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

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

where

\[
dX_{t}^{\theta^*, \alpha} = (A^* X_{t}^{\theta^*, \alpha} + B^* \alpha_t) \, dt + dW_t, \quad X_0^{\theta^*, \alpha} = x_0.
\]
Let $\theta^* = (A^*, B^*) \in \mathbb{R}^{d \times d} \times \mathbb{R}^{p \times d}$ and consider the control problem

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

where

$$dX^{\theta^*,\alpha}_t = (A^* X^{\theta^*}_t + B^* \alpha_t) \, dt + dW_t, \quad X^{\theta^*,\alpha}_0 = x_0.$$

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

$$\alpha^*_t := \phi^{\theta^*}(t, X^{\theta^*_t,\alpha^*_t}) = \arg \min_{\alpha \in \mathcal{H}^2_{\mathbb{F}}(\Omega;\mathbb{R}^p)} J(\alpha; \theta^*).$$
Let $\theta^* = (A^*, B^*) \in \mathbb{R}^{d \times d} \times \mathbb{R}^{p \times d}$ and consider the control problem

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

where

$$dX_{t}^{\theta^*, \alpha} = (A^* X_{t}^{\theta^*, \alpha} + B^* \alpha_t) \, dt + dW_t, \quad X_0^{\theta^*, \alpha} = x_0.$$

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

$$\alpha_t^* := \phi^{\theta^*} (t, X_{t}^{\theta^*, \alpha^*}) = \arg \min_{\alpha \in \mathcal{H}_2^2(\Omega; \mathbb{R}^p)} J(\alpha; \theta^*).$$

We do not know $\theta^*$ and thus cannot find $\phi^{\theta^*}$. 
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 $(X_t^m)_{t \in [0, T]}$;

\[ dX_t^m = (A^* X_t^m + B^* \varphi_m(\cdot, t, X_t^m)) \, dt + dW_t^m, \quad X_0^m = x_0 \]

and experience the (expected) cost

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

- Design $\varphi_{m+1}$ from the previous observations, $(X^n_{m})_{n=1}$. 
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 $(X_t^m)_{t \in [0, T]}$;

$$dX_t^m = (A^* X_t^m + B^* \varphi_m(\cdot, t, X_t^m)) \, dt + dW_t^m, \quad X_0^m = x_0$$

and experience the (expected) cost

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

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

The agent objective is to minimise

$$\text{Reg}(N) = \sum_{m=1}^N \left( J(\varphi_m; \theta^*) - J(\varphi^{\theta^*}; \theta^*) \right).$$
Statistical Estimate from Bayesian inference

Suppose that before the \( m \)-th episode, the posterior of \( \theta^* \) is \( N(\hat{\theta}_{m-1}, V_{m-1}) \).

At the \( m \)-th episode, the agent observes \((X_t^m)\) satisfying

\[
dX_t^m = \theta^* Z_t^m \, dt + dW_t^m, \quad X_0^m = x_0, \text{ with } Z_t^m = \left( X_t^m \right).
\]

\( \phi_m(\cdot, t, X_t^m) \).
Statistical Estimate from Bayesian inference

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

$$dX^m_t = \theta^* Z^m_t \, dt + dW^m_t, \quad X^m_0 = x_0, \text{ with } Z^m_t = \left( X^m_t \right).$$

Let consider a discretisation $(\Delta X^m_{t_1}, \Delta X^m_{t_2}, \ldots, \Delta X^m_{t_K})$ where

$$\Delta X^m_{t_k} = \theta^* Z^m_{t_k} \Delta t + \Delta W_{t_k} \sim \mathcal{N}(\theta^* Z^m_{t_k} \Delta t, \Delta t)$$
Statistical Estimate from Bayesian inference

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

$$dX_t^m = \theta^* Z_t^m \, dt + dW_t^m, \quad X_0^m = x_0, \quad \text{with} \quad Z_t^m = \left( X_t^m \right).$$

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

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

Therefore,

$$\pi \left( \theta^* \mid F_{m-1}, (\Delta X_{t_k}^m)_{k=1}^K, (Z_{t_k}^m)_{k=1}^K \right) \propto \exp \left( -\frac{1}{2} (\theta^* - \hat{\theta}_{m-1}) V_{m-1} (\theta^* - \hat{\theta}_{m-1})^\top \right) \prod_{k=1}^K \exp \left( -\frac{1}{2\Delta t} (\Delta X_{t_k}^m - \theta^* Z_{t_k}^m \Delta t)^2 \right)$$

$$\propto \exp \left( -\frac{1}{2} \theta^* \left( V_{m-1}^{-1} + \sum_{k=1}^K Z_{t_k}^m (Z_{t_k}^m)^\top \Delta t \right) \theta^*^\top \right) + \theta^* \left( V_{m-1}^{-1} \hat{\theta}_{m-1}^\top + \sum_{k=1}^K Z_{t_k}^m \Delta X_{t_k}^m \right).$$
In particular, if we send $\Delta t \to 0$, the posterior is

\[
\pi (\theta^* | \mathcal{F}_{m-1}, (X_t^m)_{t \in [0, T]}, (Z_t^m)_{t \in [0, T]}) 
\propto \exp \left( -\frac{1}{2} \theta^* \left( V_{m-1}^{-1} + \int_0^T Z_t^m (Z_t^m)^\top \, dt \right) \theta^* \right) + \theta^* \left( V_{m-1}^{-1} \hat{\theta}_{m-1} + \int_0^T Z_t^m dX_t^m \right).
\]
In particular, if we send $\Delta t \to 0$, the posterior is

$$
\pi \left( \theta^* \mid \mathcal{F}_{m-1}, (X_t^m)_{t \in [0, \tau]}, (Z_t^m)_{t \in [0, \tau]} \right) \\
\propto \exp \left( -\frac{1}{2} \theta^* \left( V_{m-1}^{-1} + \int_0^T Z_t^m (Z_t^m)^\top \, dt \right) \theta^* \right) + \theta^* \left( V_{m-1}^{-1} \hat{\theta}_{m-1}^\top + \int_0^T Z_t^m dX_t^m \right).
$$

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

$$
V_m^{-1} = V_0^{-1} + \sum_{n=1}^m \int_0^T Z_t^n (Z_t^n)^\top \, dt, \quad \text{and} \quad \hat{\theta}_m = \left( \hat{\theta}_0 V_0^{-1} + \sum_{n=1}^m \int_0^T Z_t^n dX_t^n \right) V_m.
$$
In particular, if we send $\Delta t \to 0$, the posterior is
\[
\pi \left( \theta^* \mid \mathcal{F}_{m-1}, (X_t^m)_{t \in [0, \tau]}, (Z_t^m)_{t \in [0, \tau]} \right)
\propto \exp \left( -\frac{1}{2} \theta^* \left( V^{-1}_{m-1} + \int_0^T Z_t^m (Z_t^m)^\top \, dt \right) \theta^\top \right) + \theta^* \left( V^{-1}_{m-1} \hat{\theta}^\top_{m-1} + \int_0^T Z_t^m \, dX_t^m \right).
\]

In particular, the posterior of $\theta^*$ after the $m$-th episode is $N(\hat{\theta}_m, V_m)$ where
\[
V^{-1}_m = V^{-1}_0 + \sum_{n=1}^m \int_0^T Z_t^n (Z_t^n)^\top \, dt, \quad \text{and} \quad \hat{\theta}_m = \left( \hat{\theta}_0 V^{-1}_0 + \sum_{n=1}^m \int_0^T Z_t^n \, dX_t^n \right) V_m.
\]

In comparison to the classical statistics theory, one may see
- $\hat{\theta}_m$ as a (regularised) maximum likelihood estimator.
- $V^{-1}_m$ as a (regularised) Fisher Information.
Sub-optimality of the Greedy policy

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

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

where

$$dX_t^{\theta,\alpha} = (B_1 \alpha_{1,t} + B_2 \alpha_{2,t}) \, dt + dW_t, \quad X_0^{\theta,\alpha} = x_0.$$

The optimal policy is

$$\phi^\theta(t, x) = - (1 + (B_1^2 + B_2^2)(T - t))^{-1} \begin{pmatrix} B_1 x \\ B_2 x \end{pmatrix}.$$
Sub-optimality of the Greedy policy

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

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

where

$$dX_t^{\theta,\alpha} = (B_1 \alpha_{1,t} + B_2 \alpha_{2,t}) \, dt + dW_t, \quad X_0^{\theta,\alpha} = x_0.$$ 

The optimal policy is

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

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

- Decisions are always sub-optimal provided that $B_2^* \neq 0.$
Sensitivity in Parameter Estimate

Performance Gap

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

$$dX_t^{\theta, \phi} = (AX_t^{\theta, \phi} + B\phi(t, X_t)) \, dt + dW_t, \quad X_0^{\theta^*, \phi} = x_0.$$  

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

$$J(\phi^\theta; \theta^*) - J(\phi^\theta^*; \theta^*) \lesssim \|\theta - \theta^*\|.$$  

[Guo, Hu and Zhang, 2021]

If $f$ and $g$ satisfies additional smoothness condition, then

$$J(\phi^\theta; \theta^*) - J(\phi^\theta^*; \theta^*) \lesssim \|\theta - \theta^*\|^2.$$  

[Szpruch, Treetanthiploet and Zhang, 2021]
Recall that the posterior of $\theta^*$ after the $m$-th episode is $N(\hat{\theta}_m, V_m)$ with

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

$$\hat{\theta}_m = \left( \hat{\theta}_0 V_0^{-1} + \sum_{n=1}^{m} \int_0^T Z_t^n \, dX_t^n \right) V_m$$

Therefore,

$$\|\hat{\theta}_m - \theta^*\|^2 \lesssim \|V_m\| \approx \left( \Lambda_{\min} \left( \sum_{n=1}^{m} \int_0^T Z_t^n (Z_t^n)^\top \, dt \right) \right)^{-1}.$$
Recall that the posterior of $\theta^*$ after the $m$-th episode is $N(\hat{\theta}_m, V_m)$ with

$$V_m^{-1} = V_0^{-1} + \sum_{n=1}^m \int_0^T Z^n_t (Z^n_t)^\top dt$$

$$\hat{\theta}_m = \left( \hat{\theta}_0 V_0^{-1} + \sum_{n=1}^m \int_0^T Z^n_t dX^n_t \right) V_m$$

Therefore,

$$\|\hat{\theta}_m - \theta^*\|^2 \lesssim \|V_m\| \approx \left( \Lambda_{\min} \left( \sum_{n=1}^m \int_0^T Z^n_t (Z^n_t)^\top dt \right) \right)^{-1}.$$

- **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

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

Since \( a_1 a_1^T + a_2 a_2^T + \cdots + a_p a_p^T \) is a strictly positive definite matrix,

\[
\Lambda \min \left( \int_0^T \left( \begin{array}{c} X_{t}^{\theta^*, \phi^e} \\ \phi^e(t, X_t^{\theta^*, \phi^e}) \end{array} \right) \left( \begin{array}{c} X_{t}^{\theta^*, \phi^e} \\ \phi^e(t, X_t^{\theta^*, \phi^e}) \end{array} \right)^T dt \right) \gtrsim 1.
\]

Phase Exploration Greedy Exploitation (PEGE)

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

\( k \)-th cycle
Let $\kappa(m)$ be the cycle corresponding to the $m$-th episode. Since $\Lambda_{\text{min}} \left( \sum_{n=1}^{m} \int_{0}^{T} Z_t^n (Z_t^n)^\top \, dt \right) \gtrsim \kappa(m)$,

$$
\|\hat{\theta}_m - \theta^*\|^2 \lesssim \left( \Lambda_{\text{min}} \left( \sum_{n=1}^{m} \int_{0}^{T} Z_t^n (Z_t^n)^\top \, dt \right) \right)^{-1} \lesssim \kappa(m)^{-1}.
$$
Let $\kappa(m)$ be the cycle corresponding to the $m$-th episode. Since $\Lambda_{\text{min}} \left( \sum_{n=1}^{m} \int_{0}^{T} Z_{t}^{n}(Z_{t}^{n})^{\top} \, dt \right) \gtrsim \kappa(m)$,

$$||\hat{\theta}_{m} - \theta^{*}||^{2} \lesssim \left( \Lambda_{\text{min}} \left( \sum_{n=1}^{m} \int_{0}^{T} Z_{t}^{n}(Z_{t}^{n})^{\top} \, dt \right) \right)^{-1} \lesssim \kappa(m)^{-1}.$$ 

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

$$\text{Reg}(N) = \sum_{m=1}^{N} \left( J(\varphi_{m}; \theta^{*}) - J(\phi^{\theta^{*}}; \theta^{*}) \right) \leq \sum_{m=1}^{N} \left( J(\phi^{e}; \theta^{*}) - J(\phi^{\theta^{*}}; \theta^{*}) \right) + \sum_{m=1}^{N} \left( J(\phi^{\hat{\theta}_{m}}; \theta^{*}) - J(\phi^{\theta^{*}}; \theta^{*}) \right) \lesssim \kappa(N) + \sum_{m=1}^{N} ||\hat{\theta}_{m} - \theta^{*}||^{2r} \lesssim \kappa(N) + \sum_{m=1}^{N} \kappa(m)^{-r} \Rightarrow \kappa^{*}(m) \sim m^{1+\frac{1}{r}},$$

with $\text{Reg}(N) \approx O(N^{1+r})$. Using $m \approx \sum_{k=1}^{\kappa(m)} m(k)$, we obtain $m^{*}(k) \sim k^{r}$.
Regret of the PEGE algorithm

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

\[
\text{Reg}(N) \leq C \left( N^{\frac{1}{1+r}} \left( (\ln N)^r + (\ln(\frac{1}{\delta}))^r \right) + (\ln(\frac{1}{\delta}))^{1+r} \right), \quad \forall N \geq 2
\]

Consequently,

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

When \( r = 1 \), \( \text{Reg}(N) = \tilde{O}(\sqrt{N}) \).
Noisy Exploration and LQ-Regularised Control

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

\[
J(\nu; \theta) = \mathbb{E} \left[ \int_0^T \left( \int f(t, \tilde{X}^\theta_t, a) \nu_t(da) + \varrho \mathcal{H}(\nu_t) \right) dt + g(\tilde{X}^\theta_T) \right],
\]

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

\[
d\tilde{X}^\theta_t = \int (A\tilde{X}^\theta_t + Ba) \nu_t(da) dt + dW_t, \quad \tilde{X}^\theta_0 = x_0.
\]

The optimal feedback measure is \( \nu^\theta(t, x) = N(\phi^\theta(t, x), \lambda^2) \) for some \( \lambda > 0 \).
Let consider the optimal solution of the LQ-Regularised control;

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

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

\[
d\tilde{X}_t^{\theta, \nu} = \int (A\tilde{X}_t^{\theta, \nu} + Ba) \nu_t(da) dt + dW_t, \quad \tilde{X}_0^{\theta, \nu} = x_0.
\]

The optimal feedback measure is \( \nu^{\theta}(t, x) = \mathcal{N}(\phi^{\theta}(t, x), \lambda^2) \) 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 \( \varrho_m \) to obtain \( \nu_m \).
3. Execute \( \nu_m \) through a random execution \( \varphi_m \).
4. Use an observed process \( X^m = X^{\theta, \varphi_m} \).
5. **end for**
Let $\xi_m(t) = \sum_{i=1}^{K} \zeta_{i,m} \mathbf{1}_{t \in [(i-1)h, ih)}$ where $\zeta_{i,m} \sim \text{IID } N(0, 1)$ and consider a policy

$$\varphi_m(t, x) = \phi^\theta_m(t, x) + \lambda_m \xi_m(t).$$

$$\|\hat{\theta}_m - \theta^*\|^2 \lesssim \left( \Lambda_{\text{min}} \left( \sum_{n=1}^{m} \int_{0}^{T} Z_t^n (Z_t^n)^\top dt \right) \right)^{-1} \lesssim \left( \sum_{n=1}^{m} \lambda_n^2 \right)^{-1}.$$
Exploration–Exploitation trade-off (Noisy Exploration)

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

\[
\varphi_m(t, x) = \phi_{\hat{\theta}_m}(t, x) + \lambda_m \xi_m(t).
\]

\[
\|\hat{\theta}_m - \theta^*\|_2^2 \lesssim \left( \Lambda_{\min} \left( \sum_{n=1}^{m} \int_{0}^{T} Z_t^n (Z_t^n)^	op dt \right) \right)^{-1} \lesssim \left( \sum_{n=1}^{m} \lambda_n^2 \right)^{-1}.
\]

We can now quantify the regret when

\[
J(\phi_{\theta}; \theta^*) - J(\phi_{\theta^*}; \theta^*) \lesssim \|\theta - \theta^*\|_2^2
\]

by

\[
\text{Reg}(N) = \sum_{m=1}^{N} \left( J(\varphi_m; \theta^*) - J(\phi_{\theta^*}; \theta^*) \right)
\]

\[
= \sum_{m=1}^{N} \left( J(\varphi_m; \theta^*) - J(\phi_{\hat{\theta}_m}; \theta^*) \right) + \sum_{m=1}^{N} \left( J(\phi_{\hat{\theta}_m}; \theta^*) - J(\phi_{\theta^*}; \theta^*) \right)
\]

\[
\lesssim \sum_{m=1}^{N} \lambda_m^2 + \sum_{m=1}^{N} \left( \sum_{n=1}^{m} \lambda_n^2 \right)^{-1} \Rightarrow \lambda_m^2 \sim m^{-1/2} \quad \text{with} \quad \text{Reg}(N) \approx \mathcal{O}(\sqrt{N}).
\]
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 \( 1 - \delta \),

\[
\text{Reg}(N) \leq C \sqrt{N} \text{Poly} \left( \ln N, \ln \left( \frac{1}{\delta} \right) \right).
\]

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

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.

L. Szpruch, T. Treetanthiploet, and Y. Zhang, Linear-quadratic reinforcement learning via relaxed controls, to be appeared.