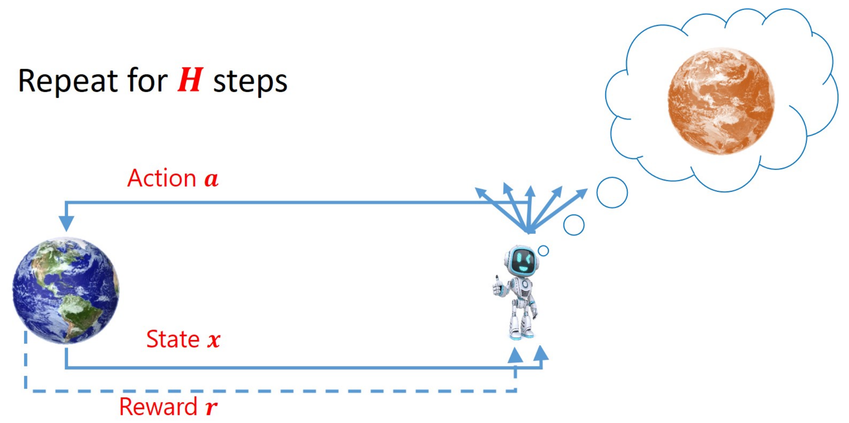
Model-based RL with Optimistic Posterior Sampling: Structural Conditions and Sample Complexity

Alekh Agarwal — Google Research

Tong Zhang — Google Research & HKUST

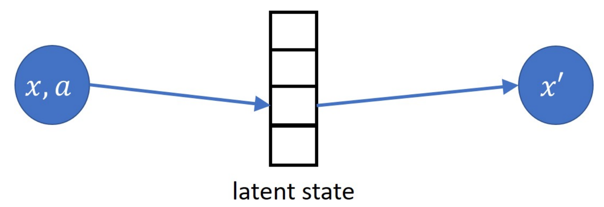
Model-based RL



Goal: Find policy π s.t. $V(\pi_\star) - V(\pi) \leq \epsilon$.
Key challenge: State and action spaces can be arbitrarily large.
Our work: Unified statistical & algorithmic framework for sample-efficiency.

Problem Setup

True MDP $M_\star = (P_\star, R_\star)$: $x^{h+1} \sim P_\star^h(\cdot|x^h, a^h)$ and $r^h \sim R_\star(\cdot|x^h, a^h)$.
Model class $\mathcal{M}$ of tuples $M = (P, R)$.
Realizability assumption: $M_\star \in \mathcal{M}$.
Optimal value function V_M under M.
Linear MDP: $P_\star(x'|x,a) = \phi_\star(x,a)^\top \mu_\star(x')$, $\phi_\star$ is known.
Low-rank MDP: $P_\star(x'|x,a) = \phi_\star(x,a)^\top \mu_\star(x')$, $\phi_\star$ is *unknown*.
KNR: $x' = W_\star \phi_\star(x,a) + \mathcal{N}(0, \sigma^2 I)$, $\phi_\star$ is known.



$P(x'|x,a) = \phi_\star(x,a)^\top \mu_\star(x')$

Model-based Optimistic Posterior Sampling (MOPS)

Require: Model class $\mathcal{M}$, prior $p_0 \in \Delta(\mathcal{M})$, policy generator π_{gen}, learning rates η, η' and optimism coefficient γ.

Set $S_0 = \emptyset$.
for $t = 1, \ldots, T$ **do**
  Observe $x_t^1 \sim \mathcal{D}$ and draw $h_t \sim \{1, \ldots, H\}$ uniformly at random.
  $L_s^h(M) = -\eta \underbrace{(R_M^h(x_s^h, a_s^h) - r_s^h)^2}_{\text{reward fit}} + \eta' \underbrace{\ln P_M^h(x_s^{h+1} \mid x_s^h, a_s^h)}_{\text{transition likelihood}}$. ▷ Likelihood
  Posterior $p_t(M) = p(M|S_{t-1}) \propto p_0(M) \exp(\sum_{s=1}^{t-1} (\gamma \underbrace{V_M(x_s^1)}_{\text{optimistic bias}} + L_s^{h_s}(M)))$.
  ▷ Optimistic posterior sampling update
  Let $\pi_t = \pi_{\mathrm{gen}}(h_t, p_t)$ ▷ policy generation
  Execute π_t for $h = 1, \ldots, h_t$, and observe $\{(x_t^h, a_t^h, r_t^h, x_t^{h+1})_{h=1}^{h_t}\}$
  Update $S_t = S_{t-1} \cup \{x_t^h, a_t^h, r_t^h, x_t^{h+1}\}$ for $h = h_t$.
end for
return $(\pi_1, \ldots, \pi_T)$.

Model fit using log-likelihood. No complicated divergences needed.
Optimism crucial to worst-case guarantees for posterior sampling.
Policy generator allows adapting exploration to problem structure.
- **Q-type, e.g. linear MDP:** $\pi_{\mathrm{gen}}(h,p) = \pi_M$, for $M \sim p$.
- **V-type, e.g. low-rank MDP:** $\pi_{\mathrm{gen}}(h,p) = \pi_M$, $M \sim p$ till $h-1$, Unif($\mathcal{A}$) at h.
- Also extends to V-type with *infinite actions.*

A Summary of the Results

General approach: MOPS is applicable whenever (near) optimal planning and likelihood/posterior are tractable.
Flexible theory: Sample complexity $\approx$ (complexity of $\mathcal{M}$)$\cdot$(MDP structure).
Strong guarantees: Bounds in most known RL settings close to optimal.
Novel decoupling generalizes most prior MDP structural assumptions.

A Regret Decomposition

Model-based Bellman error:

$$\mathcal{E}_B(M, x^h, a^h) = \underbrace{Q_M^h(x^h, a^h)}_{\text{optimal value under } M \text{ at } x^h, a^h} - \underbrace{(P_\star{}^h[r^h + V_M^h])(x^h, a^h)}_{\text{one-step backup in } M_\star \text{ of optimal value under } M}$$

Regret lemma for model-based RL:

Lemma 1 ([Sun et al., 2019]). For any context x^1 and model M, let π_M be the optimal policy in M and $\Delta V_M(x^1) = V_M(x^1) - V_\star(x^1)$. Then we have

$$\underbrace{V_\star(x^1) - V^{\pi_M}(x^1)}_{\text{Regret of } \pi_M} = \sum_{h=1}^{H} \mathop{\mathbb{E}}_{x^h, a^h \sim \pi_M} [\mathcal{E}_B(M, x^h, a^h)|x^1] - \Delta V_M(x^1).$$

High-level idea: Bound Bellman error of M using its likelihood and ΔV_M using optimism.
Key challenge: Need Bellman error control under π_M for each M.

Does a single exploration policy suffice?

A Decoupling Condition for Exploration

Definition 1 (Hellinger Decoupling of Bellman Error)**.** *Let a distribution* $p \in \Delta(\mathcal{M})$ *and a policy* $\pi(x^h, a^h|x^1)$ *be given. Then* $\mathrm{dc}^h(\epsilon, p, \pi, \alpha) = \inf_{c^h \geq 0} c^h$ *s.t.*

$$\underbrace{\mathop{\mathbb{E}}_{M \sim p}}_{\text{posterior } p} \underbrace{\mathop{\mathbb{E}}_{(x^h, a^h) \sim \pi_M(\cdot|x^1)} \mathcal{E}_B(M, x^h, a^h)}_{\text{RHS of Lemma 1}}$$
$$\leq \left(c^h \mathop{\mathbb{E}}_{M \sim p} \underbrace{\mathop{\mathbb{E}}_{(x^h, a^h) \sim \pi(\cdot|x^1)}}_{\text{policy generator } \pi} \underbrace{\ell^h(M, x^h, a^h)}_{\text{likelihood of } M} \right)^\alpha + \epsilon,$$

for $\ell^h(M, x^h, a^h) = D_H(P_M(\cdot|x^h, a^h), P_\star(\cdot|x^h, a^h))^2 + (R_M(x^h, a^h) - R_\star(x^h, a^h))^2$.

Sample Complexity of MOPS under Decoupling

Theorem 1. *Assume* $M_\star \in \mathcal{M}$ *and suppose that there exists* $0 < \alpha \leq 0.5$ *such that for all* p, $\mathrm{dc}^h(\epsilon, p, \pi_{\mathrm{gen}}(h,p), \alpha) \leq \mathrm{dc}^h(\epsilon, \alpha)$. *Define*

$$\mathrm{dc}(\epsilon, \alpha) = \left(\frac{1}{H} \sum_{h=1}^{H} \mathrm{dc}^h(\epsilon, \alpha)^{\alpha/(1-\alpha)} \right)^{(1-\alpha)/\alpha}.$$

Using $\eta = \eta' = 1/6$ *and* $\gamma \leq 0.5$*, then the following bound holds for* MOPS*:*

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[V_\star(x_t^1) - \mathop{\mathbb{E}}_{M \sim p_t} V_M(x_t^1) \right] = O\left(H \left(\frac{\mathrm{dc}(\epsilon, \alpha) \ln |\mathcal{M}|}{T} \right)^\alpha + \epsilon H \right).$$

$1/\sqrt{T}$ bound for $\alpha = 0.5$. Extends to infinite $\mathcal{M}$.

Analysis sketch: By Lemma 1:

$$\sum_{t=1}^{T} \mathbb{E} \left[V_\star(x_t^1) - \mathop{\mathbb{E}}_{M \sim p_t} V_M(x_t^1) \right]$$
$$\leq \sum_{t=1}^{T} \mathbb{E} \mathop{\mathbb{E}}_{M \sim p_t} \left[\frac{1}{\gamma} \ell^{h_t}(M, x_t^{h_t}, a_t^{h_t}) - \Delta V_M(x_1^t) \right] + T H^{\frac{1}{1-\alpha}} (\mathrm{dc}(\epsilon, \alpha)\gamma)^{\frac{\alpha}{1-\alpha}} \quad \text{(decoupling)}$$
$$\leq \gamma^{-1} \ln |\mathcal{M}| + 2\gamma T + T\, H^{\frac{1}{1-\alpha}} (\mathrm{dc}(\epsilon, \alpha)\gamma)^{\frac{\alpha}{1-\alpha}}. \quad \text{(online learning convergence)}$$

V-type Decoupling and Witness Rank

Assumption: Suppose there is a function class $\mathcal{G}$ with $g(x, a, x') \in [0, 1]$ and let $f(x, a, r, x') = r + g(x, a, x')$. Assume that for all $M, M' \in \mathcal{M}$, x^1 and h, there are maps $\psi^h(M, x^1)$ and $u^h(M', x^1)$ such that with $\|u^h(M', x^1)\|_2 \leq B_1$, and:
1. $\mathbb{E}_{x^h \sim \pi_M | x^1} \mathbb{E}_{a^h \sim \pi_{M'}(x^h)} (P_{M'}^h f)(x^h, a^h) - (P_\star^h f)(x^h, a^h) = \langle \psi^h(M, x^1), u^h(M', x^1) \rangle$, and
2. $V_M(x') \in \mathcal{G}$ for all $M \in \mathcal{M}$.

Examples: Cover in ψ gives exploration, $f = V_M$ gives Bellman error of M.
- Low-rank MDP
- Low witness rank MDPs [Sun et al., 2019]

Lemma 2. If $|\mathcal{A}| = K$ and V-type decoupling holds:

$$\mathrm{dc}(\epsilon, p, \pi_{\mathrm{gen}}(h,p), 0.5) \leq 4K \dim(\psi^h), \quad \text{where } \pi_{\mathrm{gen}}(h,p) = p \circ^h \mathrm{Unif}(\mathcal{A}).$$

Sample complexity of MOPS: $\mathcal{O}\left(\sqrt{\frac{H^2 d^2 K \ln|\mathcal{M}|}{T}}\right)$.
Improves upon earlier bound of Sun et al. [2019].
General result for infinite $\mathcal{A}$ using linear embeddability of backup errors (always holds for finite $\mathcal{A}$). Requires $\alpha = 0.25$, leading to $T^{-1/4}$ bound.

Q-type Decoupling and Linear Models

Assumption: Suppose there is a function class $\mathcal{G}$ with $g(x, a, x') \in [0, 1]$ and let $f(x, a, r, x') = r + g(x, a, x')$. Assume that for all $M \in \mathcal{M}$, x^1 and h, there are maps $\psi^h(x^h, a^h)$ and $u^h(M, f)$ such that:
1. $(P_{M'}^h f)(x^h, a^h) - (P_\star^h f)(x^h, a^h) = \langle \psi^h(x^h, a^h), u^h(M, f) \rangle$, and
2. $V_M(x') \in \mathcal{G}$ for all $M \in \mathcal{M}$.

Decomposition happens point-wise rather than in expectation.
Examples:
- Linear MDP
- Kernelized Non-linear Regulator [Kakade et al., 2020] and LQR

Lemma 3. If Q-type decoupling holds:

$$\mathrm{dc}(\epsilon, p, \pi_{\mathrm{gen}}(h,p), 0.5) \leq 4\dim(\psi^h), \quad \text{where } \pi_{\mathrm{gen}}(h,p) = \pi_M \text{ with } M \sim p.$$

KNR sample complexity: $\widetilde{\mathcal{O}}\left(\sqrt{\frac{H^2 d_\phi^2 d_{\mathcal{X}}}{T\sigma^2}}\right)$.
Sub-optimal in d_ϕ and H factors. Latter due to learning at one time-step h_t.
Similar result for linear mixture MDPs.

References

Sham M. Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. In *NeurIPS*, 2020.

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pages 2898–2933. PMLR, 2019.

Author contact: {alekhagarwal, tozhang}@google.com