Implicit Incremental Natural Actor Critic

Ryo Iwaki and Minoru Asada

Osaka University,
2-1, Yamadaoka, Suita city, Osaka, Japan
{ryo.iwaki,asada}@ams.eng.osaka-u.ac.jp

Abstract. The natural policy gradient (NPG) method is a promising approach to find a locally optimal policy parameter. The NPG method has been demonstrated remarkable successes in many fields, including the large scale applications. On the other hand, the estimation of the NPG itself requires a enormous amount of samples. Furthermore, incremental estimation of the NPG is computationally unstable. In this work, we propose a new incremental and stable algorthm for the NPG estimation. The proposed algorithm is based on the idea of *implicit temporal differences*, and we call the proposed one *implicit incremental natural actor critic* (I2NAC). Theoretical analysis indicates the stability of I2NAC and the instability of conventional incremental NPG methods. Numerical experiment shows that I2NAC is less sensitive to the value of step sizes.

Keywords: reinforcement learning, natural policy gradient, incremental natural actor critic, incremental learning, implicit update

1 Introduction

The natural policy gradient (NPG) method [10] is one of the branches of reinforcement learning (RL), which seeks a locally optimal policy by gradient ascent. By using the *natural gradient*, the plateaus in the learning can be avoided [1]. The NPG methods have demonstrated remarkale successes in many fields, such as traffic optimization [17], dialog system [9] and the high dimensional control tasks including the control of humanoid robots [3, 7, 14, 15, 18].

In this study, we focus on the incremental natural actor critic (INAC) [2, 5, 13, 22]. INAC methods have three advantages: (i) the sample complexity is $\mathcal{O}(n)$, (ii) all the update procedure can be executed by simple stochastic gradient descent, and (iii) even when the Fisher information matrix (FIM) degenerates, INAC estimates NPG by implicitly calculating the pseudo inverse of FIM [23]. However, INAC has a serious drawback: it is very difficult to tune the *step size*, and the iteration for NPG estimation is very unstable and divergent. There are many studies in the literature to improve the stabiliity of the iteraion to update the policy [8, 12, 16] and the state value function [4, 21], but, to the best of our knowledge, there are very few studies to deal with the stability of NPG iteration.

The goal of this paper is to reveal the reason why the existing INAC algorithms are unstable, and to propose an incremental and stable algorithm for the NPG estimation. The proposed method, which we refer to *implicit incremental natural actor critic* (I2NAC), is based on the idea of the *implicit stochastic gradient descent* [24] and the *implicit temporal differences* [21]. Theoretical analysis points out the stability of I2NAC and the instability of the existing INAC methods. It is shown in a classical benchmark test that I2NAC is less sensitive to the value of step sizes.

2 Background

2.1 Natural Policy Gradient

We assume that the problem is a Markov decision process (MDP). An MDP is specified by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$. $\mathcal{S}$ is a set of possible states of an environment and $\mathcal{A}$ is a set of possible actions an agent can choose, both of which could be discrete or continuous. $\mathcal{P}$ and $\mathcal{R}$ denotes the state transition probability and the bounded reward function, respectively. $\gamma \in [0, 1)$ is the discount factor. In case of model-free RL, the agent does not have the knowledge about $\mathcal{P}$ and $\mathcal{R}$.

At each discrete time step $t \in \mathbb{N}_{\geq 0}$, the agent observes the current state $\boldsymbol{s}_t \in \mathcal{S}$ and chooses the action $\boldsymbol{a}_t \in \mathcal{A}$. The state of the environment transits to the next state $\boldsymbol{s}_{t+1}$ according to $\mathcal{P}^a_{ss'} \triangleq \Pr(\boldsymbol{s}_{t+1} = \boldsymbol{s}' | \boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{a}_t = \boldsymbol{a})$, and the agent receives the reward $r_t \in \mathbb{R}$ according to $\mathcal{R}^a_s \triangleq \mathbb{E}[r_t | \boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{a}_t = \boldsymbol{a}]$. The agent's decision making is characterized by a parameterized stochastic *policy* $\pi(\boldsymbol{a}|\boldsymbol{s}; \boldsymbol{\theta}) \triangleq \Pr(\boldsymbol{a}_t = \boldsymbol{a} | \boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{\theta})$, which is a distribution over actions given the state and parameter $\boldsymbol{\theta} \in \mathbb{R}^n$. We assume that $\pi(\boldsymbol{a}|\boldsymbol{s}; \boldsymbol{\theta})$ is differentiable with respect to $\boldsymbol{\theta}$ for all $\boldsymbol{s}$ and $\boldsymbol{a}$, and allow a shorthand notation: $\pi_{\boldsymbol{\theta}} \triangleq \pi(\boldsymbol{a}|\boldsymbol{s}; \boldsymbol{\theta})$. There exists the limiting stationary state distribution $d^\pi(\boldsymbol{s})$ independent of the initial state: $d^\pi(\boldsymbol{s}) = \lim_{t\to\infty} \Pr(\boldsymbol{s}_t = \boldsymbol{s} | \boldsymbol{s}_0 = \boldsymbol{s}', \pi_{\boldsymbol{\theta}}), \forall \boldsymbol{s}' \in \mathcal{S}$.

For each policy $\pi_{\boldsymbol{\theta}}$, the state value function $V^\pi(\boldsymbol{s})$ and the state-action value function $Q^\pi(\boldsymbol{s}, \boldsymbol{a})$ are given by $V^\pi(\boldsymbol{s}) = \mathop{\mathbb{E}}_{\pi,\mathcal{P}}\left[\sum_{\tau=0}^{\infty} \gamma^\tau r_{t+\tau} | \boldsymbol{s}_t = \boldsymbol{s}\right]$ and $Q^\pi(\boldsymbol{s}, \boldsymbol{a}) = \mathop{\mathbb{E}}_{\pi,\mathcal{P}}\left[\sum_{\tau=0}^{\infty} \gamma^\tau r_{t+\tau} | \boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{a}_t = \boldsymbol{a}\right]$, respectively. The purpose of the agent is to find the (locally) optimal policy parameter $\boldsymbol{\theta}^*$ which maximizes the average reward: $J(\boldsymbol{\theta}) \triangleq \lim_{T\to\infty} \frac{1}{T} \mathop{\mathbb{E}}_{\pi,\mathcal{P}}\left[\sum_{t=0}^{T-1} r_t\right] = \sum_{\boldsymbol{s}} d^\pi(\boldsymbol{s}) \sum_{\boldsymbol{a}} \pi(\boldsymbol{a}|\boldsymbol{s}; \boldsymbol{\theta}) \mathcal{R}^a_s$.

Let $f_{\boldsymbol{w}}(\boldsymbol{s}, \boldsymbol{a})$ be a linear function approximator given by

$$f_{\boldsymbol{w}}(\boldsymbol{s}, \boldsymbol{a}) \triangleq \boldsymbol{w}^\top \boldsymbol{\psi}(\boldsymbol{s}, \boldsymbol{a}) = \boldsymbol{w}^\top \nabla_{\boldsymbol{\theta}} \ln \pi(\boldsymbol{a}|\boldsymbol{s}; \boldsymbol{\theta}), \tag{1}$$

where $|\boldsymbol{w}| = |\boldsymbol{\theta}|$ and $\boldsymbol{\psi}$ is the *characteristic eligibility*. The approximator $f_{\boldsymbol{w}}(\boldsymbol{s}, \boldsymbol{a})$ is *compatible* in the sense that the following equation holds:

$$\nabla_{\boldsymbol{w}} f_{\boldsymbol{w}}(\boldsymbol{s}, \boldsymbol{a}) = \nabla_{\boldsymbol{\theta}} \ln \pi(\boldsymbol{a}|\boldsymbol{s}; \boldsymbol{\theta}) = \frac{\nabla_{\boldsymbol{\theta}} \pi(\boldsymbol{a}|\boldsymbol{s}; \boldsymbol{\theta})}{\pi(\boldsymbol{a}|\boldsymbol{s}; \boldsymbol{\theta})}. \tag{2}$$

Assume that the following equation,

$$\mathop{\mathbb{E}}_{\boldsymbol{\theta}}\left[\left(Q^{\pi}(\boldsymbol{s},\boldsymbol{a})-b(\boldsymbol{s})-f_{\boldsymbol{w}}(\boldsymbol{s},\boldsymbol{a})\right)\nabla_{\boldsymbol{w}}f_{\boldsymbol{w}}(\boldsymbol{s},\boldsymbol{a})\right]=0 \tag{3}$$

holds, where $\mathop{\mathbb{E}}_{\boldsymbol{\theta}}[\cdot]$ denotes an expectation over the state-action pair under the current policy $\pi_{\boldsymbol{\theta}}$, that is, for an arbitrary variable $\boldsymbol{x}$, $\mathop{\mathbb{E}}_{\boldsymbol{\theta}}[\boldsymbol{x}] \triangleq \sum_{\boldsymbol{s}} d^{\pi}(\boldsymbol{s}) \sum_{\boldsymbol{a}} \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta})\,\boldsymbol{x}$, and $b(\boldsymbol{s})$ is a state-dependent arbitrary function, so called baseline. Then the *policy gradient* is given as follows [13, 20]:

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \simeq \mathop{\mathbb{E}}_{\boldsymbol{\theta}}\left[\nabla_{\boldsymbol{\theta}} \ln \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}) f_{\boldsymbol{w}}(\boldsymbol{s},\boldsymbol{a})\right]. \tag{4}$$

Thus, policy gradient can be estimated by approximating $Q^{\pi}(\boldsymbol{s},\boldsymbol{a})$ projected on to the subspace spanned by $\nabla_{\boldsymbol{\theta}} \ln \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta})$. The appropriate choise of baseline $b(\boldsymbol{s})$ reduces the variance of (4). The good choise of the baseline is the state value function $V^{\pi}(\boldsymbol{s})$. In this sense, $f_{\boldsymbol{w}}(\boldsymbol{s},\boldsymbol{a})$ approximates the *advantage* function, $A^{\pi}(\boldsymbol{s},\boldsymbol{a}) = Q^{\pi}(\boldsymbol{s},\boldsymbol{a}) - V^{\pi}(\boldsymbol{s})$. Furthermore, substituting Eq. (1) into Eq. (4) yields

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = G(\boldsymbol{\theta})\boldsymbol{w},$$

where $G(\boldsymbol{\theta})$ is the Fisher information matrix (FIM) of the policy distribution weighted by the stationary state distribution:

$$G(\boldsymbol{\theta}) \triangleq \mathop{\mathbb{E}}_{\boldsymbol{\theta}}\left[\nabla_{\boldsymbol{\theta}} \ln \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \ln \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta})^{\top}\right].$$

Thus the *natural policy gradient* [10] is given by:

$$\tilde{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = G^{-1}(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \boldsymbol{w}. \tag{5}$$

2.2 Incremental Natural Actor Critic

A number of algorithms have been proposed to estimate $\boldsymbol{w}$ satisfying Eq. (3), incrementally [2, 5, 13, 22]. In all of these algorithms, which we refer to incremental natural actor critic (INAC) algorithms, the approximation of the advantage function is performed in the form of the regression of the *temporal difference* (TD) error, δ^{π}, based on the fact that $\mathop{\mathbb{E}}_{\pi,\mathcal{P}}[\delta^{\pi}|\boldsymbol{s},\boldsymbol{a}] = A^{\pi}(\boldsymbol{s},\boldsymbol{a})$. The update of $\boldsymbol{w}$ is given by the following form:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \alpha\left(\delta_t - f_{\boldsymbol{w}}(\boldsymbol{s}_t,\boldsymbol{a}_t)\right)\boldsymbol{e}_t, \tag{6}$$

where δ_t is the approximated TD error and $\boldsymbol{e}_t$ is the *eligibility trace*.

For example, in the *natural policy gradient utilizing the temporal differences* (NTD) algorithm [13], δ_t and $\boldsymbol{e}_t$ are defined as follows, respectively:

$$\delta_t = r_t + \gamma V(\boldsymbol{s}_{t+1}) - V(\boldsymbol{s}_t),$$
$$\boldsymbol{e}_t = \sum_{\tau=0}^{t} (\gamma\lambda)^{t-\tau} \boldsymbol{\psi}_{\tau},$$

where V is the approximated state value function and $\lambda \in [0, 1]$ is the decay factor of trace. NTD and other algorithms [2, 5, 22] are different only in the definition of δ_t and $\boldsymbol{e}_t$.

3 Implicit Incremental Natural Actor Critic

In this section, first we propose an incremental NPG estimation algorithm based on the ideas of the *implicit stochastic gradient descent* [24] and the *implicit temporal differences* [21]. We start from expanding INAC update:

$$\begin{aligned}\boldsymbol{w}_{t+1} &= \boldsymbol{w}_t + \alpha \left(\delta_t - \boldsymbol{w}_t^\top \boldsymbol{\psi}_t\right) \boldsymbol{e}_t \\ &= \boldsymbol{w}_t + \alpha \left(\delta_t - \boldsymbol{w}_t^\top \boldsymbol{\psi}_t\right) \boldsymbol{e}_t + \beta \left(\boldsymbol{w}_t^\top \boldsymbol{e}_t - \boldsymbol{w}_t^\top \boldsymbol{e}_t\right) \boldsymbol{e}_t,\end{aligned}$$

where $\beta \geq \alpha$. Here we introduce the *implicit* update:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \alpha \left(\delta_t - \boldsymbol{w}_t^\top \boldsymbol{\psi}_t\right) \boldsymbol{e}_t + \beta \left(\boldsymbol{w}_t^\top \boldsymbol{e}_t - \boldsymbol{w}_{t+1}^\top \boldsymbol{e}_t\right) \boldsymbol{e}_t. \tag{7}$$

Eq. (7) is implicit in the sense that the parameter after the update, $\boldsymbol{w}_{t+1}$, appears on the both sides of equation. Note that the fixed point of Eq. (7) is the same as the fixed point of (6). It follows that

$$\left(I + \beta \boldsymbol{e}_t \boldsymbol{e}_t^\top\right) \boldsymbol{w}_{t+1} = \left(I + \beta \boldsymbol{e}_t \boldsymbol{e}_t^\top\right) \boldsymbol{w}_t + \alpha \left(\delta_t - \boldsymbol{w}_t^\top \boldsymbol{\psi}_t\right) \boldsymbol{e}_t.$$

The matrix $I + \beta \boldsymbol{e}_t \boldsymbol{e}_t^\top$ is positive definite. Finally, using the Sherman-Morrison formula, we have *implicit incremental natural actor critic* (I2NAC) algorithm:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \alpha \left(I + \beta \boldsymbol{e}_t \boldsymbol{e}_t^\top\right)^{-1} \left(\delta_t - \boldsymbol{w}_t^\top \boldsymbol{\psi}_t\right) \boldsymbol{e}_t \tag{8}$$

$$= \boldsymbol{w}_t + \alpha \left(I - \frac{\beta}{1 + \beta \|\boldsymbol{e}_t\|^2} \boldsymbol{e}_t \boldsymbol{e}_t^\top\right) \left(\delta_t - \boldsymbol{w}_t^\top \boldsymbol{\psi}_t\right) \boldsymbol{e}_t. \tag{9}$$

The difference of I2NAC from INAC is only the multiplication of the matrix $I - \frac{\beta}{1+\beta\|\boldsymbol{e}_t\|^2} \boldsymbol{e}_t \boldsymbol{e}_t^\top$. All the INAC algorithms of the form (6) can be converted into I2NAC. Note that the complexity of (9) is $\mathcal{O}(n)$, because (9) can be solved only by computing the inner products.

4 Theoretical Result

In this section, we analyze the stability of INAC and I2NAC. Similar analysis is performed in [21]. The updates (6) and (9) can be rewritten as follows, respectively:

$$\begin{aligned}\boldsymbol{w}_{t+1} &= \left(I - \alpha \boldsymbol{e}_t \boldsymbol{\psi}_t^\top\right) \boldsymbol{w}_t + \alpha \delta_t \boldsymbol{e}_t, \\ \boldsymbol{w}_{t+1} &= \left(I - \alpha E_t \boldsymbol{e}_t \boldsymbol{\psi}_t^\top\right) \boldsymbol{w}_t + \alpha \delta_t E_t \boldsymbol{e}_t,\end{aligned}$$

where $E_t \triangleq I - \frac{\beta}{1+\beta\|\boldsymbol{e}_t\|^2} \boldsymbol{e}_t \boldsymbol{e}_t^\top$. For the simplicity, we assume that the true state value function is given, thus $\delta_t = 0$. Let $\boldsymbol{w}_0$ denote the initial estimate

of the NPG, then the estimate of the NPG at time $T \in \mathbb{N}_{\geq 0}$ obtained by INAC and I2NAC can be rewritten as $\boldsymbol{w}_T = \prod_{t=0}^{T-1} \left(I - \alpha \boldsymbol{e}_t \boldsymbol{\psi}_t^\top\right) \boldsymbol{w}_0$ and $\boldsymbol{w}_T = \prod_{t=0}^{T-1} \left(I - \alpha E_t \boldsymbol{e}_t \boldsymbol{\psi}_t^\top\right) \boldsymbol{w}_0$, respectively. Thus the L2 norm of the NPG estimated by INAC and I2NAC are bounded as follows, respectively:

$$\|\boldsymbol{w}_T\|_2 \leq \prod_{t=0}^{T-1} \|I - \alpha \boldsymbol{e}_t \boldsymbol{\psi}_t^\top\|_2 \|\boldsymbol{w}_0\|_2,$$
$$\|\boldsymbol{w}_T\|_2 \leq \prod_{t=0}^{T-1} \|I - \alpha E_t \boldsymbol{e}_t \boldsymbol{\psi}_t^\top\|_2 \|\boldsymbol{w}_0\|_2.$$

If $\|I - \alpha \boldsymbol{e}_t \boldsymbol{\psi}_t^\top\|_2 \leq 1$ for all t, $\|\boldsymbol{w}_T\|_2$ stays bounded. The same argument holds for I2NAC. The following theorem gives $\|I - \alpha \boldsymbol{e}_t \boldsymbol{\psi}_t^\top\|_2$ and $\|I - \alpha E_t \boldsymbol{e}_t \boldsymbol{\psi}_t^\top\|_2$.

Theorem 1. *$\|I - \alpha \boldsymbol{e}_t \boldsymbol{\psi}_t^\top\|_2$ and $\|I - \alpha E_t \boldsymbol{e}_t \boldsymbol{\psi}_t^\top\|_2$ are given by*

$$\|I - \alpha \boldsymbol{e}_t \boldsymbol{\psi}_t^\top\|_2 = \max\{1, \sqrt{1 + \frac{\alpha^2 c_t^2 - 2\alpha d_t + \alpha c_t \sqrt{\alpha^2 c_t^2 + 4 - 4\alpha d_t}}{2}}\}, \quad (10)$$

$$\|I - \alpha E_t \boldsymbol{e}_t \boldsymbol{\psi}_t^\top\|_2 = \max\{1, \sqrt{1 + \frac{\alpha^2 \eta_t^2 c_t^2 - 2\alpha \eta_t d_t + \alpha \eta_t c_t \sqrt{\alpha^2 \eta_t^2 c_t^2 + 4 - 4\alpha \eta_t d_t}}{2}}\}, \quad (11)$$

respectively, where

$$\eta_t \triangleq \frac{1}{1 + \beta \|\boldsymbol{e}_t\|^2}, \quad c_t \triangleq \|\boldsymbol{e}_t\| \|\boldsymbol{\psi}_t\|, \quad d_t \triangleq \boldsymbol{e}_t^\top \boldsymbol{\psi}_t.$$

Proof. First we consider INAC. The norm of a real-valued matrix A is the square root of the maximum eigenvalue of $A^\top A$. We have

$$\begin{aligned}
\left(I - \alpha \boldsymbol{e}_t \boldsymbol{\psi}_t^\top\right)^\top \left(I - \alpha \boldsymbol{e}_t \boldsymbol{\psi}_t^\top\right) &= I - \alpha \boldsymbol{e}_t \boldsymbol{\psi}_t^\top - \alpha \boldsymbol{\psi}_t \boldsymbol{e}_t^\top + \alpha^2 \boldsymbol{\psi}_t \boldsymbol{e}_t^\top \boldsymbol{e}_t \boldsymbol{\psi}_t^\top \\
&= I + \boldsymbol{\psi}_t \left(\alpha^2 \boldsymbol{e}_t^\top \boldsymbol{e}_t \boldsymbol{\psi}_t^\top - \alpha \boldsymbol{e}_t^\top\right) - \alpha \boldsymbol{e}_t \boldsymbol{\psi}_t^\top \\
&\triangleq I + X. \quad (12)
\end{aligned}$$

Here we apply the following lemma (Lemma 2 in [21]).

Lemma 1. *Let $X = x_1 y_1^\top + x_2 y_2^\top \in \mathbb{R}^{n \times n}$, then the matrix X has $n-2$ eigenvalues equal to 0 and the rest 2 eigenvalues are given by*

$$\frac{x_1^\top y_1 + x_2^\top y_2 \pm \sqrt{(x_1^\top y_1 - x_2^\top y_2)^2 + 4(x_1^\top y_2)(y_1^\top x_2)}}{2}.$$

Thus, the matrix X in the righthand side of Eq. (12) has $n-2$ eigenvalues equal to 0, and the rest 2 eigenvalues are given by

$$\frac{\alpha^2 c_t^2 - 2\alpha d_t \pm \alpha c_t \sqrt{\alpha^2 c_t^2 + 4 - 4\alpha d_t}}{2},$$

where

$$c_t \triangleq \|\boldsymbol{e}_t\|\|\boldsymbol{\psi}_t\|, \quad d_t \triangleq \boldsymbol{e}_t^\top \boldsymbol{\psi}_t.$$

Therefore, the matrix in the righthand side of Eq. (12) has $n-2$ eigenvalues equal to 1, and the rest 2 eigenvalues are given by

$$1 + \frac{\alpha^2 c_t^2 - 2\alpha d_t \pm \alpha c_t \sqrt{\alpha^2 c_t^2 + 4 - 4\alpha d_t}}{2}.$$

Taking the square root of above gives $\|I - \alpha \boldsymbol{e}_t \boldsymbol{\psi}_t^\top\|_2$.

Next we consider I2NAC. Note that $E_t \boldsymbol{e}_t = \eta_t \boldsymbol{e}_t$ holds, where $\eta_t \triangleq \frac{1}{1+\beta\|\boldsymbol{e}_t\|^2}$. Therefore the same argument above holds for I2NAC by replacing $\boldsymbol{e}_t$ with $\eta_t \boldsymbol{e}_t$, and $\|I - \alpha E_t \boldsymbol{e}_t \boldsymbol{\psi}_t^\top\|_2$ can be obtained by simply replacing α with $\alpha\eta_t$ in $\|I - \alpha \boldsymbol{e}_t \boldsymbol{\psi}_t^\top\|_2$. □

Remark 1. Therorem 1 allows us to compare the stability of I2NAC with INAC. For the simplicity, by setting $\lambda = 0$, we have

$$\|I - \alpha \boldsymbol{\psi}_t \boldsymbol{\psi}_t^\top\|_2 = \max\{1, |\alpha\|\boldsymbol{\psi}_t\|^2 - 1|\} \geq 1, \tag{13}$$

$$\|I - \alpha E_t \boldsymbol{\psi}_t \boldsymbol{\psi}_t^\top\|_2 = \max\{1, |\alpha\eta_t\|\boldsymbol{\psi}_t\|^2 - 1|\} = \max\{1, |1 - \frac{\alpha\|\boldsymbol{\psi}_t\|^2}{1+\beta\|\boldsymbol{\psi}_t\|^2}|\} = 1. \tag{14}$$

The last equality in Eq. (14) holds because $\beta \geq \alpha$. Here, we assume that the policy is Gaussian, $\mathcal{N}(\mu, \sigma)$. Then the eligibility is given by $\psi_\mu = (a-\mu)/\sigma^2$ and $\psi_\sigma = ((a-\mu)^2 - \sigma^2)/\sigma^3$. In MDP, the optimal policy is deterministic. Thus, if the learning progresses successfully, $\sigma \to 0$ and $\|\psi\| \to \infty$. Therefore, (13) and (14) indicate that the iteration by INAC diverges even if the learning successes, while the iteration by I2NAC stays bounded.

5 Experimental Result

In the next experiment, we evaluate the robustness against the step size tuning. The pendulum swing up and stabilizing problem is a well known benchmark in continuous state-action space RL [6, 13]. The state of the environment consists of an angle $q \in [-\pi, \pi]$ and an angular velocity $\dot{q} \in [-15, 15]$ of pendulum, that is, $\boldsymbol{s} = (q, \dot{q})^\top$. The action of the agent is applied as a torque to the pendulum after scaling, that is, $5a = \tau \in [-5, 5]$. The dynamics of the pendulum is given by $ml^2\ddot{q} = -\mu\dot{q} + mgl\sin(q) + \tau$, where $m = l = 1, g = 9.8$ and $\mu = 0.01$, and numerically integrated with $\Delta t = 0.02$. An episode lasts for 1000 steps and the initial state in each episode is $\boldsymbol{s}_0 = (q_0, 0)^\top$, where q_0 is determined randomly. The policy parameter is not updated in the first 100 episodes, in order to avoid using the incomplete estimates of the NPG. The reward function is $\mathcal{R}(\boldsymbol{s}) = \cos(q) - (\dot{q}/15\pi)^2$, and the penalty for over-rotation does not exist. The policy is a Gaussian distribution:

$$\pi(a|\boldsymbol{s};\boldsymbol{\theta}) = \frac{1}{\sigma_{\boldsymbol{\theta}}(\boldsymbol{s})\sqrt{2\pi}} \exp\left(-\frac{(a-\mu_{\boldsymbol{\theta}}(\boldsymbol{s}))^2}{2\sigma_{\boldsymbol{\theta}}(\boldsymbol{s})^2}\right),$$

where the mean $\mu_{\boldsymbol{\theta}}(\boldsymbol{s})$ and the standard deviation $\sigma_{\boldsymbol{\theta}}(\boldsymbol{s})$ are determined by the output of a three layer fully connected neural network. The input vector is $(\cos(q), \sin(q), \dot{q})^\top$, and the hidden layer has 10 sigmoidal units. The output layer consists of two units: the mean unit has a tanh activation and the standard deviation unit has a sigmoidal activation. A small constant value $\sigma_0 = 0.01$ is added to the output of the standard deviation unit, in order to avoid the divergence of $\boldsymbol{\psi}_t$. The state value function is approximated using 7th order Fourier basis [11]. NTD and I2NAC (based on NTD iteration) are applied. We performed a grid search such that $\alpha, \alpha_{\boldsymbol{v}} \in \{10^{-1}, 5 \cdot 10^{-2}, \ldots, 10^{-4}\}, \alpha_{\boldsymbol{\theta}} \in \{10^{-4}, 5 \cdot 10^{-5}, \ldots, 10^{-7}\}$, where $\alpha_{\boldsymbol{v}}$ and $\alpha_{\boldsymbol{\theta}}$ are the step sizes for updating the parameters of the state value function and the policy, respectively. For I2NAC, the values $\{\alpha, 2\alpha, 10\alpha, 1\}$ were used for β in the grid search. The discount factor and the decay factor of trace were set to $\gamma = 0.98$ and $\lambda = 0.9$.

Figs. (1-5) shows the learning results for all the sets of the step sizes. The horizontal axes indicate the number of the episodes and the vertical axes indicate the average reward. For each set of step sizes, the result is averaged over 10 runs. If the estimate of even one run diverged, then the learning curve for the set is truncated. Therefore, if the learning diverges in many sets of the step sizes, the plot will be sparse, otherwise dense. Table 1 is the summary of the results, which shows the rate of the divergent sets. It was shown that iteration of I2NAC is much more stable and robust against the tuning of step sizes, compared to INAC. The rate of the divergent sets of I2NAC was still high, this was mainly because the parameters for the policy and state value function diverge if $\alpha_{\boldsymbol{\theta}}$ and $\alpha_{\boldsymbol{v}}$ are large. The adaptive step size methods for policy [12, 16] or state value function [4] would stabilize the learning process, but this issue is outside of the scope of this work. The larger value of β would stabilize the iteration, while the learning would be slower. However, the performance of I2NAC was less sensitive to the value of β.

Table 1: The rate of the divergent sets.

	NTD	I2NAC			
β	n/a	α	2α	10α	1
rate	88%	46%	51%	50%	43%

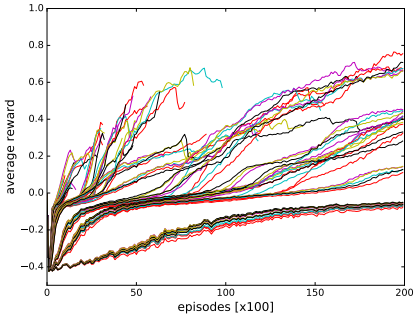

Fig. 1: NTD

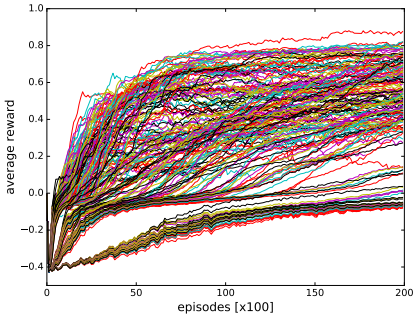

Fig. 2: I2NAC, $\beta = \alpha$

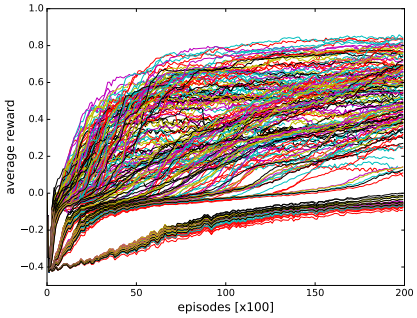

Fig. 3: I2NAC, $\beta = 2\alpha$

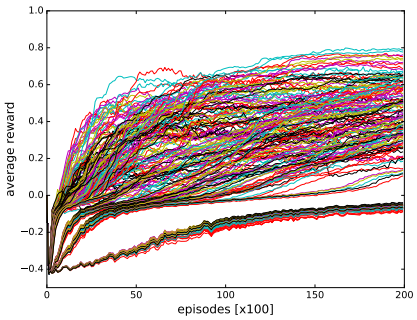

Fig. 4: I2NAC, $\beta = 10\alpha$

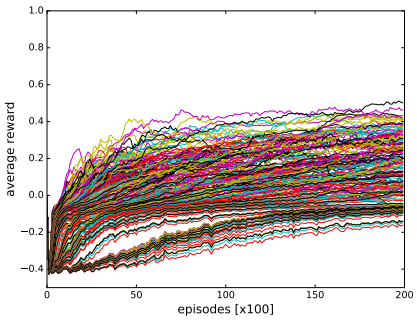

Fig. 5: I2NAC, $\beta = 1$

6 Conclusion and Outlook

In this work, we proposed incremental estimation algorithm of the NPG based on the implicit update. Theoretical analysis pointed out the stability of I2NAC and the instability of the existing INAC methods. It was shown in a classical benchmark test that I2NAC is less sensitive to the value of step sizes. The promising and straightforward future work is to extend I2NAC to the deterministic policy gradient method [19].

References

1. Amari, S.: Natural gradient works efficiently in learning. Neural Computation 10(2), 251–276 (1998)
2. Bhatnagar, S., Sutton, R.S., Ghavamzadeh, M., Lee, M.: Natural actor-critic algorithms. Automatica 45(11) (2009)
3. Chou, P.W., Maturana, D., Scherer, S.: Improving stochastic policy gradients in continuous control with deep reinforcement learning using the beta distribution. In: Proceedings of the 34th International Conference on Machine Learning, PMLR. vol. 70 (2017)
4. Dabney, W., Barto, A.G.: Adaptive step-size for online temporal difference learning. In: AAAI (2012)
5. Degris, T., Pilarski, P.M., Sutton, R.S.: Model-free reinforcement learning with continuous action in practice. In: Proceedings of the 2012 American Control Conference (2012)
6. Doya, K.: Reinforcement learning in continuous time and space. Neural Computation 12 (2000)
7. Duan, Y., Chen, X., Houthooft, R., Schulman, J., Abbeel, P.: Benchmarking deep reinforcement learning for continuous control. In: Proceedings of the 33rd International Conference on Machine Learning. pp. 1329–1338 (2016)
8. Greensmith, E., Bartlett, P.L., Baxter, J.: Variance reduction techniques for gradient estimates in reinforcement learning. Journal of Machine Learning Research 5 (2004)
9. Jurcicek, F., Thomson, B., Keizer, S., Mairesse, F., Gasic, M., Yu, K., Young, S.J.: Natural belief-critic: a reinforcement algorithm for parameter estimation in statistical spoken dialogue systems. In: INTERSPEECH. pp. 90–93 (2010)
10. Kakade, S.: A natural policy gradient. In: Advances in Neural Information Processing Systems. vol. 14 (2001)
11. Konidaris, G., Osentoski, S., Thomas, P.: Value function approximation in reinforcement learning using the fourier basis. In: Proceedings of the National Conference on Artificial Intelligence (AAAI). pp. 380–385 (2011)
12. Matsubara, T., Morimura, T., Morimoto, J.: Adaptive step-size policy gradients with average reward metric. In: Journal of Machine Learning Research- Proceedings Track. vol. 13 (2010)
13. Morimura, T., Uchibe, E., Doya, K.: Utilizing natural gradient in temporal difference reinforcement learning with eligibility traces. In: International Symposium on Information Geometry and Its Applications. pp. 256–263 (2005)
14. Peters, J., Schaal, S.: Natural actor-critic. In: Neurocomputing. vol. 71, pp. 1180–1190 (2008)
15. Peters, J., Vijayakumar, S., Schaal, S.: Reinforcement learning for humanoid robotics. In: Proceedings of the Third IEEE-RAS International Conference on Humanoid Robots. pp. 1–20. American Association for Artificial Intelligence (2003)
16. Pirotta, M., Restelli, M., Bascetta, L.: Adaptive step-size for policy gradient methods. In: Advances in Neural Information Processing Systems. pp. 1394–1402 (2013)
17. Richter, S., Aberdeen, D., Yu, J.: Natural actor-critic for road traffic optimisation. In: Schölkopf, P.B., Platt, J.C., Hoffman, T. (eds.) Advances in Neural Information Processing Systems, pp. 1169–1176. MIT Press (2007), `http://papers.nips.cc/paper/3087-natural-actor-critic-for-road-traffic-optimisation.pdf`
18. Schulman, J., Levine, S., Moritz, P., Jordan, M., Abbeel, P.: Trust region policy optimization. In: Proceedings of the 32nd International Conference on Machine Learning. pp. 1889–1897 (2015)

19. Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., Riedmiller, M.: Deterministic policy gradient algorithms. Proceedings of the 31st International Conference on Machine Learning pp. 387–395 (2014)
20. Sutton, R.S., McAllester, D., Singh, S., Mansour, Y.: Policy gradient methods for reinforcement learning with function approximation. In: Advances in Neural Information Processing Systems. vol. 12, pp. 1057–1063 (1999)
21. Tamar, A., Toulis, P., Mannor, S., Airoldi, E.M.: Implicit temporal differences. In: Neural Information Processing Systems, Workshop on Large-Scale Reinforcement Learning (2014)
22. Thomas, P.S.: Bias in natural actor-critic algorithms. In: Proceedings of The 31st International Conference on Machine Learning. pp. 441–448 (2014)
23. Thomas, P.S.: Genga: A generalization of natural gradient ascent with positive and negative convergence results. In: Proceedings of The 31st International Conference on Machine Learning. pp. 1575–1583 (2014)
24. Toulis, P., Rennie, J., Airoldi, E.M.: Statistical analysis of stochastic gradient methods for generalized linear models. Proceedings of The 31st International Conference on Machine Learning 32(1), 667–675 (2014)