Exploration-exploitation trade-off for reinforcement learning with continuous-time models

YUFEI ZHANG

Department of Statistics, London School of Economics

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

Financial/Actuarial Mathematics Seminar
16 March 2022
Fix $\theta = (A, B) \in \mathbb{R}^{d \times d} \times \mathbb{R}^{d \times p}$, and minimise

$$J(\alpha; \theta) = \mathbb{E} \left[ \int_0^T f(t, X_{t \theta}^{\alpha}, \alpha_t) \, dt + g(X_T^{\theta, \alpha}) \right],$$

over $\mathbb{R}^p$-valued adapted control processes $\alpha$, where $X^{\theta, \alpha}$ satisfies the following dynamics:

$$dX_t = (AX_t + B\alpha_t) \, dt + dW_t, \quad t \in [0, T]; \quad X_0 = x_0,$$

and $f$ and $g$ are given functions that are convex in state and strongly convex in control (with suitable regularity).
Feedback control

- When $\theta$ is known, the LC control problem admits a unique optimal control $\alpha^\theta$, given in a feedback form:

$$\alpha^\theta_t = \phi_\theta(t, X^\theta_t), \quad \text{for } dP \otimes dt \text{ a.e.},$$

and $\phi_\theta$ is in bounded in time and Lipschitz in space.

- When $\theta$ is unknown, one needs to balance exploration (learning via interactions with the random environment) and exploitation (optimal control).
Learning via trial and error

Episode 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

$$dX^m_t = (AX^m_t + B\phi^{(m)}(t, X^m_t))dt + dW^m_t, \quad t \in [0, T], \quad X^m_0 = x_0.$$ 

The expected cost for the $m$-th episode is

$$J(\phi^{(m)}; \theta) = \mathbb{E} \left[ \int_0^T f(t, X^m_t, \phi^{(m)}(t, X^m_t)) \, dt + g(X^m_T) \right].$$

- Agent constructs $\hat{\theta}^{(m)}$ using observed trajectories of $(X^i_i)_{i=1}^m$. 
Let $\Phi = (\phi^{(1)}, \phi^{(2)}, \ldots)$ be a learning algorithm. The regret of learning with $N \in \mathbb{N}$ episodes is

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

where

- $J(\phi^{(m)}; \theta)$ is the cost for the $m$-th episode,
- $J(\phi_{\theta}; \theta)$ is the optimal cost as if the parameter $\theta$ were known.

Construct $\Phi$ whose regret $R(N, \Phi)$ grows as slow as possible in terms of $N$. 

Do we need to explore?

Incomplete learning

Let \((B_1, B_2) \neq (0, 0)\), and consider the 1d SDE:

\[
dX_t = (B_1 \alpha_{1,t} + B_2 \alpha_{2,t})dt + dW_t, \quad t \in [0, T], \quad X_0 = x_0,
\]

and the cost \(J(\alpha; \theta) = \mathbb{E} \left[ \int_0^T (\alpha_{1,t}^2 + \alpha_{2,t}^2)dt + X_T^2 \right] \).

- The optimal policy is \(\phi(t, x) = -p_t Bx\), where \(B = (B_1, B_2)\top\), \(p : [0, T] \rightarrow \mathbb{R}\) solves a Riccati equation.
- Assume agent learns by only executing optimal policies.
- If \((\hat{B}_1^{(0)}, 0), \hat{B}_1^{(0)} \neq 0\), then agent executes \((\alpha_{1,t}, 0)\) and only learns about \(B_1\) in the next episode,
- and if it happens that for all \(m \in \mathbb{N}\), \((\hat{B}_1^{(m)}, 0), \hat{B}_1^{(m)} \neq 0\), the optimal model and the optimal policy will never be learned.
Assumption (Exploration policy)

There exists a policy $\phi^e$ such that

- $J(\phi^e; \theta) < \infty$, and
- 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.

- $\phi^e$ exists if and only if the action set linearly spans $\mathbb{R}^p$.
- In this case, one can choose linearly independent actions $(a_i)_{i=1}^p$, and set

$$\phi^e(t, x) = a_k, \quad (t, x) \in [t_{k-1}, t_k) \times \mathbb{R}^d, \quad k = 1, \ldots, p,$$

for a time grid $\{0 = t_0 < t_1 < \ldots < t_p = T\}$. 
Probabilistic learning

- View unknown parameters as a hidden random variable $\theta = (A, B)$.
- For chosen policy $\phi$, estimate $\theta = (A, B)$ via
  \[ dX_t = \theta Z_t^\phi dt + dW_t, \quad Z_t^\phi = (X_t, \phi(t, X_t))^\top. \]
- Agent only observes the state process $X$ (also $Z^\phi$), but not the corresponding Brownian path.
- Otherwise, choosing $\phi = \phi^e$,
  \[ \theta = \left( \int_0^T Z_s^{\phi^e} (dX_s - dW_s)^\top \right)^\top \left( \int_0^T Z_s^{\phi^e} (Z_s^{\phi^e})^\top ds \right)^{-1} \]
  where $\int_0^T Z_s^{\phi^e} (Z_s^{\phi^e})^\top ds$ is almost surely invertible.
Bayesian perspective

Likelihood function

Given fixed policy \( \phi \), estimate \( \theta = (A, B) \) via

\[
dX_t = \theta Z_t^\phi dt + dW_t, \quad Z_t^\phi = (X_t, \phi(t, X_t))^\top.
\]

Likelihood function of \( \theta \) with observation \( X \):

\[
\ell(\theta | X) \propto \exp \left( -\frac{1}{2} \theta \left( \int_0^t (Z_s^\phi)(Z_s^\phi)^\top ds \right) \theta^\top + \theta \int_0^t Z_s^\phi dX_s \right).
\]
Given prior $\pi_0(\theta) = \mathcal{N}(\hat{\theta}_0, V_0)$, the posterior of $\theta$ with observation $X$ is given by

$\pi(\theta | \mathcal{F}_T^X, \phi) \propto \ell(\theta | X)\pi_0(\theta)$

$\propto \exp \left( - \frac{1}{2} \theta \left( V_0^{-1} + \int_0^T (Z_s^\phi)(Z_s^\phi)^\top ds \right) \theta^\top + \theta \left( V_0^{-1} \hat{\theta}_0^\top + \int_0^T (Z_s^\phi)dX_s \right) \right)$.

The posterior $\pi(\theta | \mathcal{F}_T^X, \phi) = \mathcal{N}(\hat{\theta}, V)$, where

$\hat{\theta} = \mathbb{E}[\theta | \mathcal{F}_T^X, \phi] = \left( V_0^{-1} \hat{\theta}_0^\top + \int_0^T (Z_s^\phi)dX_s \right) \left( V_0^{-1} + \int_0^T (Z_s^\phi)(Z_s^\phi)^\top ds \right)^{-1}$,

$V = \text{Var}[\theta | \mathcal{F}_T^X, \phi] = \left( V_0^{-1} + \int_0^T (Z_s^\phi)(Z_s^\phi)^\top ds \right)^{-1}$. 
Algorithm 1: PEGE Algorithm

**Input:** \( m : \mathbb{N} \rightarrow \mathbb{N} \).

1. Initialize \( m = 0 \).
2. for \( k = 1, 2, \ldots \) do
   3. Execute the exploration policy \( \phi^e \) for one episode, and \( m \leftarrow m + 1 \).
   4. Update the estimate \( \hat{\theta}_m \) and set \( \bar{\theta} = \hat{\theta}_m \).
   5. for \( l = 1, 2, \ldots, m(k) \) do
      6. Execute the greedy policy \( \phi_{\bar{\theta}} \) for one episode, and \( m \leftarrow m + 1 \).
   7. end
3. end

- \( \hat{\theta}_m \) is updated based on all previous observations.
- \( m : \mathbb{N} \rightarrow \mathbb{N} \) is chosen to optimise algorithm regret order.
Let $\mathcal{E}^\Phi = \{ m \in \mathbb{N} \mid \phi^m = \phi^e \}$ and consider

$$
R(N, \Phi, \theta) = \sum_{m=1}^{N} \left( J(\phi^m; \theta) - J(\phi^e; \theta) \right)
$$

$$
= \sum_{m \in [1, N] \cap \mathcal{E}^\Phi} \left( J(\phi^e, \theta) - J(\phi^\theta, \theta) \right)
+ \sum_{m \in [1, N] \setminus \mathcal{E}^\Phi} \left( J(\phi^\theta, \theta) - J(\phi^\theta; \theta) \right)

\leq \left( |J(\phi^e, \theta)| + |J(\phi^\theta, \theta)| \right) [1, N] \cap \mathcal{E}^\Phi
+ \sum_{m \in [1, N] \setminus \mathcal{E}^\Phi} \left( J(\phi^\theta, \theta) - J(\phi^\theta; \theta) \right).
$$
Assumption (Performance gap)

Let $\theta$ take values in $\Theta$, and $\exists L_{\Theta}, \beta > 0, r \in (0, 1]$ s.t. $\forall \theta_0 \in \Theta$,

$$|J(\phi_{\theta}; \theta_0) - J(\phi_{\theta_0}; \theta_0)| \leq L_{\Theta}|\theta - \theta_0|^{2r}, \quad \forall \theta \in B_{\beta}(\theta_0),$$

where $\phi_{\theta}$ is an optimal policy with parameter $\theta$.

- The value $r$ quantifies the impact of model misspecification on learning.
Linear performance gap
Nonsmooth running cost

Theorem

Assume \( f(t, x, a) = f_0(t, x, a) + h(a) \), where

- \( f_0 : [0, T] \times \mathbb{R}^d \times \mathbb{R}^p \rightarrow \mathbb{R} \) has Lipschitz continuous derivatives, and
- \( h : \mathbb{R}^p \rightarrow \mathbb{R} \cup \{\infty\} \) is lower semicontinuous and convex.

Then the performance gap holds with \( r = 1/2 \).

- \( h \) includes (convex) indicator function, \( \ell_1 \)-norm and \( f \)-divergence.
- First prove that \( \exists C \geq 0 \) s.t. \( \forall \theta, \theta_0 \in \Theta \),

\[
|\phi_\theta(t, x) - \phi_{\theta_0}(t, x)| \leq C(1 + |x|)|\theta - \theta_0|, \quad \forall (t, x) \in [0, T] \times \mathbb{R}^d.
\]

- Prove \( |J(\phi_\theta; \theta_0) - J(\phi_{\theta_0}; \theta_0)| \leq C\|\alpha^{\phi_\theta} - \alpha^{\phi_{\theta_0}}\|_{\mathcal{H}^2} \).
Theorem

Assume $f$ has Lipschitz continuous derivatives. Then the performance gap holds with $r = 1$.

- First prove the functional $\mathcal{H}^2(\Omega; \mathbb{R}^p) \ni \alpha \mapsto J(\alpha; \theta_0) \in \mathbb{R}$ has a Lipschitz continuous (Fréchet) derivative.
- Conclude that for all $\alpha \in \mathcal{H}^2(\Omega; \mathbb{R}^p)$,

$$J(\alpha; \theta_0) - J(\alpha^{\theta_0}; \theta_0) \leq \langle \nabla_\alpha J(\alpha; \theta_0) \big|_{\alpha = \alpha^{\theta_0}}, \alpha - \alpha^{\theta_0} \rangle_{\mathcal{H}^2} + C \|\alpha - \alpha^{\theta_0}\|_{\mathcal{H}^2}^2.$$
- Entropy-regularized relaxed control problem.

**Theorem**

Assume \( f \) is of the form \( f(t, x, a) = f_0(t, x) + h_{en}(a) \), with

\[
h_{en}(a) = \begin{cases} 
\sum_{i=1}^{p} a_i \ln(a_i), & \text{if } \sum_{i=1}^{p} a_i = 1, \ a_i \geq 0, \ \forall i, \\
\infty, & \text{otherwise},
\end{cases}
\]

and \( f_0(t, \cdot) \in C_b^4(\mathbb{R}^d) \) and \( g \in C_b^4(\mathbb{R}^d) \) uniformly in \( t \).

Then the performance gap holds with \( r = 1 \).

- Prove that \( \mathbb{R}^{d \times (d+p)} \ni \theta \mapsto J(\phi_\theta; \theta_0) \in \mathbb{R} \) is \( C^2 \), and a second-order Taylor expansion around the minimizer \( \theta_0 \).
Regret Analysis

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

$$
\mathcal{R}(N, \Phi, \theta) = \sum_{m=1}^{N} \left( J(\phi(m); \theta) - J(\phi^e; \theta) \right) \\
\leq \left( \left| J(\phi^e, \theta) \right| + \left| J(\phi^e; \theta) \right| \right) \left| [1, N] \cap \mathcal{E}^\Phi \right| + \sum_{m \in [1, N] \setminus \mathcal{E}^\Phi} \left( J(\phi_{\hat{\theta}_{m-1}}, \theta) - J(\phi^e; \theta) \right) \\
\lesssim \left| [1, N] \cap \mathcal{E}^\Phi \right| + \sum_{m \in [1, N] \setminus \mathcal{E}^\Phi} |\hat{\theta}_{m-1} - \theta|^{2r}.
$$

- Quantify parameter estimation error in terms of episode numbers.
Recall after the $m$-th episode, given the chosen policies $\Phi = (\phi^{(n)})_{n=1}^{m}$, posterior distribution of $\theta$ is $\mathcal{N}(\hat{\theta}_m, V_{m}^{\Phi})$, where

$$V_{m}^{\Phi} := \left(V_{0}^{-1} + \sum_{n=1}^{m} \int_{0}^{T} Z_{s,n}^{\Phi}(Z_{s,n}^{\Phi})^\top ds\right)^{-1},$$

$$\hat{\theta}_m := \left(\hat{\theta}_{0} V_{0}^{-1} + \sum_{n=1}^{m} \left(\int_{0}^{T} Z_{s,n}^{\Phi}(dX_{s,n}^{\Phi})^\top\right)^\top\right) V_{m}^{\Phi},$$

where $\mathcal{N}(\hat{\theta}_{0}, V_{0})$ is the initial prior, and $Z_{t,n}^{\Phi} = \left(\chi_{t,\phi^{(n)}}^{X_{t,n}^{\Phi}}\right)$ for all $n = 1, \ldots, m$. 
Theorem

If $\Phi = (\phi^{(n)})_{n \in \mathbb{N}}$ is uniformly Lipschitz in space, with high prob.,

$$|\hat{\theta}_m - \theta|^2 \lesssim \frac{\ln m}{\lambda_{\min}((V_m^\Phi)^{-1})}, \quad \forall m \geq 2,$$

where $\lambda_{\min}(S)$ is the smallest eigenvalue of a symmetric matrix $S$.

As $(V_m^\Phi)^{-1} = V_0^{-1} + \sum_{n=1}^{m} \int_0^T Z_s^{\Phi,n}(Z_s^{\Phi,n})^\top ds$,

$$\lambda_{\min}((V_m^\Phi)^{-1}) \geq \sum_{m \in [1, m] \cap \mathcal{E}^\Phi} \lambda_{\min} \left( \int_0^T Z_s^{\Phi,n}(Z_s^{\Phi,n})^\top ds \right) \gtrsim |[1, m] \cap \mathcal{E}^\Phi|, \quad \text{with high probability},$$

where $\mathcal{E}^\Phi = \{m \in \mathbb{N}|\phi^{(m)} = \phi^e\}$.
Optimal regret

General case

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

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

$$ \lesssim \left| [1, N] \cap \mathcal{E}^\Phi \right| + \sum_{m \in [1, N] \setminus \mathcal{E}^\Phi} \left( \frac{\ln m}{\left| [1, m] \cap \mathcal{E}^\Phi \right|} \right)^r $$

**Theorem**

For $m(k) = \lfloor k^r \rfloor$, $k \in \mathbb{N}$, with high probability,

$$ R(N, \Phi^{\text{PGE}}, \theta) \lesssim N^{\frac{1}{1+r}} (\log N)^r, \quad \forall N \geq 2. $$

- Extend $O(\sqrt{N})$-regrets for bandits or discrete-time MDPs.
Observations from exploitation episodes can also improve parameter estimation.

**Definition**

The **self-exploration property** holds if \( \exists \eta > 0 \) such that for all \( \theta \in \mathbb{B}_\eta(\text{Range}(\theta)) \), \( \phi_\theta \) is an exploration policy, i.e.,

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

- For LQ problems, self-exploration is equivalent to full column rank of the control coefficient \( B \); see M Basei, X Guo, A Hu, Y Zhang, 2021.
Theorem

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

$$R(N, \Phi^{PEGE}, \theta) \leq N^{1-r} (\log N)^r, \quad \forall N \geq 2.$$
Consider the 3D controlled SDE (Dean et al. 2018):

\[
dX_t = (AX_t + B\alpha_t) \, dt + dW_t, \quad t \in [0, 1.5].
\]

with unknown \( A = \begin{bmatrix} 1.01 & 0.01 & 0 \\ 0.01 & 1.01 & 0.01 \\ 0 & 0.01 & 1.01 \end{bmatrix} \) and \( B = I_3 \), and a given cost

\[
J(\alpha) = E \left[ \int_0^T (0.1|X_t^\alpha|^2 + |\alpha_t|^2) \, dt \right].
\]

Sample \((A^{(0)}, B^{(0)})\) from the standard normal distribution, and run PEGE algorithm with \( m(k) = 2^k \), \( k \in \mathbb{N} \).

Perform 100 independent executions to estimate statistical properties of the algorithm.
Figure: Numerical results from 100 repeated experiments; solid lines are sample means and shallow areas are 95% confidence intervals.
Two complimentary aspects on model-based RL:

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


Convexity of cost functions

Assumption (H1)

(1) \( \exists f_0 : [0, T] \times \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R} \) and \( h : \mathbb{R}^p \to \mathbb{R} \cup \{\infty\} \) s.t.

\[
f(t, x, a) = f_0(t, x, a) + h(a)
\]

and for all \( t \in [0, T], \)

- \( f_0(t, \cdot, \cdot) \) has Lipschitz continuous derivative, \( h \) is lower semicontinuous and convex,
- \( f(t, \cdot, \cdot) \) is convex in \( x \), strongly convex in \( a \).

(2) \( g \) is convex and has Lipschitz continuous derivative.