Appendix A Continuous RL: Formulation and Well-Posedness

A.1 Exploratory Stochastic-Control

For n, m positive integers, let $b : \mathbb{R}^n \times \mathcal{A} \mapsto \mathbb{R}^n$ and $\sigma : \mathbb{R}^n \times \mathcal{A} \mapsto \mathbb{R}^{n\times m}$ be given functions, where $\mathcal{A}$ is a compact action space. A classical stochastic control problem [15, 62] is to control the state (or feature) dynamics governed by an Itô process, defined on a filtered probability space $\left(\Omega, \mathcal{F}, \mathbb{P}; \{\mathcal{F}_s^B\}_{s\geq 0}\right)$, along with an $\{\mathcal{F}_s^B\}$-Brownian motion $B = \{B_s, s \geq 0\}$:

$$\mathrm{d}X_s^a = b\left(X_s^a, a_s\right)\mathrm{d}s + \sigma\left(X_s^a, a_s\right)\mathrm{d}B_s,\ s \geq t, \quad X_t = x, \tag{29}$$

where a_s is the agent's action (control) at time s. The goal of the stochastic control (discounted objective over an infinite time horizon) is for any time-state pair (t, x) in (29), to find the optimal $\{\mathcal{F}_s^B\}_{s\geq 0}$-progressively measurable sequence of actions $a = \{a_s, s \geq t\}$ (called the optimal policy) that maximizes the expected total β-discounted reward:

$$\mathbb{E}\left[\int_t^{+\infty} e^{-\beta(s-t)} r\left(X_s^a, a_s\right)\mathrm{d}s \mid X_t^a = x\right], \tag{30}$$

where $r : \mathbb{R}^n \times \mathcal{A} \mapsto \mathbb{R}$ is the running reward of the current state and action (X_s^a, a_s), and $\beta > 0$ is a discount factor that measures the time-depreciation of the objective value (or the impatience level of the agent). Note that the state process $X^a = \{X_s^a, s \geq t\}$ depends on the starting (initial) time-state pair (t, x). For ease of notation, we denote by X^a instead of $X^{t,x,a} = \{X_s^{t,x,a}, s \geq t\}$ the solution to the SDE in (29) when there is no ambiguity.

Listed below are the standard assumptions to ensure the well-posedness of the stochastic control problem in (29)-(30).

Assumption 2. *The following conditions are assumed throughout:*
(i) b, σ, r are all continuous functions in their respective arguments;
(ii) b, σ are uniformly Lipschitz continuous in x, i.e., there exists a constant $C > 0$ such that for $\varphi \in \{b, \sigma\}$,

$$\|\varphi(x,a) - \varphi(x', a)\|_2 \leq C\|x - x'\|_2, \quad \textit{for all } a \in \mathcal{A},\ x, x' \in \mathbb{R}^n; \tag{31}$$

(iii) b, σ have linear growth in x and a, i.e., there exists a constant $C > 0$ such that for $\varphi \in \{b, \sigma\}$,

$$\|\varphi(x,a)\|_2 \leq C(1 + \|x\|_2 + \|a\|_2), \quad \textit{for all } (x,a) \in \mathbb{R}^n \times \mathcal{A}; \tag{32}$$

(iv) r has polynomial growth in x and a, i.e., there exists a constant $C > 0$ and $\mu \geq 1$ such that

$$|r(x,a)| \leq C\left(1 + \|x\|_2^\mu + \|a\|_2^\mu\right) \quad \textit{for all } (x,a) \in \mathbb{R}^n \times \mathcal{A}. \tag{33}$$

The key idea underlying *exploratory* stochastic control is to use a randomized policy (or relaxed control), i.e., apply a probability distribution to the admissible action space. To do so, let's assume the probability space is rich enough to support a uniform random variable Z that is independent of the Brownian motion $B = \{B_t\}$. We then expand the original filtered probability space to $\left(\Omega, \mathcal{F}, \mathbb{P}; \{\mathcal{F}_s\}_{s\geq 0}\right)$, where $\mathcal{F}_s = \mathcal{F}_s^B \vee \sigma(Z)$ (i.e., augment $\mathcal{F}_s^B$ with the sigma field generated by Z).

Let $\pi : \mathbb{R}^n \ni x \mapsto \pi(\cdot \mid x) \in \mathcal{P}(\mathcal{A})$ be a stationary feedback policy given the state at x, where $\mathcal{P}(\mathcal{A})$ is a suitable collection of probability distributions (with density functions). At each time s, an action a_s is generated from the distribution $\pi\left(\cdot \mid X_s^a\right)$, i.e. the policy only depends on the current state. In other words, we only consider stationary, or time-independent feedback control policies for the stochastic control problem (29)-(30).

Given a stationary policy $\pi \in \mathcal{P}(\mathcal{A})$, an initial state x, and an $\{\mathcal{F}_s\}$-progressively measurable action process $a^\pi = \{a_s^\pi, s \geq 0\}$ generated from π, the state process $X^\pi = \{X_s^\pi, s \geq 0\}$ follows:

$$\mathrm{d}X_s^\pi = b\left(X_s^\pi, a_s^\pi\right)\mathrm{d}s + \sigma\left(X_s^\pi, a_s^\pi\right)\mathrm{d}B_s,\ s \geq t, \quad X_0^\pi = x, \tag{34}$$

defined on $\left(\Omega, \mathcal{F}, \mathbb{P}; \{\mathcal{F}_s\}_{s\geq 0}\right)$. It is easy to see that the dynamics in (34) define a time-homogeneous Markov process, such that for each $t \geq 0$ and x:

$$\left(X_s^\pi \mid X_0^\pi = x\right) \overset{d}{=} \left(X_{s+t}^\pi \mid X_t^\pi = x\right),\ s \geq 0.$$

Consequently, the objective in (30) is independent of time t, and is equal to:

$$\mathbb{E}\left[\int_0^{+\infty} e^{-\beta s} r\left(X_s^\pi, a_s^\pi\right) \mathrm{d}s \mid X_0^\pi = x\right]. \tag{35}$$

Furthermore, following [58], we can add a regularizer to the objective function to encourage exploration (represented by the randomized policy), leading to

$$V(t,x;\pi) := \mathbb{E}\left[\int_t^{\infty} e^{-\beta(s-t)}\left[r\left(X_s^\pi, a_s^\pi\right) + \gamma p\left(X_s^\pi, a_s^\pi, \pi\left(\cdot \mid X_s^\pi\right),\right)\right] \mathrm{d}s \mid X_t^\pi = x\right], \tag{36}$$

where $p : \mathbb{R}^n \times \mathcal{A} \times \mathcal{P}(\mathcal{A}) \mapsto \mathbb{R}$ is the regularizer, and $\gamma \geq 0$ is a weight parameter on exploration (also known as the "temperature" parameter). For instance, in [58], p is taken as the differential entropy,

$$p(x, a, \pi(\cdot)) := -\log \pi(a),$$

and hence, the "entropy" regularizer. The same argument as before justifies that $V(t,x;\pi)$ is independent of time t. That is, for all $t \geq 0$,

$$V(t,x;\pi) \equiv V(x;\pi) := \mathbb{E}^{\mathbb{P}}\left[\int_0^{\infty} e^{-\beta s}\left[r\left(X_s^\pi, a_s^\pi\right) + \gamma p\left(X_s^\pi, a_s^\pi, \pi\left(\cdot \mid X_s^\pi\right)\right)\right] \mathrm{d}s \mid X_0^\pi = x\right]; \tag{37}$$

which is the state-value function under the policy π, $V(x;\pi)$, in (4), and which, in turn, leads to the performance function $\eta(\pi)$ in (6). Moreover, recall the main task of the continuous RL is to find (or approximate) $\eta^* = \max_\pi \ \eta(\pi)$, where $\max$ is over all admissible policies.

A.2 Controlled SDE and the HJ Equation

Note that the exploratory state dynamics in (34) is governed by a general Itô process. It is sometimes more convenient to consider an equivalent SDE representation— in the sense that its (weak) solution has the same distribution as the Itô process in (34) at each fixed time t. It is known ([58]) that when $n = m = 1$, the marginal distribution of $\{X_s^\pi, s \geq 0\}$ agrees with that of the solution to the SDE, denoted by $\{\tilde{X}_s, s \geq 0\}$:

$$\mathrm{d}\tilde{X}_s = \tilde{b}\left(\tilde{X}_s, \pi\left(\cdot \mid \tilde{X}_s\right)\right)\mathrm{d}s + \tilde{\sigma}\left(\tilde{X}_s, \pi\left(\cdot \mid \tilde{X}_s\right)\right)\mathrm{d}\tilde{B}_s, \quad \tilde{X}_0 = x,$$

where $\tilde{b}(x,\pi(\cdot)) = \int_{\mathcal{A}} b(x,a)\pi(a)\mathrm{d}a$ and $\tilde{\sigma}(x,\pi(\cdot)) = \sqrt{\int_{\mathcal{A}} \sigma^2(x,a)\pi(a)\mathrm{d}a}$. This result is easily extended to arbitrary n, m, thanks to [7, Corollary 3.7], with the precise statement presented below (assuming $n = m$ for ease of exposition).

Theorem 6. *Assume that for a policy π and for every x,*

$$\int_{\mathcal{A}} \sigma^2(x,a)\pi(a)\mathrm{d}a \in \mathbb{R}^{n\times n},$$

is positive definite. Then there exists a filtered probability space $\left(\tilde{\Omega}, \tilde{\mathcal{F}}, \left\{\tilde{\mathcal{F}}_t\right\}_{t\geq 0}, \tilde{\mathbb{P}}\right)$ that supports a continuous $\mathbb{R}^n$-valued adapted process $\tilde{X}$ and an n-dimensional Brownian motion $\tilde{B}$ satisfying

$$\mathrm{d}\tilde{X}_s = \tilde{b}\left(\tilde{X}_s, \pi\left(\cdot \mid \tilde{X}_s\right)\right)\mathrm{d}s + \tilde{\sigma}\left(\tilde{X}_s, \pi\left(\cdot \mid \tilde{X}_s\right)\right)\mathrm{d}\tilde{B}_s, \quad \tilde{X}_0 = x, \tag{38}$$

where

$$\tilde{b}(x,\pi(\cdot)) = \int_{\mathcal{A}} b(x,a)\pi(a)\mathrm{d}a, \quad \tilde{\sigma}(x,\pi(\cdot)) = \left(\int_{\mathcal{A}} \sigma^2(x,a)\pi(a)\mathrm{d}a\right)^{\frac{1}{2}}.$$

For each $s \geq 0$, the distribution of $\tilde{X}_s$ under $\tilde{\mathbb{P}}$ agrees with that of X_s^π under $\mathbb{P}$ defined in (34).

As a consequence, the state value function in (37) is identical to

$$V(x;\pi) = \mathbb{E}\left[\int_0^{\infty} e^{-\beta s}\int_{\mathcal{A}}\left[r(\tilde{X}_s, a) + \gamma p\left(\tilde{X}_s, a, \pi(\cdot \mid \tilde{X}_s)\right)\right]\pi(a \mid \tilde{X}_s)\mathrm{d}a \, \mathrm{d}s \mid \tilde{X}_0 = x\right].$$

Also define

$$\tilde{r}(x, \pi)=\int_{\mathcal{A}} r(x, a) \pi(a \mid s) \mathrm{d} a, \quad \tilde{p}(x, \pi)=\int_{\mathcal{A}} p(x, a, \pi) \pi(a \mid x) \mathrm{d} a,$$

so we can simplify the value function to

$$V(x ; \pi)=\mathbb{E}\left[\int_0^{\infty} e^{-\beta s}\left[\tilde{r}(\tilde{X}_s, \pi)+\gamma \tilde{p}\left(\tilde{X}_s^\pi, \pi(\cdot \mid \tilde{X}_s)\right)\right] \mathrm{d} s \mid \tilde{X}_0=x\right]. \tag{39}$$

Following the principle of optimality, V then satisfies the HJ equation:

$$\beta V(x ; \pi)-\tilde{b}(x, \pi) \cdot \nabla V(x ; \pi)-\frac{1}{2} \tilde{\sigma}^2(x, \pi) \circ \nabla^2 V(x ; \pi)-\tilde{r}(x, \pi)-\gamma \tilde{p}(x, \pi)=0. \tag{40}$$

To guarantee that the HJ equation in (40) characterizes the state-value function in (39), we need

Assumption 3. *Assume the following conditions hold:*
(i) b, σ, r, p are all continuous functions in their respective arguments.
(ii) b, r, p are uniformly Lipschitz continuous in x, i.e., there exists a constant $C>0$ such that for $\varphi \in\{b, r\}$,

$$\|\varphi(x, a)-\varphi(x^{\prime}, a)\|_2 \leq C\|x-x^{\prime}\|_2, \quad \text{for all } a \in \mathcal{A},\ x, x^{\prime} \in \mathbb{R}^n,$$

and

$$|p(x, a, \pi)-p(x^{\prime}, a, \pi)| \leq C\|x-x^{\prime}\|_2, \quad \text{for all } a \in \mathcal{A},\ \pi \in \mathcal{P}(\mathcal{A}),\ x, x^{\prime} \in \mathbb{R}^n.$$

(iii) $\tilde{\sigma}$ is globally bounded, i.e., there exist $0<\sigma_0<\bar{\sigma}_0$ such that

$$\sigma_0^2 \cdot I \leq \tilde{\sigma}^2(x, a) \leq \bar{\sigma}_0^2 \cdot I, \quad \text{for all } a \in \mathcal{A},\ x \in \mathbb{R}^n.$$

(iv) the SDE (38) has a weak solution which is unique in distribution.
(v) $\pi(a \mid x)$ is measurable in (x, a) and is uniformly Lipschitz continuous in x, i.e., there exists a constant $C>0$ such that

$$\int_{\mathcal{A}}|\pi(a \mid x)-\pi(a \mid x^{\prime})| d a \leq C\|x-x^{\prime}\|_2, \quad \text{for all } x, x^{\prime} \in \mathbb{R}^n.$$

Theorem 7. *Under Assumption 3, the state-value function in (39) is the unique (subquadratic) viscosity solution to the HJ equation in (40).*

Proof. By [56, Section 3.1], the HJ equation in (40) has a unique (subquadratic) viscosity solution under the conditions (i)-(iii). Further by [21, Lemma 2], the viscosity solution is the state-value function. □

Appendix B Proofs of Main Results (in §3)

B.1 Proof of Theorem 2

Recall in the proof sketch of the Theorem in §3, we have defined the operator $\mathcal{L}^\pi: C^2(\mathbb{R}^n) \mapsto C(\mathbb{R}^n)$ as

$$\left(\mathcal{L}^\pi \varphi\right)(x):=-\beta \varphi(x)+\tilde{b}(x, \pi) \cdot \nabla \varphi(x)+\frac{1}{2} \tilde{\sigma}(x, \pi)^2 \circ \nabla^2 \varphi(x),$$

which leads to the following characterization of the HJ equation:

$$-\mathcal{L}^\pi V(x ; \pi)=\tilde{r}(x, \pi)+\gamma \tilde{p}(x, \pi). \tag{41}$$

We need the following two lemmas concerning the operator $\mathcal{L}^\pi$.

Lemma 8. *For any $\varphi \in C^2(\mathbb{R}^n)$, we have*

$$\int_{\mathbb{R}^n} d_x^\pi(y)(-\mathcal{L}^\pi \varphi)(y) \mathrm{d} y=\varphi(x).$$

Proof. The left hand side of the above equation is

$$\begin{aligned}
&= \mathbb{E}\int_0^\infty e^{-\beta s}\left(\beta\varphi(\tilde{X}_s^\pi) - \tilde{b}(\tilde{X}_s^\pi,\pi)\frac{\partial\varphi}{\partial x}(\tilde{X}_s^\pi) - \frac{1}{2}\tilde{\sigma}(\tilde{X}_s^\pi,\pi)^2\frac{\partial^2\varphi}{\partial x^2}(\tilde{X}_s^\pi)\right)\mathrm{d}s \\
&= \mathbb{E}\int_0^\infty e^{-\beta s}\left[\left(\beta\varphi(\tilde{X}_s^\pi) - \tilde{b}(\tilde{X}_s^\pi,\pi)\frac{\partial\varphi}{\partial x}(\tilde{X}_s^\pi) - \frac{1}{2}\tilde{\sigma}(\tilde{X}_s^\pi,\pi)^2\frac{\partial^2\varphi}{\partial x^2}(\tilde{X}_s^\pi)\right)\mathrm{d}s - \tilde{\sigma}(\tilde{X}_s^\pi,\pi)\frac{\partial\varphi}{\partial x}(\tilde{X}_s^\pi)\mathrm{d}B_s\right] \\
&= \mathbb{E}\int_0^\infty \mathrm{d}\left(-e^{-\beta s}\varphi(\tilde{X}_s^\pi)\right) \\
&= \lim_{s\to\infty}\left(-e^{-\beta s}\varphi(\tilde{X}_s^\pi)\right) + \varphi(\tilde{X}_0^\pi) \\
&= \varphi(x),
\end{aligned}$$

where the first equality follows from the definition of the occupation time and the third equality from Itô's formula. □

Lemma 9. *Let $\pi, \hat{\pi}$ be two feedback policies. We have*

$$(\mathcal{L}^{\hat{\pi}} - \mathcal{L}^\pi)V(x;\pi) + \tilde{r}(x,\hat{\pi}) - \tilde{r}(x,\pi) - \gamma\tilde{p}(x,\pi) = \int_{\mathcal{A}(x)} \hat{\pi}(a \mid x)q(x,a;\pi)\mathrm{d}a. \tag{42}$$

Proof. By definition of $q(x,a;\pi)$ in (11), we have

$$\begin{aligned}
\text{RHS} &= \int_{\mathcal{A}(x)} \hat{\pi}(a \mid x)\left(\mathcal{H}^a\left(x, \frac{\partial V}{\partial x}(x;\pi), \frac{\partial^2 V}{\partial x^2}(x;\pi)\right) - \beta V(x;\pi)\right)\mathrm{d}a \\
&= \int_{\mathcal{A}(x)} \hat{\pi}(a \mid x)\left(b(x,a)\cdot\frac{\partial V}{\partial x}(x;\pi) + \frac{1}{2}\sigma^2(x,a)\circ\frac{\partial^2 V}{\partial x^2}(x;\pi) + r(x,a) - \beta V(x;\pi)\right)\mathrm{d}a \\
&= \tilde{r}(x,\hat{\pi}) + \mathcal{L}^{\hat{\pi}}V^\pi(x) \\
&= \tilde{r}(x,\hat{\pi}) - \tilde{r}(x,\pi) - \gamma\tilde{p}(x,\pi) + \mathcal{L}^{\hat{\pi}}V^\pi(x) - \mathcal{L}^\pi V^\pi(x) \\
&= \text{LHS}.
\end{aligned}$$

□

Proof of Theorem 2. Note that in (13), the equation to be proven, the right hand side can be written as $\int_{\mathbb{R}} d_\mu^{\hat{\pi}}(y)f(x;\pi,\hat{\pi})\mathrm{d}y$, with

$$f(x;\pi,\hat{\pi}) := \int_{\mathcal{A}} \hat{\pi}(a \mid x)\left(q(x,a;\pi) + \gamma p(x,a,\hat{\pi})\right)\mathrm{d}a.$$

From Lemma 9, we have

$$f(x;\pi,\hat{\pi}) = (\mathcal{L}^{\hat{\pi}} - \mathcal{L}^\pi)V(x;\pi) + \tilde{r}(x,\hat{\pi}) + \gamma\tilde{p}(x,\hat{\pi}) - \tilde{r}(x,\pi) - \gamma\tilde{p}(x,\pi). \tag{43}$$

On the other hand, for the left hand side of (13), we have

$$\eta(\pi) = \int_{\mathbb{R}^n} V(y;\pi)\mu(\mathrm{d}y) = \int_{\mathbb{R}^n} d_\mu^{\hat{\pi}}(y)(-\mathcal{L}^{\hat{\pi}})V(y;\pi)\mathrm{d}y, \tag{44}$$

with the second equality following from Lemma 8; and

$$\eta(\hat{\pi}) = \int_{\mathbb{R}} d_\mu^{\hat{\pi}}(y)\left[\tilde{r}(y,\hat{\pi}) + \gamma\tilde{p}(y,\hat{\pi})\right]\mathrm{d}y, \tag{45}$$

following the definition of the discounted expected occupation time; moreover, from (41), we have

$$0 = \int_{\mathbb{R}} d_\mu^{\hat{\pi}}(y)\left[(-\mathcal{L}^\pi)V(y;\pi) - \tilde{r}(y,\pi) - \gamma\tilde{p}(y,\pi)\right]\mathrm{d}y. \tag{46}$$

Hence, combining the last three equations (44,45,46), we have

$$\eta(\hat{\pi}) - \eta(\pi) = \int_{\mathbb{R}} d_\mu^{\hat{\pi}}(y)\left[(\mathcal{L}^{\hat{\pi}} - \mathcal{L}^\pi)V(y;\pi) + \tilde{r}(y,\hat{\pi}) + \gamma\tilde{p}(y,\hat{\pi}) - \tilde{r}(y,\pi) - \gamma\tilde{p}(y,\pi)\right]\mathrm{d}y. \tag{47}$$

Thus, we have shown LHS=RHS in (13). □

B.2 Proof of Theorem 3

Proof. It suffices to show the integral version of the theorem:

$$\nabla_\theta \left(\eta(\pi^\theta)\right) |_{\theta=\theta}= \int_{\mathbb{R}^n} d_\mu^{\pi^\theta}(x) \left[\int_{\mathcal{A}} \nabla_\theta \pi^\theta(a \mid x) \left(q(x, a; \pi^\theta) + \gamma p(x, a, \pi^\theta)\right) + \gamma \cdot \pi^\theta(a \mid x) \nabla_\theta p(x, a, \pi^\theta) \mathrm{d}a\right] \mathrm{d}x. \tag{48}$$

As before, we simplify notation by denoting $\eta(\pi^\theta)$ as $\eta(\theta)$ and d^{π^θ} as d^θ. Then, by Theorem 2), we have

$$\eta(\theta + \delta\theta) - \eta(\theta) = \int_{\mathbb{R}^n} d_\mu^{\theta+\delta\theta}(x) \left[\int_{\mathcal{A}} \pi^{\theta+\delta\theta}(a \mid x) \left(q(x, a; \theta) + \gamma p(x, a, \theta + \delta\theta)\right) \mathrm{d}a\right] \mathrm{d}x. \tag{49}$$

Denote

$$f(\delta\theta) = \int_{\mathcal{A}} \pi^{\theta+\delta\theta}(a \mid x) \left(q(x, a; \theta) + \gamma p(x, a, \theta + \delta\theta)\right) \mathrm{d}a.$$

Note that $f(0) = 0$, which follows from

$$\begin{aligned}
f(0) &= \int_{\mathcal{A}} \pi^\theta(a \mid x) \left(q(x, a; \theta) + \gamma p(x, a, \theta)\right) \mathrm{d}a \\
&= \int_{\mathcal{A}} \pi^\theta(a \mid x) \left(\mathcal{H}^a(x, \frac{\partial V}{\partial x}(x; \pi), \frac{\partial^2 V}{\partial x^2}(x; \pi)) - \beta V(x; \pi) + \gamma p(x, a, \theta)\right) \mathrm{d}a \\
&= -\beta V(x; \pi) + \tilde{b}(x, \pi) \cdot \nabla V(x; \pi) + \frac{1}{2}\tilde{\sigma}^2(x, \pi) \circ \nabla^2 V(x; \pi) + \tilde{r}(x, \pi) + \gamma \tilde{p}(x, \pi) \\
&= 0.
\end{aligned}$$

Thus,

$$\begin{aligned}
\eta(\theta + \delta\theta) - \eta(\theta) &= \langle d_\mu^{\theta+\delta\theta}, f(\delta\theta)\rangle \\
&= \langle d_\mu^{\theta+\delta\theta}, f(\delta\theta)\rangle - \langle d_\mu^{\theta+\delta\theta}, f(0)\rangle \\
&= \langle d_\mu^{\theta+\delta\theta}, f(\delta\theta) - f(0)\rangle \\
&= \langle d_\mu^{\theta+\delta\theta} - d_\mu^\theta, f(\delta\theta) - f(0)\rangle + \langle d_\mu^\theta, f(\delta\theta) - f(0)\rangle.
\end{aligned}$$

Dividing both sides by $\delta\theta$ completes the proof, as the first term on the last line above is of higher order than $\delta\theta$. □

B.3 Proofs of Lemma 4 and Theorem 5

We need a lemma for the perturbation bounds.

Lemma 10. *Assume that both $\tilde{\sigma}^2(x, \hat{\pi}(\cdot))$ and $\tilde{\sigma}^2(x, \pi(\cdot))$ are positive definite and*

$$\tilde{\sigma}^2(x, \pi(\cdot)), \tilde{\sigma}^2(x, \hat{\pi}(\cdot)) \geq \sigma_0^2 \cdot I.$$

where $\sigma_0 > 0$, then we have that the difference between the square root matrix is bounded by

$$\|\tilde{\sigma}(x, \hat{\pi}) - \tilde{\sigma}(x, \pi)\|_2 \leq \frac{1}{2\sigma_0}\|\tilde{\sigma}^2(x, \hat{\pi}) - \tilde{\sigma}^2(x, \pi)\|_2.$$

If we also assume that the upper bounds, i.e.

$$\tilde{\sigma}^2(x, \pi(\cdot)), \tilde{\sigma}^2(x, \hat{\pi}(\cdot)) \leq \bar{\sigma}_0^2 \cdot I.$$

by some $\bar{\sigma}_0 > \sigma_0 > 0$, then we have

$$\|\tilde{\sigma}(x, \hat{\pi}) - \tilde{\sigma}(x, \pi)\|_2 \leq \frac{\bar{\sigma}_0}{2\sigma_0}\|\hat{\pi} - \pi\|_1^{\frac{1}{2}}.$$

Proof. Consider a normalized vector x with $\|x\|_2 = 1$ is an eigenvector of $A^{\frac{1}{2}} - B^{\frac{1}{2}}$ with eigenvalue μ then

$$\begin{aligned}
x^T(A - B)x &= x^T(A^{\frac{1}{2}} - B^{\frac{1}{2}})A^{\frac{1}{2}}x + x^T B^{\frac{1}{2}}(A^{\frac{1}{2}} - B^{\frac{1}{2}})x \\
&= \mu x^T(A^{\frac{1}{2}} + B^{\frac{1}{2}})x.
\end{aligned}$$

thus, if $A, B \geq \sigma_0^2 I$, this implies

$$\mu \leq \frac{|x^T(A-B)x|}{x^T(A^{\frac{1}{2}}+B^{\frac{1}{2}})x} \leq \|A-B\|_2 \cdot \lambda_{\min}(A^{\frac{1}{2}}+B^{\frac{1}{2}})^{-1} \leq \|A-B\|_2/(2\sigma_0).$$

Furthermore, note that

$$\tilde{\sigma}^2(x,\hat{\pi}) - \tilde{\sigma}^2(x,\pi) = \int_{\mathcal{A}} \sigma^2(x,a)(\tilde{\pi}(a|x) - \pi(a|x))\mathrm{d}a.$$

so

$$\|\tilde{\sigma}^2(x,\hat{\pi}) - \tilde{\sigma}^2(x,\pi)\|_2 \leq \bar{\sigma}_0^2 \int_{\mathcal{A}} |\tilde{\pi}(a|x) - \pi(a|x)|\mathrm{d}a = \bar{\sigma}_0^2 \cdot \|\tilde{\pi}(a|x) - \pi(a|x)\|_1.$$

$\square$

Proof (of Lemma 4). Consider the Wasserstein-2 distance $W_2(\mu, v)$ between distribution μ and v as

$$W_2(\mu,\nu) = \left(\inf_{\gamma\in\Gamma(\mu,\nu)} \mathbf{E}_{(x,y)\sim\gamma}\|x-y\|_2^2\right)^{1/2},$$

where $\Gamma(\mu,\nu)$ is the set all probability measures on the product space $\mathbb{R}^n \times \mathbb{R}^n$ with the marginal distributions being μ and v, and $\|\cdot\|_2$ is the standard Euclidean distance. Denote

$$\bar{d}_\mu^\pi := \beta d_\mu^\pi.$$

We want to get an upper bound on $W_2(\bar{d}_\mu^\pi, \bar{d}_\mu^{\hat{\pi}})$ in terms of the distance between two policies π and $\hat{\pi}$. Consider a specific coupling (X_t, Y_t) below:

$$\begin{cases} \mathrm{d}X_s = \tilde{b}\left(X_s, \pi\left(\cdot \mid X_s\right)\right)\mathrm{d}s + \tilde{\sigma}\left(X_s, \pi\left(\cdot \mid X_s\right)\right)\mathrm{d}B_s, \\ \mathrm{d}Y_s = \tilde{b}\left(Y_s, \hat{\pi}\left(\cdot \mid Y_s\right)\right)\mathrm{d}s + \tilde{\sigma}\left(Y_s, \hat{\pi}\left(\cdot \mid Y_s\right)\right)\mathrm{d}B_s. \end{cases} \tag{50}$$

with $X_0 = Y_0$, which leads to a joint distribution over $\mathbb{R}^n \times \mathbb{R}^n$:

$$\tilde{\gamma} := \left\{\tilde{p}(x,y) = \int_0^\infty \frac{1}{\beta} e^{-\beta t} f_{(X_t,Y_t)}(x,y)\mathrm{d}t\right\}.$$

Hence,

$$W_2^2(\bar{d}_\mu^\pi, \bar{d}_\mu^{\hat{\pi}}) \leq \mathbb{E}_{(x,y)\sim\tilde{\gamma}}\|x-y\|_2^2 = \int_0^\infty \frac{1}{\beta} e^{-\beta s}\mathbb{E}\|X_s - Y_s\|_2^2 \mathrm{d}s. \tag{51}$$

It then boils down to estimating $\mathbb{E}\|X_s - Y_s\|_2^2$. By Itô's formula,

$$\begin{aligned} \mathrm{d}\|X_s - Y_s\|_2^2 =& 2(X_s - Y_s)^\top \left[(\tilde{b}\left(X_s,\pi\right) - \tilde{b}\left(Y_s,\hat{\pi}\right))\mathrm{d}s + (\tilde{\sigma}\left(X_s,\pi\right) - \tilde{\sigma}\left(Y_s,\hat{\pi}\right))\mathrm{d}B_s\right] \\ &+ \mathrm{Tr}\left[(\tilde{\sigma}\left(X_s,\pi\right) - \tilde{\sigma}\left(Y_s,\hat{\pi}\right))^2\right]\mathrm{d}s. \end{aligned}$$

Taking expectation on both sides yields

$$\frac{\mathrm{d}}{\mathrm{d}s}\mathbb{E}\|X_s - Y_s\|_2^2 = \underbrace{2\mathbb{E}\left[(X_s - Y_s)^\top(\tilde{b}\left(X_s,\pi\right) - \tilde{b}\left(Y_s,\hat{\pi}\right))\mathrm{d}s\right]}_{(A)} + \underbrace{\mathrm{Tr}\left[\mathbb{E}(\tilde{\sigma}\left(X_s,\pi\right) - \tilde{\sigma}\left(Y_s,\hat{\pi}\right))^2\right]}_{(B)}, \tag{52}$$

with

$$\begin{aligned} \text{(A)} &= \mathbb{E}\left[(X_s - Y_s)^\top(\tilde{b}\left(X_s,\pi\right) - \tilde{b}\left(Y_s,\pi\right))\mathrm{d}s\right] + \mathbb{E}\left[(X_s - Y_s)^\top(\tilde{b}\left(Y_s,\pi\right) - \tilde{b}\left(Y_s,\hat{\pi}\right))\mathrm{d}s\right] \\ &\leq C_{\tilde{b}} \cdot \mathbb{E}\|X_s - Y_s\|_2^2 + \frac{1}{2}\mathbb{E}\|X_s - Y_s\|_2^2 + \frac{1}{2}\mathbb{E}\|\tilde{b}\left(Y_s,\pi\right) - \tilde{b}\left(Y_s,\hat{\pi}\right)\|_2^2 \\ &\leq (C_{\tilde{b}} + \frac{1}{2}) \cdot \mathbb{E}\|X_s - Y_s\|_2^2 + \frac{1}{2}\|\tilde{b}(\cdot,\pi) - \tilde{b}(\cdot,\hat{\pi})\|_{2,\infty}^2; \end{aligned}$$

and

$$\begin{aligned} \text{(B)} &= \mathbb{E}\|\tilde{\sigma}\left(X_s,\pi\right) - \tilde{\sigma}\left(Y_s,\hat{\pi}\right)\|_F^2 \\ &\leq 2\mathbb{E}\|\tilde{\sigma}\left(X_s,\pi\right) - \tilde{\sigma}\left(Y_s,\pi\right)\|_F^2 + 2\mathbb{E}\|\tilde{\sigma}\left(Y_s,\pi\right) - \tilde{\sigma}\left(Y_s,\hat{\pi}\right)\|_F^2 \\ &\leq 2C_{\tilde{\sigma}}^2 \cdot \mathbb{E}\left\|X_s - Y_s\right\|_2^2 + 2\sup_x \|\tilde{\sigma}\left(x,\pi\right) - \tilde{\sigma}\left(x,\hat{\pi}\right)\|_F^2 \\ &:= 2C_{\tilde{\sigma}}^2 \cdot \mathbb{E}\left\|X_s - Y_s\right\|_2^2 + 2\|\tilde{\sigma}\left(\cdot,\pi\right) - \tilde{\sigma}\left(\cdot,\hat{\pi}\right)\|_{F,\infty}^2. \end{aligned}$$

Combining the above, we get

$$\frac{\mathrm{d}}{\mathrm{d}s}\mathbb{E}\|X_s - Y_s\|_2^2 \leq \underbrace{(2C_{\tilde{b}} + 1 + 2C_{\tilde{\sigma}}^2)}_{C_{\tilde{b},\tilde{\sigma}}}\mathbb{E}\|X_s - Y_s\|_2^2 + \underbrace{\|\tilde{b}(\cdot,\pi) - \tilde{b}(\cdot,\hat{\pi})\|_{2,\infty}^2 + 2\|\tilde{\sigma}(\cdot,\pi) - \tilde{\sigma}(\cdot,\hat{\pi})\|_{F,\infty}^2}_{C(\pi,\hat{\pi})}.$$

By Grönwall's inequality, we have

$$\mathbb{E}\|X_t - Y_t\|_2^2 \leq \frac{C(\pi,\hat{\pi})}{C_{\tilde{b},\tilde{\sigma}}}\left(e^{C_{\tilde{b},\tilde{\sigma}}t} - 1\right). \tag{53}$$

Substituting back into (51), we obtain

$$W_2^2(\bar{d}_\mu^\pi, \bar{d}_\mu^{\hat{\pi}}) \leq \frac{C(\pi,\hat{\pi})}{C_{\tilde{b},\tilde{\sigma}}}\int_0^\infty \frac{1}{\beta}e^{-\beta s}\left(e^{C_{\tilde{b},\tilde{\sigma}}s} - 1\right)\mathrm{d}s.$$

Thus, if $\beta > C_{\tilde{b},\tilde{\sigma}}$, we have

$$W_2(\bar{d}_\mu^\pi, \bar{d}_\mu^{\hat{\pi}}) \leq \frac{C(\pi,\hat{\pi})}{C_{\tilde{b},\tilde{\sigma}}(\beta - C_{\tilde{b},\tilde{\sigma}})\beta}.$$

Concerning the term $C(\pi,\hat{\pi})$, we have

$$\|\tilde{b}(\cdot,\pi) - \tilde{b}(\cdot,\hat{\pi})\|_{2,\infty} = \sup_x \|\tilde{b}(x,\pi) - \tilde{b}(x,\hat{\pi})\|_2 \leq \sup_x \|\hat{\pi}(\cdot|x) - \pi(\cdot|x)\|_1 \cdot \sup_{x,a}|b(x,a)|,$$

and

$$\|\tilde{\sigma}(\cdot,\pi) - \tilde{\sigma}(\cdot,\hat{\pi})\|_{F,\infty} = \sup_x \|\tilde{\sigma}(x,\pi) - \tilde{\sigma}(x,\hat{\pi})\|_F \leq \sqrt{n}\frac{\bar{\sigma}_0}{2\sigma_0}\sup_x \|\hat{\pi}(\cdot|x) - \pi(\cdot|x)\|_1^{\frac{1}{2}}.$$

Thus we have:

$$\begin{aligned} C(\pi,\hat{\pi}) &= \|\tilde{b}(\cdot,\pi) - \tilde{b}(\cdot,\hat{\pi})\|_{2,\infty}^2 + 2\|\tilde{\sigma}(\cdot,\pi) - \tilde{\sigma}(\cdot,\hat{\pi})\|_{F,\infty}^2 \\ &\leq \left(\sup_{x,a}|b(x,a)|^2 + \frac{d\cdot\bar{\sigma}_0^2}{2\sigma_0^2}\right)\max\left(\sup_x \|\hat{\pi}(\cdot|x) - \pi(\cdot|x)\|_1, \sup_x \|\hat{\pi}(\cdot|x) - \pi(\cdot|x)\|_1^{\frac{1}{2}}\right) \end{aligned}$$

which proves our upper bound. □

Proof (of Theorem 5). We have that

$$\begin{aligned} |\eta^{\hat{\pi}} - L^\pi(\hat{\pi})| = |\langle d_\mu^{\hat{\pi}} - d_\mu^\pi, f\rangle| &= \frac{\|f\|_{\dot{H}^1}}{\beta}\left|\left\langle \bar{d}_\mu^{\hat{\pi}} - \bar{d}_\mu^\pi, \frac{f}{\|f\|_{\dot{H}^1}}\right\rangle\right| \\ &\leq \frac{K}{\beta}\|\bar{d}_\mu^{\hat{\pi}} - \bar{d}_\mu^\pi\|_{\dot{H}^{-1}} \leq \frac{K\sqrt{M}}{\beta}W_2\left(\bar{d}_\mu^{\hat{\pi}}, \bar{d}_\mu^\pi\right). \end{aligned} \tag{54}$$

where $K := \sup_{\hat{\pi}}\|f\|_{\dot{H}^1} < \infty$ (more about K in the remarks below). Combining (54) with the estimate in (22) (of Lemma 4) yields the desired result in (23). [1] □

Remarks (on K). In the performance-difference bound developed above, we assume K is finite:

$$K := \|f\|_{\dot{H}^1} := \left(\int_{\mathbb{R}^n}|\nabla f(x)|^2\mathrm{d}x\right)^{\frac{1}{2}} < \infty,$$

where $f(x;\pi,\hat{\pi}) := \int_{\mathcal{A}}\hat{\pi}(a \mid x)\left(q(x,a;\pi) + p(x,a,\hat{\pi})\right)\mathrm{d}a$. The famous Poincaré inequality can provide a lower bound on this quantity; but we need an upper bound as well, i.e.,

$$K = \left(\int_{\mathbb{R}^n}|\nabla f(x)|^2\mathrm{d}x\right)^{\frac{1}{2}} \leq C\left(\int_{\mathbb{R}^n}|f(x)|^2\mathrm{d}x\right)^{\frac{1}{2}}.$$

This above is essentially a *reverse* Poincaré Inequality, which is not likely to hold (in particular, the existence of the constant C).

[1] From this proof, it's evident that there's a β missing in the denominator on the RHS of (22). Consequently, the $C(\mu,\pi,\hat{\pi})$ expression in Theorem 5 should have $2\beta^2$ (instead of 2β) in the denominator. This correction will *not* affect the two numerical examples as both had set $\beta = 1$ (as a hyper-parameter).

Should we indeed have a reverse Poincaré Inequality, then we can further bound f by

$$
\begin{aligned}
|f(x)| &= |\int_{\mathcal{A}} \left(\hat{\pi}(a \mid x) - \pi(a \mid x)\right) \left(q(x, a; \pi) + p(x, a, \hat{\pi})\right) \mathrm{d}a| \\
&\leq \int_{\mathcal{A}} |\hat{\pi}(a \mid x) - \pi(a \mid x)| \cdot |q(x, a; \pi) + p(x, a, \hat{\pi})| \,\mathrm{d}a \\
&\leq 2 \sup_a |q(x, a; \pi) + p(x, a, \hat{\pi})| \, D_{\mathrm{TV}}(\pi(\cdot \mid x), \hat{\pi}(\cdot \mid x)),
\end{aligned}
$$

and

$$
\begin{aligned}
\left(\int_{\mathbb{R}^n} |f(x)|^2 \mathrm{d}x\right)^{\frac{1}{2}} &\leq \left(\int_{\mathbb{R}^n} 4 \sup_a |q(x, a; \pi) + p(x, a, \hat{\pi})|^2 \, D^2_{\mathrm{TV}}(\pi(\cdot \mid x), \hat{\pi}(\cdot \mid x)) \mathrm{d}x\right)^{\frac{1}{2}} \\
&\leq \left(\int_{\mathbb{R}^n} 2 \sup_a |q(x, a; \pi) + p(x, a, \hat{\pi})|^2 \,\mathrm{d}x\right)^{\frac{1}{2}} \sqrt{\sup_x D_{\mathrm{KL}}(\pi(\cdot \mid x), \hat{\pi}(\cdot \mid x))},
\end{aligned}
$$

where the second inequality is from Pinsker's inequality. This way, we would have recovered a similar bound as in the discrete RL. Since we do not have the reverse Poincaré inequality, however, we have to assume that K is finite.

Appendix C Algorithms

C.1 Performance of CPPO with Square-root KL and Linear KL

Here we present a detailed version of the CPPO algorithm. For two probability distributions P and Q over the action space with density functions p and q correspondingly, the KL-divergence between these two is defined as:

$$D_{\mathrm{KL}}(P\|Q)=\int_{\mathcal{A}}\log(\frac{q(a)}{p(a)})q(a)\mathrm{d}a,$$

Denote $D_{\mathrm{KL}}(\theta,\theta_k) := \mathbb{E}_{x\sim d_\mu^{\theta_k}} D_{\mathrm{KL}}(\pi_\theta(\cdot|x)\|\pi_{\theta_k}(\cdot|x))$, to distinguish it from $\bar{D}_{\mathrm{KL}}(\theta\|\theta_k) := \mathbb{E}_{x\sim d_\mu^{\theta_k}} \sqrt{D_{\mathrm{KL}}(\pi_\theta(\cdot|x)\|\pi_{\theta_k}(\cdot|x))}$ which was used in CPPO Algorithm in 2.

Note that bounding the performance difference by the linear KL-divergence $D_{\mathrm{KL}}(\theta,\theta_k)$, instead of its square-root counterpart $\bar{D}_{\mathrm{KL}}(\theta\|\theta_k)$, will generally require stronger conditions (which may be difficult to satisfy). For completeness, we present the following algorithm, the CPPO with linear KL-divergence:

Algorithm 3 CPPO: PPO with adaptive penalty constant (linear KL-divergence)

Input: Policy parameters θ_0, critic net parameters ϕ_0

for $k=0,1,2,\cdots$ until θ_k converge **do**

  Collect a truncated trajectory $\{X_{t_i},a_{t_i},r_{t_i},p_{t_i}\}, i=1,\dots,N$ from the environment using π_{θ_k}.

  for $i=0,\dots,N-1$ **do**: Update the critic parameters as in (8)

  for $j=1,,\dots,J$ **do**: Draw i.i.d. τ_j from $\exp(\beta)$, round τ_j to the largest multiple of δ_t no larger than it, and compute the GAE estimator of $q(X_{\tau_j},a_{\tau_j})$

$$\tilde{q}(X_{\tau_j},a_{\tau_j}) := \left(r_{\tau_j}\delta_t + e^{-\beta\delta_t}V(X_{\tau_j+\delta_t}) - V(X_{\tau_j})\right)/\delta_t.$$

  Compute policy update (by taking a fixed s steps of gradient descent)

$$\theta_{k+1}=\arg\max_\theta L^{\theta_k}(\theta)-C^k_{\text{penalty}}D_{\mathrm{KL}}\left(\theta,\theta_k\right).$$

  if $D_{\mathrm{KL}}\left(\theta_{k+1},\theta_k\right)\geq(1+\epsilon)\delta$, **then** $C^{k+1}_{\text{penalty}}=2C^k_{\text{penalty}}$.

  else if $D_{\mathrm{KL}}\left(\theta_{k+1},\theta_k\right)\leq\delta/(1+\epsilon)$, **then** $C^{k+1}_{\text{penalty}}=C^k_{\text{penalty}}/2$.

A comparison between the above and Algorithm 2 (using square-root KL divergence) is presented in §D.3 below, which clearly illustrates the advantage of square-root KL divergence.

C.2 KL-divergence

We elaborate here on the KL-divergence between the current policy and the optimal policy, along with the entropy regularizer. By the performance difference formula, we have

$$\eta(\pi)-\eta(\pi^*)=\int_{\mathbb{R}^n}d^\pi_\mu(x)\left[\int_{\mathcal{A}}\pi(a\mid x)\left(q(x,a;\pi^*)-\gamma\log(\pi(a))\right)\mathrm{d}a\right]\mathrm{d}x.$$

Notice that by the definition of KL-divergence we defined before, we have

$$D_{\mathrm{KL}}(\pi^*(\cdot|x)\|\pi(\cdot|x))=\int_{\mathcal{A}}\log(\frac{\pi(a|x)}{\pi^*(a|x)})\pi(a|x)\mathrm{d}a.$$

Similar as the previous discussion of soft q-learning, π^* is optimal implies that

$$\pi^*(a\mid x)\propto\exp(\frac{q(x,a,\pi^*)}{\gamma}),$$

and the normalization constant is 1 can be proved through considering the exploratory HJB equation, see [22, 56]. Thus

$$D_{\mathrm{KL}}(\pi^*(\cdot|x)\|\pi(\cdot|x))=\int_{\mathcal{A}}\log(\pi(a|x))\pi(a|x)\mathrm{d}a-\int_{\mathcal{A}}\frac{q(x,a,\pi^*)}{\gamma}\pi(a|x)\mathrm{d}a,$$

which leads to

$$\eta(\pi) - \eta(\pi^*) = -\gamma \cdot \mathbb{E}_{x\sim d_\mu^\pi} D_{\mathrm{KL}}(\pi^*(\cdot|x) \| \pi(\cdot|x)).$$

This justifies our claim that the KL-divergence is essentially equivalent to the distance to the optimal performance.

Appendix D Experiments

D.1 Example 1

Recall, in the LQ control problem, the reward function is

$$r(x,a) = -\left(\frac{M}{2}x^2 + Rxa + \frac{N}{2}a^2 + Px + Qa\right),$$

with $M \geq 0, N > 0, R, Q, P \in \mathbb{R}$ and $R^2 < MN$, and we adopt the entropy regularizor as

$$p(x,a,\pi) = -\log(\pi(a)).$$

Furthermore, suppose that the discount rate satisfies $\beta > 2A + C^2 + \max\left(\frac{D^2R^2 - 2NR(B+CD)}{N}, 0\right)$.

The following results are readily derived from Theorem 4 of [58]. The value function of the optimal policy π^* is

$$V(x) = \frac{1}{2}k_2x^2 + k_1x + k_0, \quad x \in \mathbb{R},$$

where

$$k_2 := \frac{1}{2}\frac{\left(\rho - \left(2A + C^2\right)\right)N + 2(B+CD)R - D^2M}{(B+CD)^2 + \left(\rho - (2A+C^2)\right)D^2}$$

$$-\frac{1}{2}\frac{\sqrt{\left(\left(\rho - (2A+C^2)\right)N + 2(B+CD)R - D^2M\right)^2 - 4\left((B+CD)^2 + \left(\rho - (2A+C^2)\right)D^2\right)\left(R^2 - MN\right)}}{(B+CD)^2 + \left(\rho - (2A+C^2)\right)D^2},$$

$$k_1 := \frac{P\left(N - k_2D^2\right) - QR}{k_2B(B+CD) + (A-\rho)\left(N - k_2D^2\right) - BR},$$

and

$$k_0 := \frac{\left(k_1B - Q\right)^2}{2\rho\left(N - k_2D^2\right)} + \frac{\gamma}{2\rho}\left(\ln\left(\frac{2\pi e\gamma}{N - k_2D^2}\right) - 1\right)$$

respectively. Moreover, the optimal feedback control is Gaussian, with density function

$$\pi^*(a;x) = \mathcal{N}\left(a \mid \frac{\left(k_2(B+CD) - R\right)x + k_1B - Q}{N - k_2D^2}, \frac{\gamma}{N - k_2D^2}\right).$$

For a set of model parameters: $A = -1, B = C = 0, D = 1, M = N = Q = 2, R = P = 1, \beta = 1, \gamma = 0.1$, following the formulas and the parameterized policy $\pi_\theta(\cdot \mid x) = \mathcal{N}(\theta_1x + \theta_2, \exp(\theta_3))$, and the corresponding value function $V_\phi(x) = \frac{1}{2}\phi_2x^2 + \phi_1x + \phi_0$, we can derive the optimal parameters:

$$\phi^* = [0.71914874, -0.10555128, -0.53518376],$$

and

$$\theta^* = [-0.39444872, -0.78889745, -1.40400944].$$

Table 1: Hyper-parameter values for Example 1

Alphabet	Description	Value
T	Trajectory Truncation Length	25
β	discount factor	1
δ_t	time interval	0.005
J	batch size for sampling $\exp(\beta)$	100
α_1	learning rate for policy iteration k	0.02 when $k \leq 50$ and $0.02 \times \log(\frac{50}{k})$ when $k > 50$
α_2	learning rate for value iteration k	0.01 when $k \leq 50$ and $0.01 \times \log(\frac{50}{k})$ when $k > 50$
K	iteration threshold	2000
s	steps of gradient descent	10
δ	radius	0.0002
ϵ	tolerance level	0.5

D.2 Example 2

The model parameters are $k = 0.01, \theta = 7, \eta = 0.1, \rho = 0.3, \sigma = 1, r_f = 0.01, \ell = 5$. For both the value function and the policy parameterization, we use a 3-layer neural network, and with the initial parameters sampled form the uniform distribution over [-0.5,0.5]. We use the tanh activation function for the hidden layer.

Table 2: Hyperparameter values for Example 2

Alphabet	Description	Value
T	Trajectory Truncation Length	25
β	discount factor	1
δ_t	time interval	0.005
J	batch size for sampling $\exp(\beta)$	100
α_1	learning rate for policy iteration k	0.005 when $k \leq 50$ and $0.005 \times \log(\frac{50}{k})$ when $k > 50$
α_2	learning rate for value iteration k	0.01 when $k \leq 50$ and $0.01 \times \log(\frac{50}{k})$ when $k > 50$
K	iteration threshold	200
s	steps of gradient descent	10
δ	radius	0.025
ϵ	tolerance level	0.5

D.3 Performance of CPPO with Square-root KL and Linear KL

We compare the performance of CPPO with square-root KL-divergence (denote as CPPO), and linear KL-divergence (denoted as CPPO (nst) — non square-root) applied to the experiments in Example 1 and Example 2. Figure 4 compares the distance between the current policy parameters and the

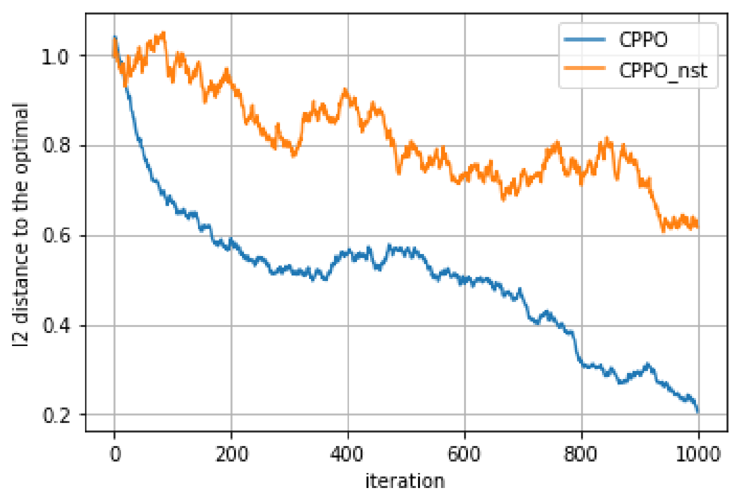

Figure 4: Performance of CPPO and CPPO (nst) to the Example 1

optimal parameters, with x-axis denoting the iteration times and y-axis denoting the L_2 distance. Figure 5 compares the current expected return, with x-axis denoting the iteration times and y-axis denoting the current performance by taking the average of 100 times of Monte Carlo evaluation. In both figures, the blue curve represents the algorithm with square-root KL-divergence as opposed to the orange one corresponding to the linear version. Both figures clearly demonstrate the advantage of the former. In particular, the linear version can suffer from getting stuck at the local optimum as demonstrated in Example 1.

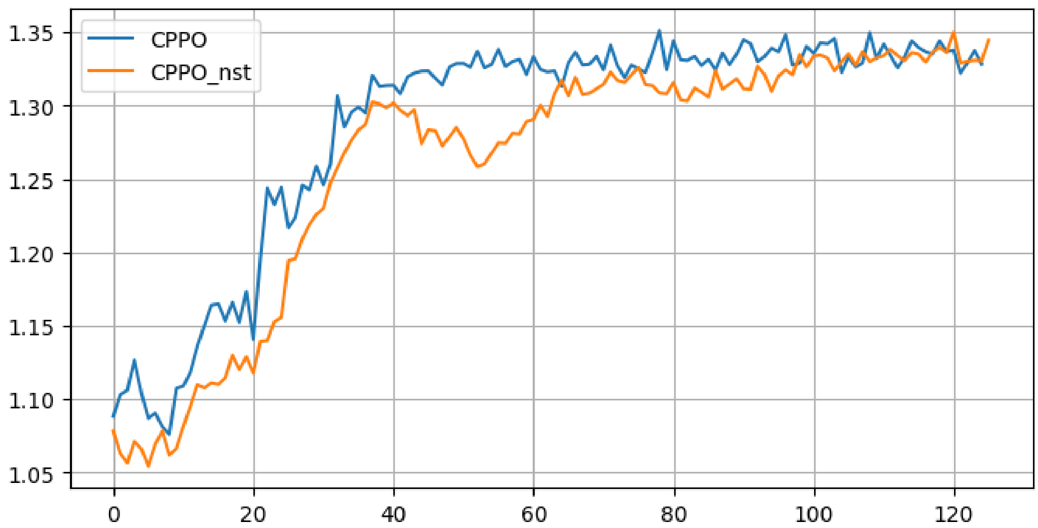

Figure 5: Performance of CPPO and CPPO (nst) to the Example 2