Exploration vs Exploitation in Reinforcement Learning: Dilemma of the Controller in an Uncertain World

Tanut (Nash) Treetanthiploet[⋆]

Joint work with Lukasz Szpruch[⋆,†] and Yufei Zhang[•]

[⋆]The Alan Turing Institute, London, [†]The University of Edinburgh, Edinburgh
[•]The London School of Economics and Political Science, London

The 9th International Colloquium on BSDEs and Mean Field Systems,

June-July 2022

Stochastic Control Problem with Linear Dynamics

Let $\theta^\star = (A^\star, B^\star) \in \mathbb{R}^{d\times d} \times \mathbb{R}^{p\times d}$ and consider the control problem

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

where

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

Stochastic Control Problem with Linear Dynamics

Let $\theta^\star = (A^\star, B^\star) \in \mathbb{R}^{d\times d} \times \mathbb{R}^{p\times d}$ and consider the control problem

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

where

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

[Guo, Hu and Zhang, 2021] There exists $\phi^{\theta^\star} : [0,T] \times \mathbb{R}^d \to \mathbb{R}^p$ such that

$$\alpha_t^\star := \phi^{\theta^\star}(t, X_t^{\theta^\star,\alpha^\star}) = \underset{\alpha\in\mathcal{H}^2_{\mathbb{F}}(\Omega;\mathbb{R}^p)}{\arg\min}\; J(\alpha;\theta^\star).$$

Stochastic Control Problem with Linear Dynamics

Let $\theta^\star = (A^\star, B^\star) \in \mathbb{R}^{d\times d} \times \mathbb{R}^{p\times d}$ and consider the control problem

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

where

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

[Guo, Hu and Zhang, 2021] There exists $\phi^{\theta^\star} : [0,T]\times\mathbb{R}^d \to \mathbb{R}^p$ such that

$$\alpha_t^\star := \phi^{\theta^\star}(t, X_t^{\theta^\star,\alpha^\star}) = \underset{\alpha\in\mathcal{H}^2_{\mathbb{F}}(\Omega;\mathbb{R}^p)}{\arg\min}\ J(\alpha;\theta^\star).$$

We do not know $\theta^\star$ and thus cannot find $\phi^{\theta^\star}$.

Episodic Learning problem

Let $\varphi_m : \Omega \times [0, T] \times \mathbb{R}^d \to \mathbb{R}^p$ be a sequence of random (feedback) function that the agent executes for each episode.

- At the end of the m-th episode, the agent observes $(X_t^m)_{t\in[0,T]}$;

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

  and experience the (expected) cost

$$J(\varphi_m; \theta^\star) := \mathbb{E}^{W^m}\left[\int_0^T f(t, X_t^m, \varphi_m(\cdot, t, X_t^m))\,\mathrm{d}t + g(X_T^m)\right].$$

- Design φ_{m+1} from the previous observations, $(X^n)_{n=1}^m$.

Episodic Learning problem

Let $\varphi_m : \Omega \times [0, T] \times \mathbb{R}^d \to \mathbb{R}^p$ be a sequence of random (feedback) function that the agent executes for each episode.

- At the end of the m-th episode, the agent observes $(X_t^m)_{t\in[0,T]}$;

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

  and experience the (expected) cost

$$J(\varphi_m; \theta^\star) := \mathbb{E}^{W^m}\left[\int_0^T f(t, X_t^m, \varphi_m(\cdot, t, X_t^m))\,\mathrm{d}t + g(X_T^m)\right].$$

- Design φ_{m+1} from the previous observations, $(X^n)_{n=1}^m$.

The agent objective is to minimise

$$\mathrm{Reg}(N) = \sum_{m=1}^N \left(J(\varphi_m; \theta^\star) - J(\phi^{\theta^\star}; \theta^\star)\right).$$

Statistical Estimate from Bayesian inference

Suppose that before the m-th episode, the posterior of $\theta^\star$ is $N(\hat{\theta}_{m-1}, V_{m-1})$.
At the m-th episode, the agent observes (X_t^m) satisfying

$$\mathrm{d}X_t^m = \theta^\star Z_t^m \,\mathrm{d}t + \mathrm{d}W_t^m, \quad X_0^m = x_0, \text{ with } Z_t^m = \begin{pmatrix} X_t^m \\ \varphi_m(\cdot, t, X_t^m) \end{pmatrix}.$$

Statistical Estimate from Bayesian inference

Suppose that before the m-th episode, the posterior of $\theta^\star$ is $N(\hat{\theta}_{m-1}, V_{m-1})$.
At the m-th episode, the agent observes (X_t^m) satisfying

$$\mathrm{d}X_t^m = \theta^\star Z_t^m \,\mathrm{d}t + \mathrm{d}W_t^m, \quad X_0^m = x_0, \text{ with } Z_t^m = \begin{pmatrix} X_t^m \\ \varphi_m(\cdot, t, X_t^m) \end{pmatrix}.$$

Let consider a discretisation $(\Delta X_{t_1}^m, \Delta X_{t_2}^m, ..., \Delta X_{t_K}^m)$ where

$$\Delta X_{t_k}^m = \theta^\star Z_{t_k}^m \Delta t + \Delta W_{t_k} \sim N\big(\theta^\star Z_{t_k}^m \Delta t, \Delta t\big)$$

Statistical Estimate from Bayesian inference

Suppose that before the m-th episode, the posterior of $\theta^\star$ is $N(\hat{\theta}_{m-1}, V_{m-1})$. At the m-th episode, the agent observes (X_t^m) satisfying

$$\mathrm{d}X_t^m = \theta^\star Z_t^m \,\mathrm{d}t + \mathrm{d}W_t^m, \quad X_0^m = x_0, \text{ with } Z_t^m = \begin{pmatrix} X_t^m \\ \varphi_m(\cdot, t, X_t^m) \end{pmatrix}.$$

Let consider a discretisation $(\Delta X_{t_1}^m, \Delta X_{t_2}^m, ..., \Delta X_{t_K}^m)$ where

$$\Delta X_{t_k}^m = \theta^\star Z_{t_k}^m \Delta t + \Delta W_{t_k} \sim N\big(\theta^\star Z_{t_k}^m \Delta t, \Delta t\big)$$

Therefore,

$$\begin{aligned}
&\pi\left(\theta^\star \middle| \mathcal{F}_{m-1}, (\Delta X_{t_k}^m)_{k=1}^K, (Z_{t_k}^m)_{k=1}^K\right) \\
&\quad\propto \exp\left(-\frac{1}{2}(\theta^\star - \hat{\theta}_{m-1}) V_{m-1}^{-1} (\theta^\star - \hat{\theta}_{m-1})^\top\right) \prod_{k=1}^K \exp\left(-\frac{1}{2\Delta t}(\Delta X_{t_k}^m - \theta^\star Z_{t_k}^m \Delta t)^2\right) \\
&\quad\propto \exp\left(-\frac{1}{2}\theta^\star \left(V_{m-1}^{-1} + \sum_{k=1}^K Z_{t_k}^m (Z_{t_k}^m)^\top \Delta t\right) \theta^{\star\top} + \theta^\star \left(V_{m-1}^{-1}\hat{\theta}_{m-1}^\top + \sum_{k=1}^K Z_{t_k}^m \Delta X_{t_k}^m\right)\right).
\end{aligned}$$

Statistical Estimate from Bayesian inference

In particular, if we send $\Delta t \to 0$, the posterior is

$$\pi\left(\theta^\star \middle| \mathcal{F}_{m-1}, (X_t^m)_{t\in[0,T]}, (Z_t^m)_{t\in[0,T]}\right)$$
$$\propto \exp\left(-\frac{1}{2}\theta^\star\left(V_{m-1}^{-1} + \int_0^T Z_t^m (Z_t^m)^\top \mathrm{d}t\right)\theta^{\star\top} + \theta^\star\left(V_{m-1}^{-1}\hat{\theta}_{m-1}^\top + \int_0^T Z_t^m \mathrm{d}X_t^m\right)\right).$$

Statistical Estimate from Bayesian inference

In particular, if we send $\Delta t \to 0$, the posterior is

$$\pi\left(\theta^\star \middle| \mathcal{F}_{m-1}, (X_t^m)_{t\in[0,T]}, (Z_t^m)_{t\in[0,T]}\right)$$
$$\propto \exp\left(-\frac{1}{2}\theta^\star\left(V_{m-1}^{-1} + \int_0^T Z_t^m (Z_t^m)^\top \mathrm{d}t\right)\theta^{\star\top} + \theta^\star\left(V_{m-1}^{-1}\hat{\theta}_{m-1}^\top + \int_0^T Z_t^m \mathrm{d}X_t^m\right)\right).$$

In particular, the posterior of $\theta^\star$ after the m-th episode is $N(\hat{\theta}_m, V_m)$ where

$$V_m^{-1} = V_0^{-1} + \sum_{n=1}^m \int_0^T Z_t^n (Z_t^n)^\top \mathrm{d}t, \quad \text{and} \quad \hat{\theta}_m = \left(\hat{\theta}_0 V_0^{-1} + \sum_{n=1}^m \int_0^T Z_t^n \mathrm{d}X_t^n\right) V_m.$$

Statistical Estimate from Bayesian inference

In particular, if we send $\Delta t \to 0$, the posterior is

$$\pi\left(\theta^\star \middle| \mathcal{F}_{m-1}, (X_t^m)_{t\in[0,T]}, (Z_t^m)_{t\in[0,T]}\right)$$
$$\propto \exp\left(-\frac{1}{2}\theta^\star\left(V_{m-1}^{-1} + \int_0^T Z_t^m (Z_t^m)^\top \mathrm{d}t\right)\theta^{\star\top} + \theta^\star\left(V_{m-1}^{-1}\hat{\theta}_{m-1}^\top + \int_0^T Z_t^m \mathrm{d}X_t^m\right)\right).$$

In particular, the posterior of $\theta^\star$ after the m-th episode is $N(\hat{\theta}_m, V_m)$ where

$$V_m^{-1} = V_0^{-1} + \sum_{n=1}^m \int_0^T Z_t^n (Z_t^n)^\top \mathrm{d}t, \quad \text{and} \quad \hat{\theta}_m = \left(\hat{\theta}_0 V_0^{-1} + \sum_{n=1}^m \int_0^T Z_t^n \mathrm{d}X_t^n\right) V_m.$$

In comparison to the classical statistics theory, one may see

- $\hat{\theta}_m$ as a (regularised) maximum likelihood estimator.
- V_m^{-1} as a (regularised) Fisher Information.

Sub-optimality of the Greedy policy

Consider the case when $\theta = (B_1, B_2)$ and

$$J(\alpha;\theta) = \mathbb{E}\left[\int_0^T (\alpha_{1,t}^2 + \alpha_{2,t}^2)\,\mathrm{d}t + (X_T^{\theta,\alpha})^2\right],$$

where

$$\mathrm{d}X_t^{\theta,\alpha} = (B_1\alpha_{1,t} + B_2\alpha_{2,t})\,\mathrm{d}t + \mathrm{d}W_t, \quad X_0^{\theta,\alpha} = x_0.$$

The optimal policy is

$$\phi^\theta(t,x) = -\left(1 + (B_1^2 + B_2^2)(T-t)\right)^{-1}\begin{pmatrix} B_1 x \\ B_2 x \end{pmatrix}.$$

Sub-optimality of the Greedy policy

Consider the case when $\theta = (B_1, B_2)$ and

$$J(\alpha;\theta) = \mathbb{E}\left[\int_0^T (\alpha_{1,t}^2 + \alpha_{2,t}^2)\,\mathrm{d}t + (X_T^{\theta,\alpha})^2\right],$$

where

$$\mathrm{d}X_t^{\theta,\alpha} = (B_1\alpha_{1,t} + B_2\alpha_{2,t})\,\mathrm{d}t + \mathrm{d}W_t, \quad X_0^{\theta,\alpha} = x_0.$$

The optimal policy is

$$\phi^\theta(t,x) = -\left(1 + (B_1^2 + B_2^2)(T-t)\right)^{-1}\begin{pmatrix} B_1 x \\ B_2 x \end{pmatrix}.$$

$$\hat{\theta}_m = (\hat{B}_{1,m}, 0) \;\Rightarrow\; \phi^{\hat{\theta}_m}(t,x) = \begin{pmatrix} K_t^m \\ 0 \end{pmatrix} x \;\Rightarrow\; \hat{\theta}_{m+1} = (\hat{B}_{1,m+1}, 0)$$

- Decisions are always sub-optimal provided that $B_2^\star \neq 0$.

Sensitivity in Parameter Estimate

Performance Gap

Let $\theta = (A, B)$ and $J(\phi;\theta) := \mathbb{E}\left[\int_0^T f(t, X_t^{\theta,\phi}, \phi(t, X_t^{\theta,\phi}))\,\mathrm{d}t + g(X_T^{\theta,\phi})\right]$, where

$$\mathrm{d}X_t^{\theta,\phi} = (AX_t^{\theta,\phi} + B\phi(t, X_t))\,\mathrm{d}t + \mathrm{d}W_t, \quad X_0^{\theta^\star,\phi} = x_0.$$

Define $\phi^\theta := \arg\min_\phi J(\phi;\theta)$. Then for a strong convex cost f and g,

$$J(\phi^\theta;\theta^\star) - J(\phi^{\theta^\star};\theta^\star) \lesssim \|\theta - \theta^\star\|.$$

[Guo, Hu and Zhang, 2021]

If f and g satisfies additional smoothness condition, then

$$J(\phi^\theta;\theta^\star) - J(\phi^{\theta^\star};\theta^\star) \lesssim \|\theta - \theta^\star\|^2.$$

[Szpruch, Treetanthiploet and Zhang, 2021]

From Estimation Error to Learning

Recall that the posterior of $\theta^\star$ after the m-th episode is $N(\hat{\theta}_m, V_m)$ with

$$\begin{aligned} V_m^{-1} &= V_0^{-1} + \sum_{n=1}^{m} \int_0^T Z_t^n (Z_t^n)^\top \mathrm{d}t \\ \hat{\theta}_m &= \left(\hat{\theta}_0 V_0^{-1} + \sum_{n=1}^{m} \int_0^T Z_t^n \mathrm{d}X_t^n \right) V_m \end{aligned} \quad ; \quad Z_t^n = \begin{pmatrix} X_t^n \\ \varphi_n(\cdot, t, X_t^n) \end{pmatrix}.$$

Therefore,

$$\|\hat{\theta}_m - \theta^\star\|^2 \lesssim \|V_m\| \approx \left(\Lambda_{\min} \left(\sum_{n=1}^{m} \int_0^T Z_t^n (Z_t^n)^\top \mathrm{d}t \right) \right)^{-1}.$$

From Estimation Error to Learning

Recall that the posterior of $\theta^\star$ after the m-th episode is $N(\hat{\theta}_m, V_m)$ with

$$V_m^{-1} = V_0^{-1} + \sum_{n=1}^{m} \int_0^T Z_t^n (Z_t^n)^\top \mathrm{d}t$$

$$\hat{\theta}_m = \left(\hat{\theta}_0 V_0^{-1} + \sum_{n=1}^{m} \int_0^T Z_t^n \mathrm{d}X_t^n \right) V_m \quad ; \quad Z_t^n = \begin{pmatrix} X_t^n \\ \varphi_n(\cdot, t, X_t^n) \end{pmatrix}.$$

Therefore,

$$\|\hat{\theta}_m - \theta^\star\|^2 \lesssim \|V_m\| \approx \left(\Lambda_{\min} \left(\sum_{n=1}^{m} \int_0^T Z_t^n (Z_t^n)^\top \mathrm{d}t \right) \right)^{-1}.$$

- **Phase Exploration:** Dedicating some episodes for exploration.
- **Noisy Exploration:** Taking optimal policies with noise for exploration.

Phase Exploration

Let $\{a_1, a_2, ..., a_p\} \subseteq \mathbb{R}^p$ be linearly independent and

$$\phi^e(t,x) := a_k \quad ; \quad t \in \left[(k-1)\left(\tfrac{T}{p}\right), k\left(\tfrac{T}{p}\right)\right).$$

Since $a_1 a_1^\top + a_2 a_2^\top + \cdots + a_p a_p^\top$ is a strictly positive definite matrix,

$$\Lambda_{\min}\left(\int_0^T \begin{pmatrix} X_t^{\theta^\star,\phi^e} \\ \phi^e(t, X_t^{\theta^\star,\phi^e}) \end{pmatrix} \begin{pmatrix} X_t^{\theta^\star,\phi^e} \\ \phi^e(t, X_t^{\theta^\star,\phi^e}) \end{pmatrix}^\top \mathrm{d}t\right) \gtrsim 1.$$

Phase Exploration Greedy Exploitation (PEGE)

1: **for** $k = 1, 2, ...$ **do**
2: Execute the feedback policy ϕ^e.
3: **for** $l = 1, 2, ..., \mathfrak{m}(k)$ **do**
4: Execute the feedback policy $\phi^{\hat{\theta}_m}$.
5: **end for**
6: **end for**

k-th cycle

Exploration–Exploitation trade-off (Phase-based)

Let $\kappa(m)$ be the cycle corresponding to the m-th episode.
Since $\Lambda_{\min}\left(\sum_{n=1}^{m}\int_0^T Z_t^n(Z_t^n)^\top \mathrm{d}t\right) \gtrsim \kappa(m)$,

$$\|\hat{\theta}_m - \theta^\star\|^2 \lesssim \left(\Lambda_{\min}\left(\sum_{n=1}^{m}\int_0^T Z_t^n(Z_t^n)^\top \mathrm{d}t\right)\right)^{-1} \lesssim \kappa(m)^{-1}.$$

Exploration–Exploitation trade-off (Phase-based)

Let $\kappa(m)$ be the cycle corresponding to the m-th episode.
Since $\Lambda_{\min}\left(\sum_{n=1}^{m}\int_0^T Z_t^n (Z_t^n)^\top \mathrm{d}t\right) \gtrsim \kappa(m)$,

$$\|\hat{\theta}_m - \theta^\star\|^2 \lesssim \left(\Lambda_{\min}\left(\sum_{n=1}^{m}\int_0^T Z_t^n (Z_t^n)^\top \mathrm{d}t\right)\right)^{-1} \lesssim \kappa(m)^{-1}.$$

Suppose that $J(\phi^\theta;\theta^\star) - J(\phi^{\theta^\star};\theta^\star) \lesssim \|\theta-\theta^\star\|^{2r}$. Then

$$\begin{aligned}
\mathrm{Reg}(N) &= \sum_{m=1}^{N}\left(J(\varphi_m;\theta^\star) - J(\phi^{\theta^\star};\theta^\star)\right) \\
&\le \sum_{m=1|\varphi_m=\phi^e}^{N}\left(J(\phi^e;\theta^\star) - J(\phi^{\theta^\star};\theta^\star)\right) + \sum_{m=1|\varphi_m\neq\phi^e}^{N}\left(J(\phi^{\hat{\theta}_m};\theta^\star) - J(\phi^{\theta^\star};\theta^\star)\right) \\
&\lesssim \kappa(N) + \sum_{m=1}^{N}\|\hat{\theta}_m-\theta^\star\|^{2r} \lesssim \kappa(N) + \sum_{m=1}^{N}\kappa(m)^{-r} \quad \Rightarrow \quad \kappa^*(m)\sim m^{\frac{1}{1+r}},
\end{aligned}$$

with $\mathrm{Reg}(N)\approx\mathcal{O}(N^{\frac{1}{1+r}})$. Using $m\approx\sum_{k=1}^{\kappa(m)}\mathfrak{m}(k)$, we obtain $\mathfrak{m}^*(k)\sim k^r$.

Regret of the PEGE algorithm

Regret of the PEGE algorithm

Suppose that $J(\phi^{\theta};\theta^{\star}) - J(\phi^{\theta^{\star}};\theta^{\star}) \lesssim \|\theta-\theta^{\star}\|^{2r}$. Then for the PEGE algorithm with $\mathfrak{m}(k) = \lfloor k^r \rfloor, \forall k \in \mathbb{N}$, there exists a constant $C \geq 0$ such that for all $\delta \in (0,1)$, the regret satisfies with probability at least $1-\delta$,

$$\mathrm{Reg}(N) \leq C\Big(N^{\frac{1}{1+r}}\big((\ln N)^r + \big(\ln(\tfrac{1}{\delta})\big)^r\big) + \big(\ln(\tfrac{1}{\delta})\big)^{1+r}\Big), \quad \forall N \geq 2$$

Consequently,

$$\mathbb{E}[\mathrm{Reg}(N)] \leq C N^{\frac{1}{1+r}} (\ln N)^r, \quad \forall N \geq 2.$$

When $r=1$, $\mathrm{Reg}(N) = \tilde{\mathcal{O}}(\sqrt{N})$.

Noisy Exploration and LQ-Regularised Control

Let consider the optimal solution of the LQ-Regularised control;

$$J(\nu;\theta) = \mathbb{E}\left[\int_0^T \left(\int f(t, \tilde{X}_t^{\theta,\nu}, a)\nu_t(\mathrm{d}a) + \varrho\mathcal{H}(\nu_t)\right)\mathrm{d}t + g(\tilde{X}_T^{\theta^\star,\nu})\right],$$

when f and g are quadratic, $\mathcal{H}(\nu) := \int \ln\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu_{Leb}}\right)\mathrm{d}\nu$ and

$$\mathrm{d}\tilde{X}_t^{\theta,\nu} = \int (A\tilde{X}_t^{\theta,\nu} + Ba)\nu_t(\mathrm{d}a)\,\mathrm{d}t + \mathrm{d}W_t, \quad \tilde{X}_0^{\theta,\nu} = x_0.$$

The optimal feedback measure is $\nu^\theta(t,x) = N(\phi^\theta(t,x), \lambda^2)$ for some $\lambda > 0$.

Noisy Exploration and LQ-Regularised Control

Let consider the optimal solution of the LQ-Regularised control;

$$J(\nu;\theta) = \mathbb{E}\left[\int_0^T \left(\int f(t,\tilde{X}_t^{\theta,\nu},a)\nu_t(\mathrm{d}a) + \varrho\mathcal{H}(\nu_t)\right)\mathrm{d}t + g(\tilde{X}_T^{\theta^\star,\nu})\right],$$

when f and g are quadratic, $\mathcal{H}(\nu) := \int \ln\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu_{Leb}}\right)\mathrm{d}\nu$ and

$$\mathrm{d}\tilde{X}_t^{\theta,\nu} = \int (A\tilde{X}_t^{\theta,\nu} + Ba)\nu_t(\mathrm{d}a)\,\mathrm{d}t + \mathrm{d}W_t, \quad \tilde{X}_0^{\theta,\nu} = x_0.$$

The optimal feedback measure is $\nu^\theta(t,x) = N(\phi^\theta(t,x),\lambda^2)$ for some $\lambda > 0$.

Learning with Regularised Control

1: **for** $m = 1,2,\ldots$ **do**
2: Solve a regularised control problem with $\hat{\theta}_m$ and hyper-parameter ϱ_m to obtain ν_m.
3: Execute ν_m through a random execution φ_m.
4: Use an observed process $X^m = X^{\theta,\varphi_m}$.
5: **end for**

Exploration–Exploitation trade-off (Noisy Exploration)

Let $\xi_m(t) = \sum_{i=1}^{K} \zeta_{i,m} \mathbf{1}_{t \in [(i-1)h, ih)}$ where $\zeta_{i,m} \sim_{IID} N(0,1)$ and consider a policy

$$\varphi_m(t,x) = \phi^{\hat{\theta}_m}(t,x) + \lambda_m \xi_m(t).$$

$$\|\hat{\theta}_m - \theta^\star\|^2 \lesssim \left(\Lambda_{\min}\left(\sum_{n=1}^{m} \int_0^T Z_t^n (Z_t^n)^\top \mathrm{d}t\right)\right)^{-1} \lesssim \left(\sum_{n=1}^{m} \lambda_n^2\right)^{-1}.$$

Exploration–Exploitation trade-off (Noisy Exploration)

Let $\xi_m(t) = \sum_{i=1}^{K} \zeta_{i,m} \mathbf{1}_{t\in[(i-1)h,ih)}$ where $\zeta_{i,m} \sim_{IID} N(0,1)$ and consider a policy

$$\varphi_m(t,x) = \phi^{\hat{\theta}_m}(t,x) + \lambda_m \xi_m(t).$$

$$\|\hat{\theta}_m - \theta^\star\|^2 \lesssim \left(\Lambda_{\min}\left(\sum_{n=1}^{m}\int_0^T Z_t^n (Z_t^n)^\top \mathrm{d}t\right)\right)^{-1} \lesssim \left(\sum_{n=1}^{m}\lambda_n^2\right)^{-1}.$$

We can now quantify the regret when $J(\phi^\theta;\theta^\star) - J(\phi^{\theta^\star};\theta^\star) \lesssim \|\theta-\theta^\star\|^2$ by

$$\begin{aligned}
\mathrm{Reg}(N) &= \sum_{m=1}^{N}\left(J(\varphi_m;\theta^\star) - J(\phi^{\theta^\star};\theta^\star)\right) \\
&= \sum_{m=1}^{N}\left(J(\varphi_m;\theta^\star) - J(\phi^{\hat{\theta}_m};\theta^\star)\right) + \sum_{m=1}^{N}\left(J(\phi^{\hat{\theta}_m};\theta^\star) - J(\phi^{\theta^\star};\theta^\star)\right) \\
&\lesssim \sum_{m=1}^{N}\lambda_m^2 + \sum_{m=1}^{N}\left(\sum_{n=1}^{m}\lambda_n^2\right)^{-1} \quad \Rightarrow \quad \lambda_m^2 \sim m^{-1/2} \quad \text{with} \quad \mathrm{Reg}(N) \approx \mathcal{O}(\sqrt{N}).
\end{aligned}$$

Regret of the Regularised Control Algorithm

Regret of the Regularised Control Algorithm

Suppose that f and g are quadratic. Then by choosing an appropriate $(\varrho_m)_{m\in\mathbb{N}}$ and execution increment, there exists a constant $C \geq 0$ such that for all $\delta \in (0,1)$, the regret for learning with regularised control satisfies with probability at least $1-\delta$,

$$\mathrm{Reg}(N) \leq C\sqrt{N}\mathrm{Poly}\Big(\ln N, \ln\left(\tfrac{1}{\delta}\right)\Big).$$

and $\mathbb{E}[\mathrm{Reg}(N)] \leq C\sqrt{N}\mathrm{Poly}(\ln N)$.

NB. This result also holds for a different regularised control problem where the divergence between episodes is penalised to the Hamiltonian to ensure that our policy does not change too much between episodes.

References:

- L. Szpruch, T. Treetanthiploet, and Y. Zhang, Exploration-exploitation trade-off for continuous-time episodic reinforcement learning with linear-convex models, arXiv preprint arXiv:2112.10264, (2021).
- L. Szpruch, T. Treetanthiploet, and Y. Zhang, Linear-quadratic reinforcement learning via relaxed controls, to be appeared.