Bilinear Exponential Family of MDPs: Frequentist Regret Bound with Tractable Exploration & Planning

Reda Ouhamma*, Odalric Maillard*, Debabrota Basu*

* Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France.

Setting and contributions

Episodic RL

- Bilinear exponential family (BEF) model:

$$\mathbb{P}(\tilde{s} \mid s, a) = \exp\left(\psi(\tilde{s})^\top M_{\theta^p} \varphi(s,a) - Z^p_{s,a}(\theta^p)\right)$$
$$\mathbb{P}(r \mid s, a) = \exp\left(r\, B^\top M_{\theta^r} \varphi(s,a) - Z^r_{s,a}(\theta^r)\right).$$

- Minimizes (pseudo-)regret:

$$\mathcal{R}(K) \triangleq \sum_{k=1}^{K} \left(V^{\pi^\star}_{\theta,1}(s_1^k) - V^{\pi^t}_{\theta,1}(s_1^k)\right).$$

Contributions: We investigate episodic RL problem with unknown rewards and transition. Our contributions are:

1. A Linear value observation for BEF transitions, generalizing the Gaussian transition observation of [1]
2. An algorithm with: tractable exploration, tractable planning, and a $\tilde{\mathcal{O}}(\sqrt{d^3H^3K})$ regret upper-bound.
3. A clipping-free algorithm thanks to an improved elliptical lemma.

$\mathtt{BEF-RLSVI}$ algorithm

$\mathtt{BEF-RLSVI}$ is similar to $\mathtt{RLSVI}$, and is clipping-free.

Algorithm 1 $\mathtt{BEF-RLSVI}$

Input: failure rate δ, constants α^p, η and $(x_k)_{k\in[K]} \in \mathbb{R}^+$
for episode $k = 1, 2, \dots$ **do**
Observe initial state s_1^k
Sample noise $\xi_k \sim \mathcal{N}\left(0, x_k (G^p)^{-1}\right)$ such that

$$G^p = \frac{\eta}{\alpha^p}\mathbb{A} + \sum_{\tau=1}^{k-1}\sum_{h=1}^{H}(\varphi(s_h^\tau, a_h^\tau)^\top A_i^\top A_j \varphi(s_h^\tau, a_h^\tau))_{i,j\in[d]}$$

Perturb reward parameter: $\tilde{\theta}^r(k) = \hat{\theta}^r(k) + \xi_k$
Compute $(\tilde{Q}_h^k)_{h\in[H]}$ via Bellman-backtracking, see Algorithm 2
for $h = 1, \dots, H$ **do**
Pull action $a_h^k = \arg\max_a \tilde{Q}_h^k(s_h^k, a)$
Observe reward $r(s_h^k, a_h^k)$ and state s_{h+1}^k.
end for
Update the penalized ML estimators $\hat{\theta}^p(k), \hat{\theta}^r(k)$
end for

Unlike optimistic approaches, exploration here is explicit and efficient as it does not involve a high-dimensional optimization.

Algorithm 2 Bellman Backtracking

Input Parameters $\hat{\theta}^p, \hat{\theta}^r$, initialize $\tilde{\theta} = (\tilde{\theta}^r, \hat{\theta}^p)$ and $\forall s, V_{H+1}(s) = 0$
for steps $h = H-1, H-2, \cdots, 0$ **do**
Calculate $Q_{\tilde{\theta},h}(s,a) = \mathbb{E}^{\tilde{\theta}^r}_{s,a}[r] + \langle \phi^p(s,a), \int V_{\tilde{\theta},h+1}(s')\mu^p(s')ds'\rangle_{\mathcal{H}}$.
end for

Why is $\mathtt{BEF-RLSVI}$ tractable?

Planning:

For an MDP of the BEF, we can write the state-action value function linearly, at step h:

$$\tilde{Q}^\pi_h(s,a) = \mathbb{E}^{\tilde{\theta}^r}[r(s,a)] + \left\langle \phi^p(s,a), \int_{\mathcal{S}} \mu^p(\tilde{s})\tilde{V}h + 1^\pi(\tilde{s})d\tilde{s}\right\rangle.$$

Key facts:

- ϕ^p and ψ^p are in an RKHS, *i.e.* infinite dimensional
- Using Random Fourier Transform entails $\mathcal{O}(pH^2K\log(HK))$ dimensional approximations of ϕ^p and ψ^p
- Therefore, the planning has a $\mathcal{O}(pH^3K\log(HK))$ complexity, pseudo-polynomial in p, H and K, thus tractable.

Maximum likelihood estimation:

There are different methods to approximate ML estimator:

- Integral approximation techniques:
  - Simulated annealing and importance sampling
  - MCMC techniques for approximating the partition function.
  - Optimizing a different objective, *the contrastive divergence*, yields a good approximation.
- If the natural parameter and support of the distribution are bounded, an ϵ-approximation can be derived in $O(\mathrm{poly}(k/\epsilon))$
- Score matching: avoids approximating the partition function. Under certain conditions, the estimation can be solved in $O(d^3)$

Regret bound

Theorem (regret bound): Let $G_{s,a} \triangleq (\varphi(s,a)^\top A_i^\top A_j \varphi(s,a))_{i,j\in[d]}$ and $\mathbb{A} \triangleq (\mathrm{tr}(A_i A_j^\top))_{i,j\in[d]}$. Under regularity of the Hessian and assuming

1. $\max\{\|\theta^r\|_{\mathbb{A}}, \|\theta^p\|_{\mathbb{A}}\} \leq B_{\mathbb{A}}, \ \|\mathbb{A}^{-1}G_{s,a}\| \leq B_{\varphi,\mathbb{A}}$ and $\mathbb{E}^{\theta^r}[r(s,a)] \in [0,1]$
2. The noise $\xi_k \sim \mathcal{N}(0, x_k(G^p)^{-1})$ satisfies $x_k \gtrsim dH^2$

then for all $\delta \in (0,1]$, with probability at least $1 - 7\delta$,

$$\mathcal{R}(K) = \mathcal{O}(\sqrt{d^3H^3K}).$$

Tightness of regret upper-bound:

- A lower bound for episodic RL with continuous state-action spaces is still missing.
- For tabular RL, [2] proves a lower bound of order $\Omega(\sqrt{H^3SAK})$
- A tabular MDP is also a BEF model with $d = S^2 \times A$
- $\mathtt{BEF-RLSVI}$'s yields $R(K) = O(\sqrt{(S^2A)^3H^3K})$, **tight in H and K.**

Interesting proof bits

Optimism: Key reasons for choosing $\mathtt{RLSVI}$-type algorithms:

- Perturbing the reward estimation guarantees optimism with a constant probability
- A constant probability of optimism is enough to control the value function approximation error

Transportation: Using transportation inequalities instead of the simulation lemma (*c.f.* Lemma 1 in [1]) reduces a $\sqrt{H}$ regret factor

Elliptical lemma:

- Leveraging the boundedness of the true value function enables using an improved elliptical lemma ($\sqrt{H}$ less than [3])
- The norm of features can only be large $\mathcal{O}(d)$ times, thus, we can omit clipping and reduce the regret by $\sqrt{d}$ compared to [4].

Approximate planning:

- To guarantee a tractable planning, we approximate the transition with $(1/\sqrt{H^2K})$-error. Using mis-specification style analysis, we show that the approximation doesn't hinder the regret bound.
- Using a Linear-RL algorithm directly on top of the approximation would lead to a linear regret.

Conclusion

- For episodic RL with BEF transitions and rewards:
  1. We propose $\mathtt{BEF-RLSVI}$ that achieves a $\mathcal{O}(\sqrt{d^3H^3K})$ regret
  2. We show that tractable planning and exploration are possible
  3. We give the second example of continuous linear MDPs in literature, although both are infinite dimensional
- For linear RL style analyses: The occurrences of values outside the plausible range, *e.g.* $\hat{V} \notin [0, H]$, are finite
- Future work:
  1. The paper could be complemented by experimental evaluations on relevant tasks.
  2. The tractability of planning can be extended to any shift invariant kernel: this can lead to interesting generalizations.

Acknowledgements: The authors acknowledge the funding of the French National Research Agency, the French Ministry of Higher Education and Research, Inria, the MEL and the I-Site ULNE regarding project R-PILOTE-19-004-APPRENF

References

[1] Tongzheng Ren, Tianjun Zhang, Csaba Szepesvári, and Bo Dai. A free lunch from the noise: Provable and practical exploration for representation learning. In *Uncertainty in Artificial Intelligence*. PMLR, 2022.

[2] Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*. PMLR, 2021.

[3] Sayak Ray Chowdhury, Aditya Gopalan, and Odalric-Ambrym Maillard. Reinforcement learning in parametric mdps with exponential families. In *AISTATS*. PMLR, 2021.

[4] Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *AISTATS*. PMLR, 2020.