Some recent progress in continuous-time reinforcement learning

Yufei Zhang

Statistics Research Showcase Day

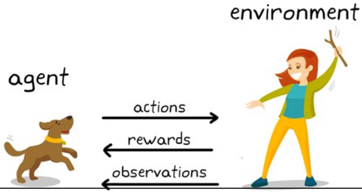
based on joint works with Matteo Basei (EDF R&D), Xin Guo (UC Berkeley), Anran Hu (UC Berkeley), Lukasz Szpruch (U of Edinburgh, Turing) and Tanut Treetanthiploet (Turing).

15 June 2022

- Stochastic control problems are ubiquitous.
- Continuous-time models well understood in this community.
  - optimal trading, dynamic hedging, autonomous driving, robots.
- Reinforcement learning (RL) methods increasingly popular.
- Analysis restricted to discrete-time models.

My research:

- systematically understand the performance of artificial agents in a continuous-time environment.



Learning algorithm focuses on policy, i.e., a function mapping system states to actions.

For a system with unknown parameter θ, issues in RL:

- Model identification: how to learn the parameter θ?
  Examples: consumer behaviour for online retailer, price impact factor in optimal execution.

Learning algorithm focuses on policy, i.e., a function mapping system states to actions.

For a system with unknown parameter θ, issues in RL:

- Model identification: how to learn the parameter θ?
  Examples: consumer behaviour for online retailer, price impact factor in optimal execution.
- Robustness of policies w.r.t. θ.

Learning algorithm focuses on policy, i.e., a function mapping system states to actions.

For a system with unknown parameter θ, issues in RL:

- Model identification: how to learn the parameter θ?
  Examples: consumer behaviour for online retailer, price impact factor in optimal execution.
- Robustness of policies w.r.t. θ.
- Convergence rate analysis:
  - critical for understanding algorithm efficiency.

Overview of related works

Discrete-time RL (partial)

- Tabular MDPs (finite states and actions): Watkins and Dayan 1992 (Q-learning), Williams 1992 (policy gradient), Jaksch, Ortner and Auer 2009, and many others.
- Infinite states and actions: (LQ-RL, $T = \infty$)
  - Sublinear regret: Abbasi-Yadkori and Szepesvari 2011, Dean et al 2018, Mania, Tu and Recht 2019, Cohen, Koren and Mansour 2019
  - Logarithmic regret (for special cases): Faradonbeh, Tewari and Michailidis 2020, Cassel, Cohen and Koren 2020, Lale, Azizzadenesheli, Hassibi and Anandkumar 2020

- Algorithm design: Modares and Lewis 2014, Doya 2000, Tallec, Blier and Ollivier 2019, Jia and Zhou 2021a, 2021b
- Asymptotic convergence analysis
  - LQ-RL ($T=\infty$): Duncan, Guo and Pasik-Duncan 1999, Rizvi and Lin 2018, Pang, Bian and Jiang 2020

— No convergence rate/regret analysis

- Algorithm design: Modares and Lewis 2014, Doya 2000, Tallec, Blier and Ollivier 2019, Jia and Zhou 2021a, 2021b
- Asymptotic convergence analysis
  - LQ-RL ($T=\infty$): Duncan, Guo and Pasik-Duncan 1999, Rizvi and Lin 2018, Pang, Bian and Jiang 2020

— No convergence rate/regret analysis

This talk:

- analyses regret of continuous-time RL over finite-time horizon.
- starts with linear-quadratic models and then extends to linear-convex models.

Fix $(A^\star, B^\star) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times k}$, minimise

$$J(\alpha; \theta^\star) = \mathbb{E}\left[\int_0^T \left((X_t^{\theta^\star,\alpha})^\top Q X_t^{\theta^\star,\alpha} + \alpha_t^\top R \alpha_t \right) \mathrm{d}t \right],$$

where $X^{\theta^\star,\alpha}$ satisfies the dynamics with parameter $\theta^\star = (A^\star, B^\star)$:

$$\mathrm{d}X_t = (A^\star X_t + B^\star \alpha_t)\,\mathrm{d}t + \mathrm{d}W_t, \quad X_0 = x_0,$$

and α is adapted to the information generated by $X^{\theta^\star,\alpha}$. Here Q and R are given positive definite matrices.

- When $\theta^\star$ is known, the optimal control $\alpha^{\theta^\star}$ is given by:

$$\alpha_t^{\theta^\star} = \phi_t^{\theta^\star}(X_t^{\theta^\star}),$$

where

  - $X^{\theta^\star}$ is the optimal state process satisfying

$$\mathrm{d}X_t = (A^\star X_t + B^\star \phi_t^{\theta^\star}(X_t))\mathrm{d}t + \mathrm{d}W_t,$$

  - $\phi_t^{\theta^\star}(x) = K_t^{\theta^\star} x$, where $K^{\theta^\star}$ solves a Riccati ODE associated with $\theta^\star$.

- When $\theta^\star$ is known, the optimal control $\alpha^{\theta^\star}$ is given by:

$$\alpha_t^{\theta^\star} = \phi_t^{\theta^\star}(X_t^{\theta^\star}),$$

  where
  - $X^{\theta^\star}$ is the optimal state process satisfying

$$\mathrm{d}X_t = (A^\star X_t + B^\star \phi_t^{\theta^\star}(X_t))\mathrm{d}t + \mathrm{d}W_t,$$

  - $\phi_t^{\theta^\star}(x) = K_t^{\theta^\star} x$, where $K^{\theta^\star}$ solves a Riccati ODE associated with $\theta^\star$.
- When $\theta^\star$ is unknown, one needs to balance exploration (learning via interactions with the random environment) and exploitation (optimal control).

- Let $\hat{\theta}^{(m-1)}$ be the estimated parameter before m-th episode.

- Let $\hat{\theta}^{(m-1)}$ be the estimated parameter before m-th episode.
- Given $\hat{\theta}^{(m-1)}$, agent exercises a policy $\phi^{(m)}$ (which may depend on $\hat{\theta}^{(m-1)}$ or not) and observes a trajectory of

$$\mathrm{d}X_t^m = (A^\star X_t^m + B^\star \phi_t^{(m)}(X_t^m))\mathrm{d}t + \mathrm{d}W_t^m,$$

Cost for the m-th episode is

$$J(\phi^{(m)}; \theta^\star) = \mathbb{E}\left[\int_0^T \left((X_t^m)^\top Q X_t^m + \phi_t^{(m)}(X_t^m)^\top R \phi_t^{(m)}(X_t^m)\right) \mathrm{d}t\right].$$

Learning via trial and error

Episodic setting

- Let $\hat{\theta}^{(m-1)}$ be the estimated parameter before m-th episode.
- Given $\hat{\theta}^{(m-1)}$, agent exercises a policy $\phi^{(m)}$ (which may depend on $\hat{\theta}^{(m-1)}$ or not) and observes a trajectory of

$$\mathrm{d}X_t^m = (A^\star X_t^m + B^\star \phi_t^{(m)}(X_t^m))\mathrm{d}t + \mathrm{d}W_t^m,$$

  Cost for the m-th episode is

$$J(\phi^{(m)};\theta^\star) = \mathbb{E}\left[\int_0^T \left((X_t^m)^\top Q X_t^m + \phi_t^{(m)}(X_t^m)^\top R \phi_t^{(m)}(X_t^m)\right)\, \mathrm{d}t\right].$$

- Agent constructs $\hat{\theta}^{(m)}$ using observed trajectories of $(X^i)_{i=1}^m$.

- Let $\hat{\theta}^{(m-1)}$ be the estimated parameter before m-th episode.
- Given $\hat{\theta}^{(m-1)}$, agent exercises a policy $\phi^{(m)}$ (which may depend on $\hat{\theta}^{(m-1)}$ or not) and observes a trajectory of

$$\mathrm{d}X_t^m = (A^\star X_t^m + B^\star \phi_t^{(m)}(X_t^m))\mathrm{d}t + \mathrm{d}W_t^m,$$

  Cost for the m-th episode is

$$J(\phi^{(m)};\theta^\star) = \mathbb{E}\left[\int_0^T \left((X_t^m)^\top Q X_t^m + \phi_t^{(m)}(X_t^m)^\top R \phi_t^{(m)}(X_t^m)\right)\,\mathrm{d}t\right].$$

- Agent constructs $\hat{\theta}^{(m)}$ using observed trajectories of $(X^i)_{i=1}^m$.
- Performance criteria – regret up to episode $N \in \mathbb{N}$:

$$\mathcal{R}(N) = \sum_{m=1}^N \left(J(\phi^{(m)};\theta^\star) - J(\phi_{\theta^\star};\theta^\star)\right).$$

- Take actions $(\phi^{(1)}, \phi^{(2)}, \ldots)$ to learn $(A^\star, B^\star)$ and minimise $\mathcal{R}$.

Theorem

$\theta^\star$ is identifiable under ϕ^e iff

- *if $u \in \mathbb{R}^d$ and $v \in \mathbb{R}^p$ satisfy $u^\top x + v^\top \phi^e(t,x) = 0$ for all $(t,x) \in [0,T] \times \mathbb{R}^d$, then u and v are zero vectors.*

Theorem

$\theta^\star$ is identifiable under ϕ^e iff

- *if $u \in \mathbb{R}^d$ and $v \in \mathbb{R}^p$ satisfy $u^\top x + v^\top \phi^e(t,x) = 0$ for all $(t,x) \in [0,T] \times \mathbb{R}^d$, then u and v are zero vectors.*

For any given $\theta = (A,B)$, the greedy policy $\phi_\theta(t,x) = K_t^\theta x$ identifies $\theta^\star$ iff B is full column rank.

- Self-exploration: if $B^\star$ is full column rank, the greedy policy $\phi_{\theta^\star}$ explores the environment, and hence explicit exploration is not required.

- Given fixed policy ϕ, estimate $\theta^\star = (A^\star, B^\star)$ via

$$\mathrm{d}X_t = \theta^\star Z_t^\phi \mathrm{d}t + \mathrm{d}W_t, \quad Z_t^\phi = (X_t, \phi(t, X_t))^\top.$$

- Agent only observes the state process X (also Z^ϕ), but not the corresponding Brownian path.
- View unknown $\theta^\star$ as a hidden random variable, and observe the log-likelihood of $\theta^\star$ is quadratic.

Likelihood function

- Discrete time approximation is given by

$$X_{k+1} - X_k = \theta^\star Z_k^\phi \Delta t + \sqrt{\Delta t}\xi_k\,, \quad \text{and} \quad \xi_k \sim_{IID} N(0,1).$$

- Discrete time approximation is given by

$$X_{k+1} - X_k = \theta^\star Z_k^\phi \Delta t + \sqrt{\Delta t}\xi_k\,, \quad \text{and} \quad \xi_k \sim_{IID} N(0,1).$$

- One can compute likelihood function

$$\ell(\theta^\star \mid X_1, \dots, X_n) \propto$$
$$\exp\left(-\frac{1}{2}\theta^\star \Big(\sum_{k=0}^{n-1} Z_k^\phi (Z_k^\phi)^\top \Delta t \Big) (\theta^\star)^\top + \theta^\star \sum_{k=0}^{n-1} Z_k^\phi (X_{k+1} - X_k) \right),$$

- Discrete time approximation is given by

$$X_{k+1} - X_k = \theta^\star Z_k^\phi \Delta t + \sqrt{\Delta t}\xi_k\,, \quad \text{and} \quad \xi_k \sim_{IID} N(0,1).$$

- One can compute likelihood function

$$\ell(\theta^\star \mid X_1, \dots, X_n) \propto$$

$$\exp\left(-\frac{1}{2}\theta^\star\Big(\sum_{k=0}^{n-1} Z_k^\phi (Z_k^\phi)^\top \Delta t\Big)(\theta^\star)^\top + \theta^\star \sum_{k=0}^{n-1} Z_k^\phi (X_{k+1} - X_k)\right),$$

which leads to

$$\ell(\theta^\star \mid X) \propto \exp\left(-\frac{1}{2}\theta^\star\Big(\int_0^t (Z_s^\phi)(Z_s^\phi)^\top \mathrm{d}s\Big)(\theta^\star)^\top + \theta^\star \int_0^t Z_s^\phi \mathrm{d}X_s\right).$$

Posterior distribution

Given a prior $\theta^\star \sim N(\hat{\theta}^{(m-1)}, V^{(m-1)})$, the posterior of $\theta^\star$ based on observation Z^ϕ is also Gaussian $N(\hat{\theta}^{(m)}, V^{(m)})$.

Given a prior $\theta^\star \sim N(\hat{\theta}^{(m-1)}, V^{(m-1)})$, the posterior of $\theta^\star$ based on observation Z^ϕ is also Gaussian $N(\hat{\theta}^{(m)}, V^{(m)})$.

Theorem

For all $\delta > 0$,

$$\left|\hat{\theta}^{(m)} - \theta^\star\right|^2 \leq \frac{1}{\lambda_{\min}((V^{(m)})^{-1})} \,\mathrm{poly}(\ln m, \ln\left(\tfrac{1}{\delta}\right)), \quad \forall m \geq 2.$$

- $\lambda_{\min}((V^{(m)})^{-1})$ increases if ϕ^e is executed.
- Sub-exponential concentration: Z_t^ϕ is Gaussian, and hence $\int_0^T Z_t^\phi (Z_t^\phi)^\top \mathrm{d}t$ and $\int_0^T Z_t^\phi \mathrm{d}X_t$ are sub-exponential.

Algorithm 1: PEGE Algorithm

Input: $\mathfrak{m} : \mathbb{N} \to \mathbb{N}$.

```
Initialize m = 0.
for k = 1, 2, ... do
    Execute the exploration policy φ^e for one episode, and m ← m + 1.
    Update the estimate θ̂^(m) and set θ̄ = θ̂^(m).
    for l = 1, 2, ..., 𝔪(k) do
        Execute the greedy policy φ_θ̄ for one episode, and m ← m + 1.
    end
end
```

- $\mathfrak{m} : \mathbb{N} \to \mathbb{N}$ determines the exploration and exploitation trade-off.

Let $\mathcal{E}^\Phi = \{m \in \mathbb{N} \mid \phi^{(m)} = \phi^e\}$ and consider

$$
\begin{aligned}
\mathcal{R}(N) &= \sum_{m=1}^{N} \left(J(\phi^{(m)}; \theta^\star) - J(\phi_{\theta^\star}; \theta^\star) \right) \\
&= \sum_{m \in [1,N] \cap \mathcal{E}^\Phi} \left(J(\phi^e, \theta^\star) - J(\phi_\theta; \theta^\star) \right) + \sum_{m \in [1,N] \setminus \mathcal{E}^\Phi} \left(J(\phi_{\hat{\theta}(m-1)}, \theta^\star) - J(\phi_{\theta^\star}; \theta^\star) \right)
\end{aligned}
$$

Let $\mathcal{E}^\Phi = \{m \in \mathbb{N} \mid \phi^{(m)} = \phi^e\}$ and consider

$$\begin{aligned}\mathcal{R}(N) &= \sum_{m=1}^{N} \left(J(\phi^{(m)}; \theta^\star) - J(\phi_{\theta^\star}; \theta^\star)\right) \\ &= \sum_{m\in[1,N]\cap\mathcal{E}^\Phi} \left(J(\phi^e, \theta^\star) - J(\phi_\theta; \theta^\star)\right) + \sum_{m\in[1,N]\setminus\mathcal{E}^\Phi} \left(J(\phi_{\hat\theta(m-1)}, \theta^\star) - J(\phi_{\theta^\star}; \theta^\star)\right)\end{aligned}$$

Theorem (Performance gap)

For all $\varepsilon > 0$, $\exists C_\varepsilon \geq 0$,

$$|J(\phi_\theta; \theta^\star) - J(\phi_{\theta^\star}; \theta^\star)| \leq C_\varepsilon |\theta - \theta^\star|^2, \quad \forall \theta \in \mathbb{B}_\varepsilon(\theta^\star).$$

Let $\mathcal{E}^\Phi = \{m \in \mathbb{N} \mid \phi^{(m)} = \phi^e\}$ and consider

$$
\begin{aligned}
\mathcal{R}(N) &= \sum_{m=1}^{N} \left(J(\phi^{(m)}; \theta^\star) - J(\phi_{\theta^\star}; \theta^\star) \right) \\
&= \sum_{m \in [1,N] \cap \mathcal{E}^\Phi} \left(J(\phi^e, \theta^\star) - J(\phi_\theta; \theta^\star) \right) + \sum_{m \in [1,N] \setminus \mathcal{E}^\Phi} \left(J(\phi_{\hat{\theta}(m-1)}, \theta^\star) - J(\phi_{\theta^\star}; \theta^\star) \right)
\end{aligned}
$$

Theorem (Performance gap)

For all $\varepsilon > 0$, $\exists C_\varepsilon \geq 0$,

$$|J(\phi_\theta; \theta^\star) - J(\phi_{\theta^\star}; \theta^\star)| \leq C_\varepsilon |\theta - \theta^\star|^2, \quad \forall \theta \in \mathbb{B}_\varepsilon(\theta^\star).$$

This gives $\mathcal{R}(N) \lesssim \left| [1, N] \cap \mathcal{E}^\Phi \right| + \sum_{m \in [1,N] \setminus \mathcal{E}^\Phi} |\hat{\theta}^{(m-1)} - \theta^\star|^2$.

Theorem

For $\mathfrak{m}(k) = k$, $k \in \mathbb{N}$, with high probability,

$$\mathcal{R}(N) \le C N^{\frac{1}{2}} (\log N)^2, \quad \forall N \ge 2.$$

Optimal regret for LQ-RL

Theorem

For $\mathfrak{m}(k) = k$, $k \in \mathbb{N}$, *with high probability,*

$$\mathcal{R}(N) \leq CN^{\frac{1}{2}}(\log N)^2, \quad \forall N \geq 2.$$

If self-exploration property holds, then by setting $\mathfrak{m}(k) = 2^k$, $k \in \mathbb{N}$, *with high probability,*

$$\mathcal{R}(N) \leq C\big(\log N\big)^2, \quad \forall N \geq 2.$$

Theorem

For $\mathfrak{m}(k) = k$, $k \in \mathbb{N}$, *with high probability,*

$$\mathcal{R}(N) \le CN^{\frac{1}{2}}(\log N)^2, \quad \forall N \ge 2.$$

If self-exploration property holds, then by setting $\mathfrak{m}(k) = 2^k$, $k \in \mathbb{N}$, *with high probability,*

$$\mathcal{R}(N) \le C\big(\log N\big)^2, \quad \forall N \ge 2.$$

- Discrete time observations and actions have the same regret order, with an additional discretization error.

Let $\theta^\star = (A^\star, B^\star) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times k}$ be fixed but unknown, minimise

$$J(\alpha; \theta^\star) = \mathbb{E}\left[\int_0^T f_t(X_t^{\theta^\star,\alpha}, \alpha_t)\,\mathrm{d}t + g(X_T^{\theta^\star,\alpha})\right],$$

over stochastic processes α, where $X^{\theta^\star,\alpha}$ satisfies the dynamics with $\theta^\star$:

$$\mathrm{d}X_t = (A^\star X_t + B^\star \alpha_t)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t + \int_{\mathbb{R}^p\setminus\{0\}} \gamma(u)\,\widetilde{N}(\mathrm{d}t, \mathrm{d}u),$$

f and g are convex in state and strongly convex in control.

- f can be nonsmooth, and includes action constraints, ℓ^1-norm or entropy regularisers.

	LQ-RL	LC-RL
Greedy policy	Linear	Nonlinear
Policy characterisation	Riccati ODE	BSDE
Performance gap	Quadratic	Linear/quadratic
Parameter estimation	Bayesian	Least-squares
Estimation error	Sub-exponential r.v.	Sub-Weibull r.v.

- Consider the 3d controlled SDE:

$$\mathrm{d}X_t = (AX_t + B\alpha_t)\,\mathrm{d}t + \mathrm{d}W_t,\ \ t \in [0, 1.5].$$

with unknowns A, B from Dean et al. 2018, and a given cost

$$J(\alpha) = \mathbb{E}\left[\int_0^T (0.1|X_t^\alpha|^2 + |\alpha_t|^2)\,\mathrm{d}t\right].$$

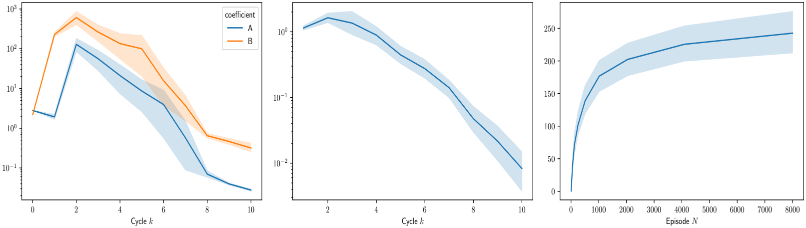
- Consider the 3d controlled SDE:

$$\mathrm{d}X_t = (AX_t + B\alpha_t)\,\mathrm{d}t + \mathrm{d}W_t,\ \ t \in [0, 1.5].$$

with unknowns A, B from Dean et al. 2018, and a given cost

$$J(\alpha) = \mathbb{E}\left[\int_0^T (0.1|X_t^\alpha|^2 + |\alpha_t|^2)\,\mathrm{d}t\right].$$

- Run PEGE algorithm with $\mathfrak{m}(k) = 2^k$, $k \in \mathbb{N}$.
- Perform 100 independent executions to estimate statistical properties of the algorithm.



(a) Relative param. error (b) Relative cost error (c) Regret

Figure: Numerical results from 100 repeated experiments; solid lines are sample means and shallow areas are 95% confidence intervals.

Summary

Two complimentary aspects on model-based RL:

- Finite-sample analysis of parameter estimation (statistical learning theory) and performance gap analysis of greedy policy (control theory).
- A phase-based learning algorithm with optimal regrets for linear-convex models.

Summary

Two complimentary aspects on model-based RL:

- Finite-sample analysis of parameter estimation (statistical learning theory) and performance gap analysis of greedy policy (control theory).
- A phase-based learning algorithm with optimal regrets for linear-convex models.

(1) Basei, Guo, Hu, Zhang, *Logarithmic regret for episodic continuous-time linear-quadratic reinforcement learning over a finite-time horizon*, JMLR, to appear, 2020.

(2) Szpruch, Treetanthiploet and Zhang, *Exploration-exploitation trade-off for continuous-time episodic reinforcement learning with linear-convex models*, arXiv preprint, 2021.