An Online Policy Gradient Algorithm for Markov Decision Processes with Continuous States and Actions

Yao Ma[1], Tingting Zhao[1], Kohei Hatano[2], and Masashi Sugiyama[1]

[1] Tokyo Institute of Technology,
2-12-1 O-okayama, Meguro, Tokyo 152-8552, Japan
{yao@sg.,tingting@sg.,sg@}cs.titech.ac.jp
[2] Kyushu University,
744 Motooka, Nishi, Fukuoka, 819-0395, Japan
hatano@inf.kyushu-u.ac.jp

Abstract. We consider the learning problem under an online Markov decision process (MDP), which is aimed at learning the time-dependent decision-making policy of an agent that minimizes the regret — the difference from the best fixed policy. The difficulty of online MDP learning is that the reward function changes over time. In this paper, we show that a simple online policy gradient algorithm achieves regret $O(\sqrt{T})$ for T steps under a certain concavity assumption and $O(\log T)$ under a strong concavity assumption. To the best of our knowledge, this is the first work to give an online MDP algorithm that can handle continuous state, action, and parameter spaces with guarantee. We also illustrate the behavior of the online policy gradient method through experiments.

Keywords: Markov decision process, Online learning.

1 Introduction

The *Markov decision process* (MDP) is a popular framework of reinforcement learning for sequential decision making [6], where an agent takes an action depending on the current state, moves to the next state, receives a reward based on the last transition, and this process is repeated T times. The goal is to find an optimal decision-making policy (i.e., a conditional probability density of action given state) that maximizes the expected sum of rewards over T steps.

In the standard MDP formulation, the reward function is fixed over iterations. On the other hand, in this paper, we consider an online MDP scenario where the reward function changes over time — it can be altered even adversarially. The goal is to find the best *time-dependent* policy that minimizes the *regret*, the difference from the best fixed policy. We expect the regret to be $o(T)$, by which the difference from the best fixed policy vanishes as T goes to infinity.

The *MDP expert* algorithm (MDP-E), which chooses the current best action at each state, was shown to achieve regret $O(\sqrt{T \log |A|})$ [1,2], where $|A|$

T. Calders et al. (Eds.): ECML PKDD 2014, Part II, LNCS 8725, pp. 354–369, 2014.

denotes the cardinality of the action space. Although this bound does not explicitly depend on the cardinality of the state space, the algorithm itself needs an expert algorithm for each state. Another algorithm called the *lazy follow-the-perturbed-leader* (lazy-FPL) divides the time steps into short periods and policies are updated only at the end of each period using the average reward function [8]. This lazy-FPL algorithm was shown to have regret $O(T^{3/4+\epsilon} \log T(|S| + |A|)|A|^2)$ for $\epsilon \in (0, 1/3)$. The online MDP algorithm called the *online relative entropy policy search* is considered in [9], which was shown to have regret $O(L^2\sqrt{T \log(|S||A|/L)})$ for state space with L-layered structure. However, the regret bounds of these algorithms explicitly depend on $|S|$ and $|A|$, and the algorithms cannot be directly implemented for problems with continuous state and action spaces. The *online algorithm for Markov decision processes* was shown to have regret $O(\sqrt{T \log |\Pi|} + \log |\Pi|)$ with changing transition probability distributions, where $|\Pi|$ it the cardinality of the policy set [11]. Although sub-linear bounds still hold for continuous policy spaces, the algorithm cannot be used with infinite policy candidates directly.

In this paper, we propose a simple *online policy gradient* (OPG) algorithm that can be implemented in a straightforward manner for problems with continuous state and action spaces[1]. Under the assumption that the expected average reward function is concave, we prove that the regret of our OPG algorithm is $O(\sqrt{T}(F^2 + N))$, which is independent of the cardinality of the state and action spaces, but is dependent on the diameter F and dimension N of the parameter space. Furthermore, regret $O(N^2 \log T)$ is also proved under a strongly concavity assumption on the expected average reward function. We numerically illustrate the superior behavior of the proposed OPG in continuous problems over MDP-E with different discretization schemes.

2 Online Markov Decision Process

In this section, we formulate the problem of online MDP learning.

An online MDP is specified by

- State space $\boldsymbol{s} \in S$, which could be either continuous or discrete.
- Action space $\boldsymbol{a} \in A$, which could be either continuous or discrete.
- Transition density $p(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a})$, which represents the conditional probability density of next state $\boldsymbol{s}'$ given current state $\boldsymbol{s}$ and action $\boldsymbol{a}$ to be taken.
- Reward function sequence $r_1, r_2, \ldots, r_T$, which are fixed in advance and will not change no matter what action is taken.

An online MDP algorithm produces a stochastic policy $\pi(\boldsymbol{a}|\boldsymbol{s}, t)$[2], which is a conditional probability density of action $\boldsymbol{a}$ to be taken given current state $\boldsymbol{s}$ at

[1] Our OPG algorithm can also be seen as an extension of the *online gradient descent* algorithm [10] to online MDPs problems, by decomposing the objective function.

[2] The stochastic policy incorporates exploratory actions, and exploration is usually required for getting a better policy in the learning process.

time step t. In other words, an online MDP algorithm $\mathcal{A}$ outputs parameter $\boldsymbol{\theta} = [\theta^{(1)}, \ldots, \theta^{(N)}]^\top \in \Theta \subset \mathbb{R}^N$ of stochastic policy $\pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta})$.

Thus, algorithm $\mathcal{A}$ gives a sequence of policies:

$$\pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}_1), \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}_2), \ldots, \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}_T).$$

We denote the expected cumulative rewards over T time steps of algorithm $\mathcal{A}$ by

$$R_{\mathcal{A}}(T) = \mathbb{E}\left[\sum_{t=1}^{T} r_t(\boldsymbol{s}_t, \boldsymbol{a}_t)\Big|\mathcal{A}\right].$$

Suppose that there exists $\boldsymbol{\theta}^*$ such that policy $\pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}^*)$ maximizes the expected cumulative rewards:

$$R_{\boldsymbol{\theta}^*}(T) = \mathbb{E}\left[\sum_{t=1}^{T} r_t(\boldsymbol{s}_t, \boldsymbol{a}_t)\Big|\boldsymbol{\theta}^*\right] = \sup_{\boldsymbol{\theta}\in\Theta} \mathbb{E}\left[\sum_{t=1}^{T} r_t(\boldsymbol{s}_t, \boldsymbol{a}_t)\Big|\boldsymbol{\theta}\right],$$

where $\mathbb{E}$ denotes the expectation. Our goal is to design algorithm $\mathcal{A}$ that minimizes the *regret* against the best offline policy defined by

$$L_{\mathcal{A}}(T) = R_{\boldsymbol{\theta}^*}(T) - R_{\mathcal{A}}(T).$$

If the regret is bounded by a sub-linear function with respect to T, the algorithm $\mathcal{A}$ is shown to be asymptotically as powerful as the best offline policy.

3 Online Policy Gradient (OPG) Algorithm

In this section, we introduce an online policy gradient algorithm for solving the online MDP problem.

Different from the previous works, we do not use the expert algorithm in our method, because it is not suitable to handling continuous state and action problems. Instead, we consider a gradient-based algorithm which updates the parameter of policy $\boldsymbol{\theta}$ along the gradient direction of the expected average reward function at time step t.

More specifically, we assume that the target MDP $\{S, A, p, \pi, r\}$ is *ergodic*. Then it has the unique stationary state distribution $d_{\boldsymbol{\theta}}(\boldsymbol{s})$:

$$d_{\boldsymbol{\theta}}(\boldsymbol{s}) = \lim_{T\to\infty} p(\boldsymbol{s}_T = \boldsymbol{s}|\boldsymbol{\theta}).$$

Note that the stationary state distribution satisfies

$$d_{\boldsymbol{\theta}}(\boldsymbol{s}') = \int_{\boldsymbol{s}\in S} d_{\boldsymbol{\theta}}(\boldsymbol{s}) \int_{\boldsymbol{a}\in A} \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}) p(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) \mathrm{d}\boldsymbol{a}\mathrm{d}\boldsymbol{s}.$$

Let $\rho_t(\boldsymbol{\theta})$ be the expected average reward function of policy $\pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta})$ at time step t:

$$\rho_t(\boldsymbol{\theta}) = \int_{\boldsymbol{s}\in S} d_{\boldsymbol{\theta}}(\boldsymbol{s}) \int_{\boldsymbol{a}\in A} r_t(\boldsymbol{s},\boldsymbol{a}) \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}) \mathrm{d}\boldsymbol{a}\mathrm{d}\boldsymbol{s}. \tag{1}$$

Then our *online policy gradient (OPG) algorithm* is given as follows:

– Initialize policy parameter $\boldsymbol{\theta}_1$.
– for $t = 1$ to ∞
  1. Observe current state $\boldsymbol{s}_t = \boldsymbol{s}$.
  2. Take action $\boldsymbol{a}_t = \boldsymbol{a}$ according to current policy $\pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}_t)$.
  3. Observe reward r_t from the environment.
  4. Move to next state $\boldsymbol{s}_{t+1}$.
  5. Update the policy parameter as

$$\boldsymbol{\theta}_{t+1} = P\left(\boldsymbol{\theta}_t + \eta_t \nabla_{\boldsymbol{\theta}} \rho_t(\boldsymbol{\theta}_t)\right), \tag{2}$$

  where $P(\boldsymbol{\vartheta}) = \arg\min_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{\vartheta} - \boldsymbol{\theta}\|$ is the projection function, $\eta_t = \frac{1}{\sqrt{t}}$ is the step size, and $\nabla_{\boldsymbol{\theta}} \rho_t(\boldsymbol{\theta})$ is the gradient of $\rho_t(\boldsymbol{\theta})$:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} \rho_t(\boldsymbol{\theta}) &\equiv \left[\frac{\partial \rho_t(\boldsymbol{\theta})}{\partial \theta^{(1)}}, \ldots, \frac{\partial \rho_t(\boldsymbol{\theta})}{\partial \theta^{(N)}}\right]^\top \\ &= \int_{\boldsymbol{s}\in S} \int_{\boldsymbol{a}\in A} d_{\boldsymbol{\theta}}(\boldsymbol{s}) \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}) (\nabla_{\boldsymbol{\theta}} \ln d_{\boldsymbol{\theta}}(\boldsymbol{s}) + \nabla_{\boldsymbol{\theta}} \ln \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta})) \\ &\qquad \times r_t(\boldsymbol{s},\boldsymbol{a}) \mathrm{d}\boldsymbol{a}\mathrm{d}\boldsymbol{s}.\end{aligned}$$

If it is time-consuming to obtain the exact stationary state distribution, gradients estimated by a reinforcement learning algorithm may be used instead in practice.

When the reward function does not changed over time, the OPG algorithm is reduced to the ordinary policy gradient algorithm [7], which is an efficient and natural algorithm for continuous state and action MDPs. The OPG algorithm can also be regarded as an extension of the *online gradient descend* algorithm [10], which maximizes $\sum_{t=1}^{T} \rho_t(\boldsymbol{\theta}_t)$, not $\mathbb{E}\left[\sum_{t=1}^{T} r_t(\boldsymbol{s}_t, \boldsymbol{a}_t)|\mathcal{A}\right]$. As we will prove in Section 4, the regret bound of the OPG algorithm is $O(\sqrt{T})$ under a certain concavity assumption and $O(\log T)$ under a strong concavity assumption. Unlike previous works, this bound does not depend on the cardinality of state and action spaces. Therefore, the OPG algorithm would be suitable to handling continuous state and action online MDPs.

4 Regret Analysis under Concavity

In this section, we provide a regret bound for the OPG algorithm.

4.1 Assumptions

First, we introduce the assumptions required in the proofs. Some assumptions have already been used in related works for discrete state and action MDPs, and we extend them to continuous state and action MDPs.

Assumption 1. *For two arbitrary distributions d and d' over S and for every policy parameter $\boldsymbol{\theta}$, there exists a positive number τ such that*

$$\int_{\boldsymbol{s}\in S}\int_{\boldsymbol{s}'\in S}|d(\boldsymbol{s})-d'(\boldsymbol{s})|p(\boldsymbol{s}'|\boldsymbol{s};\boldsymbol{\theta})\mathrm{d}\boldsymbol{s}'\mathrm{d}\boldsymbol{s}\leq e^{-1/\tau}\int_{\boldsymbol{s}\in S}|d(\boldsymbol{s})-d'(\boldsymbol{s})|\mathrm{d}\boldsymbol{s},$$

where

$$p(\boldsymbol{s}'|\boldsymbol{s};\boldsymbol{\theta})=\int_{\boldsymbol{a}\in A}\pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta})p(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\mathrm{d}\boldsymbol{a},$$

and τ is called the mixing time *[1,2].*

Assumption 2. *For two arbitrary policy parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ and for every $\boldsymbol{s}\in S$, there exists a constant $C_1>0$ depending on the specific policy model π such that*

$$\int_{\boldsymbol{a}\in A}|\pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta})-\pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}')|\mathrm{d}\boldsymbol{a}\leq C_1\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_1.$$

The Gaussian policy is a common choice in continuous state and action MDPs. Below, we consider the Gaussian policy with mean $\mu(\boldsymbol{s})=\boldsymbol{\theta}^\top\phi(\boldsymbol{s})$ and standard deviation σ, where $\boldsymbol{\theta}$ is the policy parameter and $\phi(\boldsymbol{s}):S\to\mathbb{R}^N$ is the basis function. The KL-divergence between these two policies is

$$\begin{aligned}D(p(\cdot|s;\theta)||p(\cdot|s;\theta'))&=\int_{a\in A}\mathcal{N}_{\theta,\sigma}(a)\left\{\log\mathcal{N}_{\theta,\sigma}(a)-\log\mathcal{N}_{\theta',\sigma}(a)\right\}\mathrm{d}a\\&=\int_{a\in A}\mathcal{N}_{\theta,\sigma}(a)\left\{\frac{1}{2\sigma^2}\left(-(a-\theta)^2+(a-\theta')^2\right)\right\}\mathrm{d}a\\&=\frac{\|\phi(\boldsymbol{s})\|_\infty}{2\sigma}\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|^2.\end{aligned}$$

By Pinsker's inequality, the following inequality holds:

$$\|p(\cdot|s,\theta)-p(\cdot|s,\theta')\|_1\leq\frac{\|\phi(\boldsymbol{s})\|_\infty}{\sigma}\|\theta-\theta'\|_1. \tag{3}$$

This implies that the Gaussian policy model satisfies Assumption 2 with $C_1=\frac{\|\phi(\boldsymbol{s})\|_\infty}{\sigma}$. Note that we do not specify any policy model in the analysis, and therefore other stochastic policy models could also be used in our algorithm.

Assumption 3. *All the reward functions in online MDPs are bounded. For simplicity, we assume that the reward functions satisfy*

$$r_t(\boldsymbol{s},\boldsymbol{a})\in[0,1],\forall\boldsymbol{s}\in S,\forall\boldsymbol{a}\in A,\forall t=1,\ldots,T.$$

Assumption 4. *For all $t=1,\ldots,T$, the second derivative of the expected average reward function satisfies*

$$\nabla^2_\theta\rho_t(\boldsymbol{\theta})\leq 0. \tag{4}$$

This assumption means that the expected average reward function is concave, which is currently our sufficient condition to guarantee the $O(\sqrt{T})$-regret bound for the OPG algorithm.

4.2 Regret Bound

We have the following theorem.

Theorem 1. *The regret against the best offline policy of the OPG algorithm is bounded as*

$$L_{\mathcal{A}}(T) \leq \sqrt{T}\frac{F^2}{2} + \sqrt{T}C_2N + 2\sqrt{T}\tau^2 C_1 C_2 N + 4\tau,$$

where F is the diameter of Θ and $C_2 = \frac{2C_1 - C_1 e^{-1/\tau}}{1 - e^{-1/\tau}}$.

To prove the above theorem, we decompose the regret in the same way as the previous work [1,2,3,4]:

$$\begin{aligned} L_{\mathcal{A}}(T) =& R_{\boldsymbol{\theta}^*}(T) - R_{\mathcal{A}}(T) \\ \leq& \left(R_{\boldsymbol{\theta}^*}(T) - \sum_{t=1}^{T} \rho_t(\boldsymbol{\theta}^*) \right) + \left(\sum_{t=1}^{T} \rho_t(\boldsymbol{\theta}^*) - \sum_{t=1}^{T} \rho_t(\boldsymbol{\theta}_t) \right) \\ &+ \left(\sum_{t=1}^{T} \rho_t(\boldsymbol{\theta}_t) - R_{\mathcal{A}}(T) \right). \end{aligned} \tag{5}$$

In the OPG method, $\rho_t(\boldsymbol{\theta})$ is used for optimization, and the expected average reward is calculated by the stationary state distribution $d_{\boldsymbol{\theta}}(\boldsymbol{s})$ of the policy parameterized by $\boldsymbol{\theta}$. However, the expected reward at time step t is calculated by $d_{\boldsymbol{\theta},t}$, which is the state distribution at time step t following policy $\pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta})$. This difference affects the first and third terms of the decomposed regret (5).

Below, we bound each of the three terms in Lemma 1, Lemma 2, and Lemma 3, which are proved later.

Lemma 1.

$$\left| R_{\boldsymbol{\theta}^*}(T) - \sum_{t=1}^{T} \rho_t(\boldsymbol{\theta}^*) \right| \leq 2\tau.$$

The first term has already been analyzed for discrete state and action online MDPs in [1,2], and we extended it to continuous state and action spaces in Lemma 1.

Lemma 2. *The expected average reward function satisfies*

$$\left| \sum_{t=1}^{T} (\rho_t(\boldsymbol{\theta}^*) - \rho_t(\boldsymbol{\theta}_t)) \right| \leq \sqrt{T}\frac{F^2}{2} + \sqrt{T}C_2N.$$

Lemma 2 is obtained by using the result of [10].

Lemma 3.

$$\left| R_{\mathcal{A}}(T) - \sum_{t=1}^{T} \rho_t(\boldsymbol{\theta}_t) \right| \leq 2\tau^2 C_1 C_2 N\sqrt{T} + 2\tau.$$

Lemma 3 is similar to Lemma 5.2 in [2], but our bound does not depend on the cardinality of state and action spaces.

Combining Lemma 1, Lemma 2, and Lemma 3, we can immediately obtain Theorem 1.

If the reward function is strongly concave for all $t = 1, \ldots, T$, the bound of the OPG algorithm is $O(\log T)$ which is proved in Section 5.

4.3 Proof of Lemma 1

The following proposition holds, which can be obtained by recursively using Assumption 1:

Proposition 1. *For any policy parameter $\boldsymbol{\theta}$, the state distribution $d_{\boldsymbol{\theta},t}$ at time t and stationary state distribution $d_{\boldsymbol{\theta}}$ satisfy*

$$\int_{\boldsymbol{s}\in S} |d_{\theta,t}(\boldsymbol{s}) - d_{\theta}(\boldsymbol{s})| \mathrm{d}s \leq 2e^{-t/\tau}.$$

The first part of the regret bound in Theorem 1 is caused by the difference between the state distribution at time t and the stationary state distribution following the best offline policy parameter $\boldsymbol{\theta}^*$.

$$\begin{aligned}
\left| R_{\boldsymbol{\theta}^*}(T) - \sum_{t=1}^{T} \rho_t(\boldsymbol{\theta}^*) \right| &= \left| \sum_{t=1}^{T} \left[\int_{\boldsymbol{s}\in S} d_{\theta^*,t}(\boldsymbol{s}) \int_{\boldsymbol{a}\in A} r_t(\boldsymbol{s},\boldsymbol{a}) \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}^*) \mathrm{d}\boldsymbol{s}\mathrm{d}\boldsymbol{a} \right.\right. \\
&\quad \left.\left. - \int_{\boldsymbol{s}\in S} d_{\theta^*}(\boldsymbol{s}) \int_{\boldsymbol{a}\in A} r_t(\boldsymbol{s},\boldsymbol{a}) \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}^*) \mathrm{d}\boldsymbol{s}\mathrm{d}\boldsymbol{a} \right] \right| \\
&\leq \sum_{t=1}^{T} \int_{s\in S} |d_{\theta^*,t}(\boldsymbol{s}) - d_{\theta^*}(\boldsymbol{s})| \, \mathrm{d}\boldsymbol{s} \\
&\leq 2 \sum_{t=1}^{T} e^{-t/\tau} \\
&\leq 2\tau,
\end{aligned}$$

which concludes the proof.

4.4 Proof of Lemma 2

The following proposition is a continuous extension of Lemma 6.3 in [2]:

Proposition 2. *For two policies with different parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, an arbitrary distribution d over S, and the constant $C_1 > 0$ given in Assumption 2, it holds that*

$$\int_{\boldsymbol{s}\in S} d(\boldsymbol{s}) \int_{\boldsymbol{s}'\in S} |p(\boldsymbol{s}'|\boldsymbol{s};\boldsymbol{\theta}) - p(\boldsymbol{s}'|\boldsymbol{s};\boldsymbol{\theta}')| \mathrm{d}\boldsymbol{s}'\mathrm{d}\boldsymbol{s} \leq C_1 \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_1,$$

where

$$p(\boldsymbol{s}'|\boldsymbol{s};\boldsymbol{\theta}) = \int_{\boldsymbol{a}\in A} \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}) p(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) \mathrm{d}\boldsymbol{a}.$$

Then we have the following proposition, which is proved in Section 4.6:

Proposition 3. *For all $t = 1, \ldots, T$, the expected average reward function $\rho_t(\boldsymbol{\theta})$ for two different parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ satisfies*

$$|\rho_t(\boldsymbol{\theta}) - \rho_t(\boldsymbol{\theta}')| \leq C_2 \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_1.$$

From Proposition 3, we have the following proposition:

Proposition 4. *Let*

$$\boldsymbol{\theta} = [\theta^{(1)}, \ldots, \theta^{(i)}, \ldots, \theta^{(N)}],$$
$$\boldsymbol{\theta}' = [\theta^{(1)}, \ldots, \theta^{(i)'}, \ldots, \theta^{(N)}],$$

and suppose that the expected average reward $\rho_t(\boldsymbol{\theta})$ for all $t = 1, \ldots, T$ is Lipschitz continuous with respect to each dimension $\theta^{(i)}$. Then we have

$$|\rho_t(\boldsymbol{\theta}) - \rho_t(\boldsymbol{\theta}')| \leq C_2 |\theta^{(i)} - \theta^{(i)'}|, \forall i = 1, \ldots, N.$$

Form Proposition 4, we have the following proposition:

Proposition 5. *For all $t = 1, \ldots, T$, the partial derivative of expected average reward function $\rho_t(\boldsymbol{\theta})$ with respect to $\theta^{(i)}$ is bounded as*

$$\left| \frac{\partial \rho_t(\boldsymbol{\theta})}{\partial \theta^{(i)}} \right| \leq C_2, \forall i = 1, \ldots, N,$$

and $\|\nabla_{\boldsymbol{\theta}} \rho_t(\boldsymbol{\theta})\|_1 \leq N C_2$.

From Proposition 5, the result of online convex optimization [10] is applicable to the current setup. More specifically we have

$$\sum_{t=1}^{T} (\rho_t(\boldsymbol{\theta}^*) - \rho_t(\boldsymbol{\theta}_t)) \leq \frac{F^2}{2}\sqrt{T} + C_2 N \sqrt{T},$$

which concludes the proof.

4.5 Proof of Lemma 3

The following proposition holds, which can be obtained from Assumption 2 and

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}\|_1 \leq \eta_t \|\nabla_{\boldsymbol{\theta}} \rho_t(\boldsymbol{\theta}_t)\|_1 \leq C_2 N \eta_t.$$

Proposition 6. *Consecutive policy parameters $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_{t+1}$ given by the OPG algorithm satisfy*

$$\int_{\boldsymbol{a}\in A} |\pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}_t) - \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}_{t+1})| \mathrm{d}\boldsymbol{a} \leq C_1 C_2 N \eta_t.$$

From Proposition 2 and Proposition 6, we have the following proposition:

Proposition 7. *For consecutive policy parameters $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_{t+1}$ given by the OPG algorithm and arbitrary transition probability density $p(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})$, it holds that*

$$\begin{aligned}&\int_{\boldsymbol{s}\in S} d(\boldsymbol{s}) \int_{\boldsymbol{s}'\in S} \int_{\boldsymbol{a}\in A} p(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) \\ &\times |\pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}_t) - \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}_{t+1})| \mathrm{d}\boldsymbol{a} \mathrm{d}\boldsymbol{s}' \mathrm{d}\boldsymbol{s} \leq C_1 C_2 N \eta_t.\end{aligned}$$

Then the following proposition holds, which is proved in Section 4.6 following the same line as Lemma 5.1 in [2]:

Proposition 8. *The state distribution $d_{\mathcal{A},t}$ given by algorithm $\mathcal{A}$ and the stationary state distribution $d_{\boldsymbol{\theta}_t}$ of policy $\pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}_t)$ satisfy*

$$\int_{\boldsymbol{s}\in S} |d_{\boldsymbol{\theta}_t}(\boldsymbol{s}) - d_{\mathcal{A},t}(\boldsymbol{s})| \mathrm{d}\boldsymbol{s} \leq 2\tau^2 \eta_{t-1} C_1 C_2 N + 2e^{-t/\tau}.$$

Although the original bound given in [1,2] depends on the cardinality of the action space, it is not the case in the current setup.

Then the third term of the decomposed regret (5) is expressed as

$$\begin{aligned}\left| R_{\mathcal{A}}(T) - \sum_{t=1}^{T} \rho_t(\boldsymbol{\theta}_t) \right| &= \left| \sum_{t=1}^{T} \int_{\boldsymbol{s}\in S} d_{\mathcal{A},t}(\boldsymbol{s}) \int_{\boldsymbol{a}\in A} r_t(\boldsymbol{s},\boldsymbol{a}) \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}_t) \mathrm{d}\boldsymbol{a} \mathrm{d}\boldsymbol{s} \right. \\ &\qquad \left. - \sum_{t=1}^{T} \int_{\boldsymbol{s}\in S} d_{\boldsymbol{\theta}_t}(\boldsymbol{s}) \int_{\boldsymbol{a}\in A} r_t(\boldsymbol{s},\boldsymbol{a}) \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}_t) \mathrm{d}\boldsymbol{a} \mathrm{d}\boldsymbol{s} \right| \\ &\leq \sum_{t=1}^{T} \int_{\boldsymbol{s}\in S} |d_{\mathcal{A},t}(\boldsymbol{s}) - d_{\pi_t}(\boldsymbol{s})| \mathrm{d}\boldsymbol{s} \\ &\leq 2\tau^2 C_1 C_2 N \sum_{t=1}^{T} \eta_t + 2 \sum_{t=1}^{T} e^{-t/\tau} \\ &\leq 2\tau^2 C_1 C_2 N \sqrt{T} + 2\tau,\end{aligned}$$

which concludes the proof.

4.6 Proof of Proposition 3

For two different parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, we have

$$\begin{aligned}
|\rho_t(\boldsymbol{\theta}) - \rho_t(\boldsymbol{\theta}')| = \Bigg| & \int_{\boldsymbol{s}\in S} d_{\boldsymbol{\theta}}(\boldsymbol{s}) \int_{\boldsymbol{a}\in A} \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}) r_t(\boldsymbol{s},\boldsymbol{a}) \mathrm{d}\boldsymbol{a}\mathrm{d}\boldsymbol{s} \\
& - \int_{\boldsymbol{s}\in S} d_{\boldsymbol{\theta}'}(\boldsymbol{s}) \int_{\boldsymbol{a}\in A} \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}') r_t(\boldsymbol{s},\boldsymbol{a}) \mathrm{d}\boldsymbol{a}\mathrm{d}\boldsymbol{s} \Bigg| \\
\leq & \int_{\boldsymbol{s}\in S} |d_{\boldsymbol{\theta}}(\boldsymbol{s}) - d_{\boldsymbol{\theta}'}(\boldsymbol{s})| \int_{\boldsymbol{a}\in A} \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}) r_t(\boldsymbol{s},\boldsymbol{a}) \mathrm{d}\boldsymbol{a}\mathrm{d}\boldsymbol{s} \\
& + \int_{\boldsymbol{s}\in S} d_{\boldsymbol{\theta}'}(\boldsymbol{s}) \int_{\boldsymbol{a}\in A} |\pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}) - \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}')| \, r_t(\boldsymbol{s},\boldsymbol{a}) \mathrm{d}\boldsymbol{a}\mathrm{d}\boldsymbol{s}.
\end{aligned} \tag{6}$$

The first equation comes from Eq.(1), and the second inequality is obtained from the triangle inequality. Since Assumption 2 and Assumption 3 imply

$$\int_{\boldsymbol{s}\in S} d_{\boldsymbol{\theta}'}(\boldsymbol{s}) \int_{\boldsymbol{a}\in A} |\pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}) - \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}')| r_t(\boldsymbol{s},\boldsymbol{a}) \mathrm{d}\boldsymbol{a}\mathrm{d}\boldsymbol{s} \leq C_1 \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_1,$$

and also

$$\int_{\boldsymbol{a}\in A} \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}) r_t(\boldsymbol{s},\boldsymbol{a}) \mathrm{d}\boldsymbol{a} \leq 1,$$

Eq.(6) can be written as

$$\begin{aligned}
|\rho_t(\boldsymbol{\theta}) - \rho_t(\boldsymbol{\theta}')| &\leq \int_{\boldsymbol{s}\in S} |d_{\boldsymbol{\theta}}(\boldsymbol{s}) - d_{\boldsymbol{\theta}'}(\boldsymbol{s})| \mathrm{d}\boldsymbol{s} + C_1 \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_1 \\
&= \int_{\boldsymbol{s}\in S} \int_{\boldsymbol{s}'\in S} |d_{\boldsymbol{\theta}}(\boldsymbol{s}') p(\boldsymbol{s}|\boldsymbol{s}';\boldsymbol{\theta}) - d_{\boldsymbol{\theta}'}(\boldsymbol{s}') p(\boldsymbol{s}|\boldsymbol{s}';\boldsymbol{\theta}')| \mathrm{d}\boldsymbol{s}'\mathrm{d}\boldsymbol{s} \\
&\quad + C_1 \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_1 \\
&\leq \int_{\boldsymbol{s}\in S} \int_{\boldsymbol{s}'\in S} |d_{\boldsymbol{\theta}}(\boldsymbol{s}') p(\boldsymbol{s}|\boldsymbol{s}';\boldsymbol{\theta}) - d_{\boldsymbol{\theta}'}(\boldsymbol{s}') p(\boldsymbol{s}|\boldsymbol{s}';\boldsymbol{\theta})| \mathrm{d}\boldsymbol{s}'\mathrm{d}\boldsymbol{s} \\
&\quad + \int_{\boldsymbol{s}\in S} \int_{\boldsymbol{s}'\in S} d_{\boldsymbol{\theta}'}(\boldsymbol{s}') |p(\boldsymbol{s}|\boldsymbol{s}';\boldsymbol{\theta}) - p(\boldsymbol{s}|\boldsymbol{s}';\boldsymbol{\theta}')| \mathrm{d}\boldsymbol{s}'\mathrm{d}\boldsymbol{s} \\
&\quad + C_1 \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_1 \\
&\leq e^{-1/\tau} \int_{\boldsymbol{s}\in S} |d_{\boldsymbol{\theta}}(\boldsymbol{s}) - d_{\boldsymbol{\theta}'}(\boldsymbol{s})| \mathrm{d}\boldsymbol{s} + 2C_1 \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_1.
\end{aligned}$$

The second equality comes from the definition of the stationary state distribution, and the third inequality can be obtained from the triangle inequality. The last inequality follows from Assumption 1 and Proposition 2. Thus, we have

$$|\rho_t(\boldsymbol{\theta}) - \rho_t(\boldsymbol{\theta}')| \leq \frac{2C_1 - C_1 e^{-1/\tau}}{1 - e^{-1/\tau}} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_1,$$

which concludes the proof.

4.7 Proof of Proposition 8

This proof is following the same line as Lemma 5.1 in [2].

$$
\begin{aligned}
&\int_{\boldsymbol{s}\in S} |d_{\mathcal{A},k}(\boldsymbol{s}) - d_{\boldsymbol{\theta}_t}(\boldsymbol{s})| \mathrm{d}\boldsymbol{s} \\
&= \int_{\boldsymbol{s}\in S}\int_{\boldsymbol{s}'\in S} |d_{\mathcal{A},k-1}(\boldsymbol{s}')p(\boldsymbol{s}|\boldsymbol{s}';\boldsymbol{\theta}_k) - d_{\boldsymbol{\theta}_t}(\boldsymbol{s}')p(\boldsymbol{s}|\boldsymbol{s}';\boldsymbol{\theta}_t)| \,\mathrm{d}\boldsymbol{s}'\mathrm{d}\boldsymbol{s} \\
&\leq \int_{\boldsymbol{s}\in S}\int_{\boldsymbol{s}'\in S} |d_{\mathcal{A},k-1}(\boldsymbol{s}')p(\boldsymbol{s}|\boldsymbol{s}';\boldsymbol{\theta}_t) - d_{\boldsymbol{\theta}_t}(\boldsymbol{s}')p(\boldsymbol{s}|\boldsymbol{s}';\boldsymbol{\theta}_t)| \,\mathrm{d}\boldsymbol{s}'\mathrm{d}\boldsymbol{s} \\
&\quad + \int_{\boldsymbol{s}\in S}\int_{\boldsymbol{s}'\in S} |d_{\mathcal{A},k-1}(\boldsymbol{s}')p(\boldsymbol{s}|\boldsymbol{s}';\boldsymbol{\theta}_k) - d_{\mathcal{A},k-1}(\boldsymbol{s}')p(\boldsymbol{s}|\boldsymbol{s}';\boldsymbol{\theta}_t)| \,\mathrm{d}\boldsymbol{s}'\mathrm{d}\boldsymbol{s} \\
&\leq e^{-1/\tau} \int_{\boldsymbol{s}\in S} |d_{\mathcal{A},k-1}(\boldsymbol{s}) - d_{\boldsymbol{\theta}_t}(\boldsymbol{s})| \,\mathrm{d}\boldsymbol{s} + 2(t-k)C_1C_2N\eta_{t-1}. \qquad (7)
\end{aligned}
$$

The first equation comes from the definition of the stationary state distribution, and the second inequality can be obtained by the triangle inequality. The third inequality holds from Assumption 1 and

$$
\begin{aligned}
&\int_{\boldsymbol{s}\in S}\int_{\boldsymbol{s}'\in S} |d_{\mathcal{A},k-1}(\boldsymbol{s}')p(\boldsymbol{s}|\boldsymbol{s}';\boldsymbol{\theta}_k) - d_{\mathcal{A},k-1}(\boldsymbol{s}')p(\boldsymbol{s}|\boldsymbol{s}';\boldsymbol{\theta}_t)| \,\mathrm{d}\boldsymbol{s} \\
&\leq C_1\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_k\|_1 \\
&\leq C_1 \sum_{i=k}^{t-1} \eta_i \|\nabla_{\boldsymbol{\theta}}\rho_i(\boldsymbol{\theta}_i)\|_1 \\
&\leq 2(t-k)C_1C_2N\eta_{t-1}.
\end{aligned}
$$

Recursively using Eq.(7), we have

$$
\begin{aligned}
\int_{\boldsymbol{s}\in S} |d_{\mathcal{A},t}(\boldsymbol{s}) - d_{\pi_t}(\boldsymbol{s})| \mathrm{d}\boldsymbol{s} &\leq 2\sum_{k=2}^{t} e^{-(t-k)/\tau}(t-k)C_1C_2N\eta_{t-1} + 2e^{-t/\tau} \\
&\leq 2\tau^2 C_1C_2N\eta_{t-1} + 2e^{-t/\tau},
\end{aligned}
$$

which concludes the proof.

5 Regret Analysis under Strong Concavity

In this section, we derive a shaper regret bound for the OPG algorithm under a strong concavity assumption.

Theorem 1 shows the theoretical guarantee of the OPG algorithm with the concave assumption. If the expected reward function is strongly concave, i.e.,

$$\nabla_\theta^2 \rho_t \leq -HI_N,$$

where H is a positive constant and I_N is the $N \times N$ identity matrix, we have following theorem.

Theorem 2. *The regret against the best offline policy of the OPG algorithm is bounded as*

$$L_{\mathcal{A}}(T) \leq \frac{C_2^2 N^2}{2H}(1+\log T) + \frac{2\tau^2 C_1 C_2 N}{H}\log T + 4\tau,$$

with step size $\eta_t = \frac{1}{Ht}$.

We again consider the same decomposition as Eq.(5), the first term of the regret bound is exactly the same as Lemma 1. The second and third parts are given by the following propositions.

Given the strongly concavity assumption and step size $\eta_t = \frac{1}{Ht}$, the following proposition holds:

Proposition 9.

$$\sum_{t=1}^{T}(\rho_t(\boldsymbol{\theta}^*) - \rho_t(\boldsymbol{\theta}_t)) \leq \frac{C_2^2 N^2}{2H}(1+\log T).$$

The proof is following the same line as [12], i.e., by the Taylor approximation, the expected average reward function can be decomposed as

$$\begin{aligned}
&\rho_t(\boldsymbol{\theta}^*) - \rho_t(\boldsymbol{\theta}_t) \\
&= \nabla_\theta \rho_t(\boldsymbol{\theta}_t)^\top(\boldsymbol{\theta}^* - \boldsymbol{\theta}_t) + \frac{1}{2}(\boldsymbol{\theta}_t)^\top(\boldsymbol{\theta}^* - \boldsymbol{\theta}_t)^\top \nabla_\theta^2 \rho_t(\boldsymbol{\xi}_t)(\boldsymbol{\theta}_t)^\top(\boldsymbol{\theta}^* - \boldsymbol{\theta}_t) \\
&\leq \nabla_\theta \rho_t(\boldsymbol{\theta}_t)^\top(\boldsymbol{\theta}^* - \boldsymbol{\theta}_t) - \frac{H}{2}\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_t\|^2. \qquad (8)
\end{aligned}$$

Given the parameter updating rule,

$$\nabla_\theta \rho_t(\boldsymbol{\theta}^* - \boldsymbol{\theta}_t) = \frac{1}{2\eta_t}\left((\boldsymbol{\theta}^* - \boldsymbol{\theta}_t)^2 - (\boldsymbol{\theta}^* - \boldsymbol{\theta}_{t+1})^2\right) + \eta_t \|\nabla_\theta \rho_t(\boldsymbol{\theta}_t)\|^2,$$

summing up all T terms of (8) and setting $\eta_t = \frac{1}{Ht}$ yield

$$\begin{aligned}
\sum_{t=1}^{T}(\rho_t(\boldsymbol{\theta}^* - \boldsymbol{\theta}_t)) &\leq \sum_{t=1}^{T}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} - H\right)\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_t\|^2 + \|\nabla_t \rho_t(\boldsymbol{\theta}_t)\|^2 \sum_{t=1}^{T}\eta_t \\
&\leq \frac{C_2^2 N^2}{2H}(1+\log T).
\end{aligned}$$

From the proof of Lemma 3, the bound of the third part with the strongly concavity assumption is given by following proposition.

Proposition 10.

$$\sum_{t=1}^{T}\rho_t(\boldsymbol{\theta}_t) - R_{\mathcal{A}}(T) \leq \frac{2\tau^2 C_1 C_2 N}{H}\log T + 2\tau. \qquad (9)$$

The result of Proposition 10 is obtained by following the same line as the proof of Lemma 3 with different step sizes. Combining Lemma 1, Proposition 9, and Proposition 10, we can obtain Theorem 2.

6 Experiments

In this section, we illustrate the behavior of the OPG algorithm.

6.1 Target Tracking

The task is to let an agent track an abruptly moving target located in one-dimensional real space $S = \mathbb{R}$. The action space is also one-dimensional real space $A = \mathbb{R}$, and we can change the position of the agent as $s' = s + a$. The reward function is given by evaluating the distance between the agent and target as

$$r_t(s, a) = e^{-|s+a-\mathrm{tar}(t)|},$$

where $\mathrm{tar}(t)$ denotes the position of the target at time step t. Because the target is moving abruptly, the reward function is also changing abruptly. As a baseline method for comparison, we consider the MDP-E algorithm [1,2], where the exponential weighted average algorithm is used as the best expert. Since MDP-E can handle only discrete states and actions, we discretize the state and action space. More specifically, the state space is discretized as

$$(-\infty, -6], (-6, -6 + c], (-6 + c, -6 + 2c], \ldots, (6, +\infty),$$

and the action space is discretized as

$$-6, -6 + c, -6 + 2c, \ldots, 6.$$

We consider the following 5 setups for c:

$$c = 6, 2, 1, 0.5, 0.1.$$

In the experiment, the stationary state distribution and the gradient are estimated by policy gradient theorem estimator[5]. $I = 20$ independent experiments are run with $T = 100$ time steps, and the average return $J(T)$ is used for evaluating the performance:

$$J(T) = \frac{1}{I} \sum_{i=1}^{I} \left[\sum_{t=1}^{T} r_t(s_t, a_t) \right].$$

The results are plotted in Figure 1, showing that the OPG algorithm works better than the MDP-E algorithm with the best discretization resolution. This illustrates the advantage of directly handling continuous state and action spaces without discretization.

6.2 Linear-Quadratic Regulator

The *linear-quadratic regulator* (LQR) is a simple system, where the transition dynamics is linear and the reward function is quadratic. A notable advantage of LQR is that we can compute the best offline parameter [5]. Here, an online LQR system is simulated to illustrate the parameter update trajectory of the OPG algorithm.

Let state and action spaces be one-dimensional real: $S = \mathbb{R}$ and $A = \mathbb{R}$. Transition is deterministically performed as

$$s' = s + a.$$

The reward function is defined as

$$r_t(s, a) = -\frac{1}{2}Q_t s^2 - \frac{1}{2}R_t a^2,$$

where $Q_t \in \mathbb{R}$ and $R_t \in \mathbb{R}$ are chosen from $\{1, \ldots, 10\}$ uniformly for each t. Thus, the reward function is changing abruptly.

We use the Gaussian policy with mean parameter $\mu \cdot s$ and standard deviation parameter $\sigma = 0.1$, i.e., $\theta = \mu$. The best offline parameter is given by $\theta^* = -0.98$, and the initial parameter for the OPG algorithm is set at $\theta_1 = -0.5$.

In the top graph of Figure 2, a parameter update trajectory of OPG in an online LQR problem is plotted by the red line, and the best offline parameter is denoted by the black line. This shows that the OPG solution quickly approaches the best offline parameter.

Next, we also include the Gaussian standard deviation σ in the policy parameter, i.e., $\boldsymbol{\theta} = (\mu, \sigma)^\top$. When σ takes a value less than 0.001 during gradient update iterations, we project it back to 0.001. A parameter update trajectory is plotted in the bottom graph of Figure 2, showing again that the OPG solution smoothly approaches the best offline parameter along μ.

7 Conclusion

In this paper, we proposed an online policy gradient method for continuous state and action online MDPs, and showed that the regret of the proposed method is $O(\sqrt{T})$ under a certain concavity assumption. A notable fact is that the regret bound does not depend on the cardinality of state and action spaces, which makes the proposed algorithm suitable in handling continuous states and actions. Furthermore, we also established the $O(\log T)$ regret bound under a strongly concavity assumption. Through experiments, we illustrated that directly handling continuous state and action spaces by the proposed method is more advantageous than discretizing them.

Our future work will extend the current theoretical analysis to non-concave expected average reward functions, where gradient-based algorithms suffer from the local optimal problem. Another important challenge is to develop an effective method to estimate the stationary state distribution which is required in our algorithm.

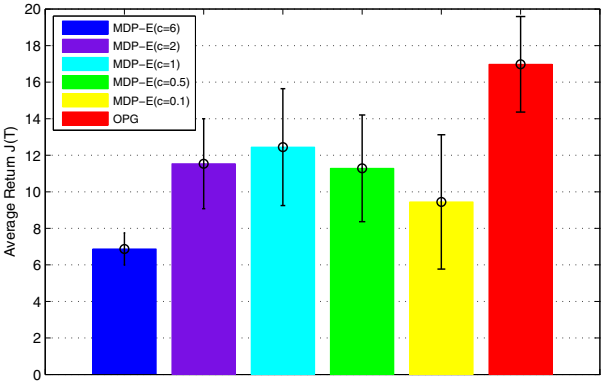

Fig. 1. Average returns of the OPG algorithm and the MDP-E algorithm with different discretization resolution c

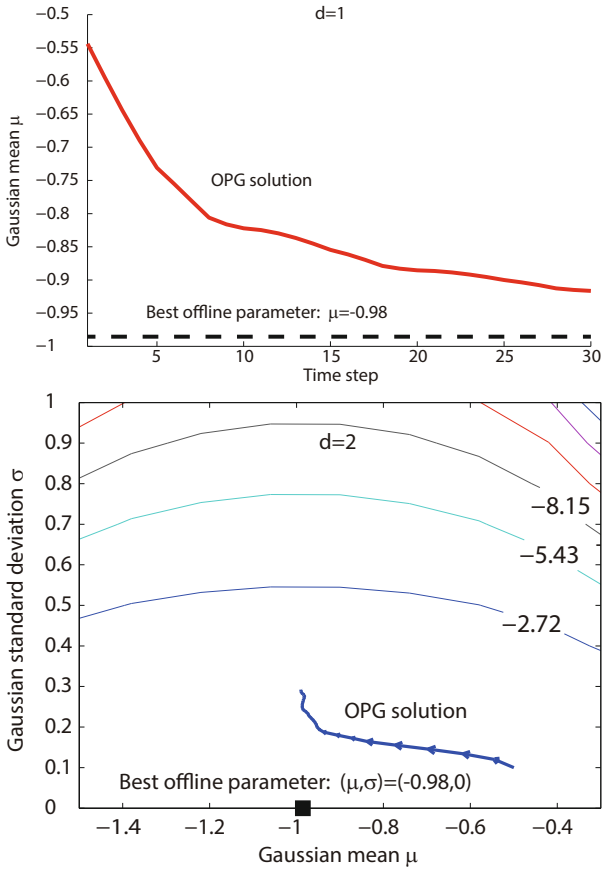

Fig. 2. Trajectory of the OPG solutions and the best offline parameter

Acknowledgments. YM was supported by the MEXT scholarship and the CREST program. KH was supported by MEXT KAKENHI 25330261 and 24106010. MS was supported by KAKENHI 23120004.

References

1. Even-Dar, E., Kakade, S.M., Mansour, Y.: Experts in a Markov Decision Process. In: Advances in Neural Information Processing System 17, pp. 401–408. MIT Press, Cambridge (2005)
2. Even-Dar, E., Karade, S.M., Mansour, Y.: Online Markov Decision Processes. Mathematics of Operations Research 34(3), 726–736 (2009)
3. Neu, G., György, A., Szepesvári, C., Antos, A.: Online Markov Decision Processes under Bandit Feedback. In: Advances in Neural Information Processing Systems 23, pp. 1804–1812 (2010)
4. Neu, G., György, A., Szepesvári, C.: The Online Loop-free Stochastic Shortest-path Problem. In: Conference on Learning Theory, pp. 231–243 (2010)
5. Peter, J., Schaal, S.: Policy Gradient Methods for Robotics. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (2006)
6. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press (1998)
7. Williams, R.J.: Simple Statistical Gradient-following Algorithms for Connetionist Reinforcement Learning. Machine Learning 8(3-4), 229–256 (1992)
8. Yu, J.Y., Mannor, S., Shimkin, N.: Markov Decision Processes with Arbitrary Reward Processes. Mathematics of Operations Research 34(3), 737–757 (2009)
9. Zimin, A., Neu, G.: Online Learning in Episodic Markovian Decision Processes by Relative Entropy Policy Search. In: Advances in Neural Information Processing Systems 26, pp. 1583–1591 (2013)
10. Zinkevich, M.: Online Convex Programming and Generalized Infinitesimal Gradient Ascent. In: International Conference on Machine Learning, pp. 928–936. AAAI Press (2003)
11. Abbasi-Yadkori, Y., Bartlett, P., Kanade, V., Seldin, Y., Szepesvári, C.: Online Learning in Markov Decision Processes with Adversarially Chosen Transition Probability Distributions. In: Advances in Neural Information Processing Systems 26 (2013)
12. Hazan, E., Agarwal, A., Kale, S.: Logarithmic Regret Algorithms for Online Convex Optimization. Machine Learning 69, 169–192 (2007)