (x_h^k, a_h^k)_{h \in [H]}$*
3. *Observes episode reward $\nu^k = \sum_{h \in [H]} r_h^k$*

Regular Bandit feedback: Observe $(r_h^k)_{h \in [H]}$

Performance criterion:

$$\text{Regret} = \sum_{k \in [K]} V^\star(x_1) - V^{\pi^k}(x_1)$$

Where $V^\pi(x_1) = \mathbb{E}_\pi\Big[\sum_{h \in [H]} r_h\Big]$ and $V^\star(x_1) = \max_\pi V^\pi(x_1)$.

Related works

Efroni et al. (2021)	ABF, tabular MDP, $O(\sqrt{X^4 A^3 H^3 K})$ regret (computationally efficient, TS+VI)
Chatterji et al. (2021)	Binary trajectory feedback in tabular MDP (computationally intractable algorithm)
Xu et al. (2022)	ABF in offline RL (computationally intractable algorithm)
Wu & Sun (2023)	Preference-based ABF, linear MDPs, $O(\sqrt{d^6 H^5 \kappa K})$ regret (computationally efficient, TS+VI)

Our Contributions

First computationally efficient algorithms for linear MDPs with ABF

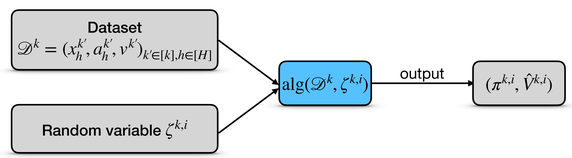
- **Randomized Ensemble Least Squares Value Iteration (RE-LSVI) Algorithm:** $\tilde{O}(d^5 H^7 K)$ regret w.h.p
- **Randomized Ensemble Policy Optimization (REPO):** $O(d^5 H^9 K)$ regret w.h.p
- **REPO (simplified for tabular MDP):** $O(\sqrt{X^4 A^3 H^7 K})$ regret w.h.p (first PO method for ABF)

Technical contributions

- Simple loose value truncation mechanism to avoid biasing the randomized bonus
- Ensemble mechanism reduces high probability optimism to obtaining constant probability

Weakly-Optimistic Algorithms

Consider an algorithm alg with the following protocol



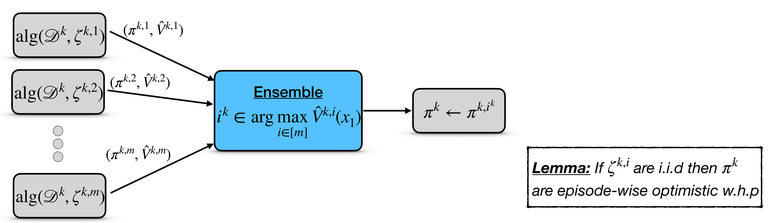
Episode-wise weakly-optimistic algorithms (Randomized Value-Iteration (VI), Thompson-Sampling (TS)):

- $\Pr(\hat{V}^{k,i}(x_1) \geq V^\star(x_1)) \geq c$ where c is a universal constant.
- The events $\{\hat{V}^{k,i}(x_1) \geq V^\star(x_1)\}_{i \in [m]}$ are statistically independent.

Epoch-wise weakly-optimistic algorithms (Randomized Policy-Optimization (PO)):

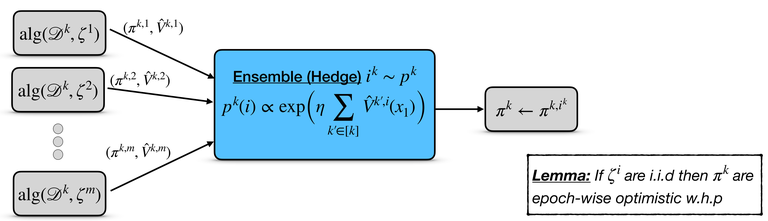
- $\Pr\Big(\sum_{k \in [K]} \hat{V}^{k,i}(x_1) \geq K V^\star(x_1)\Big) \geq c$ where c is a universal constant.
- The events $\Big\{\sum_{k \in [K]} \hat{V}^{k,i}(x_1) \geq K V^\star(x_1)\Big\}_{i \in [m]}$ are statistically independent.

Ensemble for episode-wise optimistic algorithms



Lemma: *If $\zeta^{k,i}$ are i.i.d then π^k are episode-wise optimistic w.h.p*

Ensemble for epoch-wise optimistic algorithms



Lemma: *If ζ^i are i.i.d then π^k are epoch-wise optimistic w.h.p*

Randomized PO and VI as a Weakly Optimistic Building Block (alg)

Covariances: $\Lambda^k = H \cdot I + \sum_{\tau \in \mathcal{D}^k} \phi^\tau (\phi^\tau)^T, \quad \Lambda_h^k = I + \sum_{\tau \in \mathcal{D}^k} \phi_h^\tau (\phi_h^\tau)^T$

Least-squares estimates: $\hat{\theta}^k = (\Lambda^k)^{-1} \sum_{\tau \in \mathcal{D}^k} \phi^\tau \nu^\tau, \quad \hat{\psi}_h^k V = (\Lambda_h^k)^{-1} \sum_{\tau \in \mathcal{D}^k} \phi_h^\tau V(x_{h+1}^\tau)$

Randomized Least Squares Value Iteration

For episode $k = 1,2,\ldots,K$ do:

1. **Sample**
$$\zeta^{k,i} \sim \mathcal{N}(0, \beta_r^2 (\Lambda^k)^{-1})$$
2. **For $h = H, \ldots 1$ do:**
$$w_h^{k,i} = \hat{\theta}_h^k + \hat{\psi}_h^k \hat{V}_{h+1}^{k,i}, \; b_h^k(x,a) = \beta_p \|\phi(x,a)\|_{(\Lambda_h^k)^{-1}}$$
$$\hat{Q}_h^{k,i}(x,a) = \text{clip}_{(H+1-h)\bar{\beta}}\Big[\phi(x,a)^T \Big(w_h^{k,i} + \zeta_h^{k,i} + b_h^k(x,a)\Big)\Big]$$
$$\hat{V}_h^{k,i}(x) = \max_{a \in \mathcal{A}} \hat{Q}_h^{k,i}(x,a)$$
$$\pi_h^{k,i}(x) \in \arg\max_{a \in \mathcal{A}} \hat{Q}_h^{k,i}(x,a)$$

Randomized Least Squares Policy Optimization

For episode $k = 1,2,\ldots,K$ do:

1. **Sample**
$$\zeta^{k,i} \sim \mathcal{N}(0, \beta_r^2 (\Lambda^k)^{-1} + \beta_p^2 \text{diag}(\Lambda_1^k, \ldots, \Lambda_H^k)^{-1})$$
2. **For $h = H, \ldots 1$ do:**
$$w_h^{k,i} = \hat{\theta}_h^k + \hat{\psi}_h^k \hat{V}_{h+1}^{k,i}$$
$$\hat{Q}_h^{k,i}(x,a) = \bar{\phi}(x,a)^T \Big(w_h^{k,i} + \zeta_h^{k,i}\Big)$$
$$\hat{V}_h^{k,i}(x) = \sum_{a \in \mathcal{A}} \pi_h^k(a \mid x) \hat{Q}_h^{k,i}(x,a)$$
$$\pi_h^{k+1,i}(a \mid x) \propto \pi_h^{k,i}(a \mid x) \exp(\eta \hat{Q}_h^{k,i}(x,a))$$

Analysis

Randomized bonus relaxation:

$$\arg\max_{\hat{\rho}} \beta \|\hat{\rho}\|_{\hat{\Lambda}^{-1}} + \sum_h \hat{\rho}_h^T \hat{r}_h \;\text{------>}\; \arg\max_{\hat{\rho}} \hat{\rho}^T \zeta + \sum_h \hat{\rho}_h^T \hat{r}_h \text{ where } \zeta \sim \mathcal{N}(0, \beta^2 \Lambda^{-1})$$

Weak optimism:

$$\Pr(\zeta^T \hat{\rho}^\star \geq \beta \|\hat{\rho}^\star\|_{\Lambda^{-1}}) \geq 1/9, \quad \Pr(\max_{\hat{\rho}} \zeta^T \hat{\rho} \leq \beta \sqrt{dH} \|\hat{\rho}\|_{\Lambda^{-1}}) \geq 1 - \delta$$

Optimism via ensemble:

$$\Pr(\max_{i \in [m]} \zeta_i^T \hat{\rho}^\star < \beta \|\hat{\rho}^\star\|_{\Lambda^{-1}}) = \prod_{i \in [m]} \Pr(\zeta_i^T \hat{\rho}^\star < \beta \|\hat{\rho}^\star\|_{\Lambda^{-1}}) \leq (8/9)^m \leq \delta \text{ where } m = 9 \log(1/\delta)$$

$$\Pr(\max_{i \in [m], \hat{\rho}} \zeta^T \hat{\rho} \leq \beta \sqrt{dH \log m} \|\hat{\rho}\|_{\Lambda^{-1}}) \geq 1 - \delta$$

Regret decomposition: $\text{regret} = \underbrace{\sum_{k \in [K]} V^\star(x_1) - \hat{V}_1^{k,i_k}(x_1)}_{(i)} + \underbrace{\sum_{k \in [K]} \hat{V}_1^{k,i_k}(x_1) - V^{\pi^k}(x_1)}_{(ii)}$

Value difference:

$$V^\star(x_1) - \hat{V}_1^{k,i}(x_1) = \mathbb{E}_{P,\pi^\star} \underbrace{\sum_{h \in [H]} \hat{Q}_h^{k,i}(x_h, \pi_h^\star(x_h)) - \hat{Q}_h^{k,i}(x_h, \pi_h^{k,i}(x_h))}_{(**) \leq 0} + \underbrace{\mathbb{E}_{P,\pi^\star} \sum_{h \in [H]} \phi(x_h, a_h)^T (\theta_h + \psi_h \hat{V}_{h+1}^{k,i}) - \hat{Q}_h^{k,i}(x_h, a_h)}_{(*)}$$

Concentration bounds:

$$\|(\psi_h - \widehat{\psi}_h^k) \hat{V}_{h+1}^{k,i}\|_{\Lambda_h^k} \leq \beta_p, \quad |\phi(x,a)^T (\hat{\theta}_h^k + \zeta_h^{k,i})| \leq \bar{\beta}, \;\; \forall x \in \mathcal{X}, a \in \mathcal{A}$$

$$\|(\theta_h - \hat{\theta}_h^k)\|_{\Lambda^k} \leq \beta_r, \qquad \max_{i \in [m]} \phi^{\pi^\star T} \zeta^{k,i} \geq \beta_r \|\phi^{\pi^\star}\|_{(\Lambda^k)^{-1}}$$

Loose clipping:

$$\phi(x,a)^T (w_h^{k,i} + \zeta_h^{k,i}) + \beta_p \|\phi(x,a)\|_{(\Lambda_h^k)^{-1}} = \phi(x,a)^T (\hat{\theta}_h^k + \zeta_h^{k,i} + \hat{\psi}_h^k \hat{V}_{h+1}^{k,i}) + \beta_p \|\phi(x,a)\|_{(\Lambda_h^k)^{-1}}$$

$$\geq \phi(x,a)^T (\hat{\theta}_h^k + \zeta_h^{k,i} + \psi_h \hat{V}_{h+1}^{k,i}) + (\beta_p - \|(\hat{\psi}_h^k - \psi_h) \hat{V}_{h+1}^{k,i}\|_{\Lambda_h^k}) \|\phi(x,a)\|_{(\Lambda_h^k)^{-1}} \geq \phi(x,a)^T (\hat{\theta}_h^k + \zeta_h^{k,i} + \psi_h \hat{V}_{h+1}^{k,i})$$

$$\implies \hat{Q}_h^{k,i}(x,a) \geq \phi(x,a)^T (\hat{\theta}_h^k + \zeta_h^{k,i} + \psi_h \hat{V}_{h+1}^{k,i})$$

Bounding (i):

$$(*) \leq \mathbb{E}_{P,\pi^\star} \sum_{h \in [H]} \phi(x_h, a_h)^T (\theta_h - \hat{\theta}_h^k - \zeta_h^{k,i}) = \phi^{\pi^{\star T}} (\theta - \hat{\theta}^k - \zeta^{k,i})$$

$$(i) = \min_{i \in [m]} V^\star(x_1) - \hat{V}_1^{k,i}(x_1) \leq \phi^{\pi^{\star T}} (\theta - \hat{\theta}^k) - \max_{i \in [m]} \phi^{\pi^{\star T}} \zeta^{k,i} \leq \beta_r \|\phi^{\pi^\star}\|_{(\Lambda^k)^{-1}} - \max_{i \in [m]} \phi^{\pi^{\star T}} \zeta^{k,i} \leq 0$$