Time Varying Nonlinear Policy Gradients

Evangelos A. Theodorou[1], Krishnamurthy Dvijotham[2] and Emo Todorov[3]

***Abstract*— We derive Policy Gradients(PGs) with time varying parameterizations for nonlinear diffusion processes affine in noise. The resulting policies have the form of reward weighted gradient. The analysis is in continuous time and includes the case of linear and nonlinear parameterizations. Examples on stochastic control problems for diffusions processes are provided.**

I. Introduction

Improving the performance of an initial controller/policy for dynamical systems with nonlinear and stochastic dynamics by using PGs has been a research topic in control theory and machine learning [7], [8]. The general approach relies on 1) the parameterization of the initial policy 2) Monte Carlo simulations of the stochastic dynamics to approximate the gradient of a performance criterion and 3) updates of the policy parameters in the direction of the gradient to improve performance.

A classical approach to solve nonlinear stochastic optimal control problems is based on stochastic dynamic programming [1], [7]. This method results in globally optimal feedback policies but with the cost of, exponential in the number of states, requirements in terms of memory and computational complexity. Alternatively, the use of PGs methods for nonlinear stochastic optimal control compromises global optimality in favor of feasibility and scalability especially for dynamical systems with many states and degrees of freedom. Moreover there are ways to optimize feedback policies with PGs by treating control gains as parameters and then optimizing these parameters to improve performance.

Despite the vast amount of work on PGs [2]–[6], [10], [11], most algorithms are derived in discrete time and for linear policy parameterizations. In this paper we extent our recent work on free energy policy gradients [9] to policy parameterization with time varying parameters. The resulting update rules are different when compared against policy parameterizations with time independent parameter. We provide results in the form of propositions for a number of cases of PGs depending on the form of the nonlinearity of the stochastic dynamics and the policy parameterization.

The paper is organized as follows: in Section (II) we provide the formulation of the problem. In Section (III) we derive the nonlinear and time varying nonlinear PGs and discuss special cases that include cost function of the form of free energy. Section (IV) includes the cases of linear time varying PGs. In Section (V) we provide a numerical examples and in Section (VI) we conclude.

[1] Evangelos A. Theodorou is Assistant Professor with the School of Aerospace Engineering, Georgia Institute of Technology, Atlanta. evangelos.theodorour@ae.gatech.edu

[2] Krishnamurthy Dvijotham is graduate student in Department of Computer Science and Engineering, University of Washington, Seattle. dvij@cs.washington.edu

[3] Emo Todorov is Associate professor with the Departments of Computer Science and Engineering, and Applied Math, University of Washington, Seattle. todorov@cs.washington.edu

II. Problem Formulation

We consider the stochastic dynamical system of the form:

$$d\mathbf{x}(t) = \mathbf{F}(\mathbf{x}, \mathbf{u}, t)dt + \mathbf{C}(\mathbf{x}, t)d\mathbf{w}(t) \tag{1}$$

in which $\mathbf{x} \in \Re^n$ is the state, $\mathbf{u} \in \Re^p$ the controls and $d\mathbf{w} \in \Re^n$ is brownian noise. The functions $\mathbf{F}(\mathbf{x}, \mathbf{u}, t)$ and $\mathbf{C}(\mathbf{x}, t)$ are defined as $\mathbf{F} : \Re^n \times \Re^p \times \Re \rightarrow \Re^n$ and $\mathbf{C}(\mathbf{x}, t) : \Re^n \times \Re \rightarrow \Re^n \times \Re^{n \times n}$.

In this paper we will assume that control is parameterized as $\mathbf{u} = \mathbf{\Phi}(\mathbf{x}, \boldsymbol{\theta}(t))$ with the term $\boldsymbol{\theta}(t)$ denoting the time varying parameter. Let $\vec{\mathbf{x}} = [\mathbf{x}(t_0), ..., \mathbf{x}(t_N)]$ and $\vec{\mathbf{\Theta}}$ containing all the parameters $\boldsymbol{\theta}(t_i) \in \Re^{\nu \times 1}, \forall t_i \in [t_0, t_{N-1}]$ and thus $\vec{\mathbf{\Theta}} = [\boldsymbol{\theta}(t_0), \boldsymbol{\theta}(t_1), ..., \boldsymbol{\theta}(t_{N-1})]$ with $\vec{\mathbf{\Theta}} \in \Re^{\nu \times N}$. Moreover let $\vec{\mathbf{u}}$ denoting the control trajectory during the time horizon from t_0 to t_N. Therefore $\vec{\mathbf{u}} = [\mathbf{\Phi}(\mathbf{x}(t_0), \boldsymbol{\theta}(t_0)), ..., \mathbf{\Phi}(\mathbf{x}(t_{N-1}), \boldsymbol{\theta}(t_{N-1}))]$.

We consider the objective function $\xi(\mathbf{x}(t_0), \vec{\mathbf{\Theta}})$ where $\mathbf{x}(t_0)$ is the initial state of the trajectories $\vec{\mathbf{x}}$ sampled based on (1). The objective function is defined as follows:

$$\begin{aligned}\xi(\mathbf{x}(t_0), \vec{\mathbf{\Theta}}) &= \frac{1}{\rho} \log J(\mathbf{x}(t_0), \vec{\mathbf{\Theta}}) \\ &= \frac{1}{\rho} \log \int S\Big(\vec{\mathbf{x}}, \vec{\mathbf{u}}(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}}\Big) d\mathbb{P}(\vec{\mathbf{x}}; \vec{\mathbf{\Theta}})\end{aligned}$$

The term $J(\mathbf{x}(t_0), \vec{\mathbf{\Theta}})$ is defined as $J(\mathbf{x}(t_0), \vec{\mathbf{\Theta}}) = \mathbb{E}_{\mathbb{P}(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}})}\Big(S(\vec{\mathbf{x}}, \vec{\mathbf{u}})\Big)$. The symbol $\mathbb{E}_{\mathbb{P}(\vec{\mathbf{x}}; \vec{\mathbf{\Theta}})}$ denotes the expectation under the probability measure $\mathbb{P}(\vec{\mathbf{x}}; \vec{\mathbf{\Theta}})$ which corresponds to (1) and it is parameterized by $\vec{\mathbf{\Theta}}$ due to control parameterization. Thus the expectation $\mathbb{E}_{\mathbb{P}(\vec{\mathbf{x}}; \vec{\mathbf{\Theta}})}$ is taken with respect to trajectories generated by (1). The term $S(\vec{\mathbf{x}}, \vec{\mathbf{u}}(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}})$ is a functional which depends on the state and control trajectories. The case in which $S(\vec{\mathbf{x}}, \vec{\mathbf{u}}(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}})$ is defined as $S(\vec{\mathbf{x}}, \vec{\mathbf{u}}(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}}) = \exp\Big(\rho \mathcal{L}(\vec{\mathbf{x}}, \vec{\mathbf{u}}(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}})\Big)$ is of particular interest. Under this definition the objective function (2) is transformed to:

$$\xi(\mathbf{x}(t_0), \vec{\mathbf{\Theta}}) = \frac{1}{\rho} \log \int \exp\Big(\rho \mathcal{L}(\vec{\mathbf{x}}, \vec{\mathbf{u}}(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}})\Big) d\mathbb{P}(\vec{\mathbf{x}}; \vec{\mathbf{\Theta}}) \tag{2}$$

In the next section we find the gradient of the function above with respect to parameters $\vec{\mathbf{\Theta}}$. Since the control sequence $\vec{\mathbf{u}}$ depends on the state trajectories $\vec{\mathbf{x}}$ and parameters $\vec{\mathbf{\Theta}}$ in the rest of the analysis we use the notation $\mathcal{L}(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}})$ and $\mathcal{S}(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}})$ for $\mathcal{L}(\vec{\mathbf{x}}, \vec{\mathbf{u}}(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}}))$ and $\mathcal{S}(\vec{\mathbf{x}}, \vec{\mathbf{u}}(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}}))$ respectively.

III. Time Varying Nonlinear Policy Gradients.

Let $\delta\vec{\mathbf{\Theta}}^{(j)}(t_i)$ defined as $\delta\vec{\mathbf{\Theta}}^{(j)}(t_i) = \delta\theta[0_{\nu\times 1}, ..., \mathbf{e}^{(j)}(t_i), ..., 0_{\nu\times 1}]$ with $\mathbf{e}^{(j)}(t_i)$ a vector with zero elements besides the j^{th} element that equals 1 and therefore $\mathbf{e}^{(j)}(t_i) = [0, ...1, ...0]^T$. The perturbation $\delta\vec{\mathbf{\Theta}}^{(j)}(t_i)$ corresponds to the variation of the j^{th} parameter at time instant t_i. Next we consider the gradient of the objective function $\xi(\mathbf{x}(t_0), \vec{\mathbf{\Theta}})$ as expressed in (2) and we will have:

$$\lim_{\delta\theta_j(t_i)\to 0} \frac{\delta\xi(\mathbf{x}(t_0), \vec{\mathbf{\Theta}})}{\delta\theta_j(t_i)} = \frac{1}{\rho J(\mathbf{x}(t_0), \vec{\mathbf{\Theta}})} \lim_{\delta\theta_j(t_i)\to 0} \frac{\delta J(\mathbf{x}, \vec{\mathbf{\Theta}})}{\delta\theta_j(t_i)}$$
$$= \frac{1}{\rho J(\mathbf{x}(t_0), \vec{\mathbf{\Theta}})}$$
$$\times \lim_{\delta\theta_j(t_i)\to 0} \left(\frac{J(\mathbf{x}(t_0), \vec{\mathbf{\Theta}} + \delta\vec{\mathbf{\Theta}}^{(j)}(t_i)) - J(\mathbf{x}, \vec{\mathbf{\Theta}})}{\delta\theta_j(t_i)} \right) \quad (3)$$

We work with the expression inside the parenthesis in the last equation. To keep the notation short we also define $\mathbb{P}_0 = \mathbb{P}(\vec{\mathbf{x}}; \vec{\mathbf{\Theta}})$ and $\mathbb{P}_1 = \mathbb{P}(\vec{\mathbf{x}}; \vec{\mathbf{\Theta}} + \delta\vec{\mathbf{\Theta}}^{(j)}(t_i))$ the probability measures corresponding to trajectories generated by (1) under the policy parameterization $\vec{\mathbf{u}}\left(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}}\right)$ and $\vec{\mathbf{u}}\left(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}} + \delta\vec{\mathbf{\Theta}}^{(j)}(t_i)\right)$. More precisely we will have that:

$$\frac{\delta J}{\delta\theta_j(t_i)} = \frac{\mathbb{E}_{\mathbb{P}_1}\left(S(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}} + \delta\vec{\mathbf{\Theta}}^{(j)}(t_i)) \right) - \mathbb{E}_{\mathbb{P}_0}\left(S(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}}) \right)}{\delta\theta_j(t_i)}$$
$$= \frac{\mathbb{E}_{\mathbb{P}_0}\left(S(\mathbf{x}, \vec{\mathbf{\Theta}} + \delta\vec{\mathbf{\Theta}}^{(j)}(t_i)) \frac{d\mathbb{P}_1}{d\mathbb{P}_0} \right) - \mathbb{E}_{\mathbb{P}_0}\left(S(\mathbf{x}, \vec{\mathbf{\Theta}}) \right)}{\delta\theta_j(t_i)}$$
$$= \frac{\mathbb{E}_{\mathbb{P}_0}\left(S(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}} + \delta\vec{\mathbf{\Theta}}^{(j)}(t_i)) \frac{d\mathbb{P}_1}{d\mathbb{P}_0} - S(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}}) \right)}{\delta\theta_j(t_i)}$$

We add and subtract the term $\frac{1}{\delta\theta_j(t_i)} S(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}}) \frac{d\mathbb{P}_1}{d\mathbb{P}_0}$ in the equation above. Therefore we will have the expression that follows:

$$\frac{\delta J}{\delta\theta_j(t_i)} =$$
$$= \mathbb{E}_{\mathbb{P}_0}\left(\frac{S(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}} + \delta\vec{\mathbf{\Theta}}^{(j)}(t_i)) \frac{d\mathbb{P}_1}{d\mathbb{P}_0} - S(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}}) \frac{d\mathbb{P}_1}{d\mathbb{P}_0}}{\delta\theta_j(t_i)} \right)$$
$$+ \mathbb{E}_{\mathbb{P}_0}\left(\frac{S(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}}) \frac{d\mathbb{P}_1}{d\mathbb{P}_0} - S(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}})}{\delta\theta_j(t_i)} \right)$$

Thus the resulting gradient of J takes the form:

$$\frac{\delta J}{\delta\theta_j(t_i)} =$$
$$= \frac{\mathbb{E}_{\mathbb{P}_0}\left[\left(S(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}} + \delta\vec{\mathbf{\Theta}}^{(j)}(t_i)) - S(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}}) \right) \frac{d\mathbb{P}_1}{d\mathbb{P}_0} \right]}{\delta\theta_j(t_i)}$$
$$+ \frac{\mathbb{E}_{\mathbb{P}_0}\left[S(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}}) \left(\frac{d\mathbb{P}_1}{d\mathbb{P}_0} - 1 \right) \right]}{\delta\theta_j(t_i)}$$

We incorporate the last line into (3) and we have:

$$\lim_{\delta\theta(t_i)\to 0} \frac{\delta\xi(\mathbf{x}(t_0), \vec{\mathbf{\Theta}})}{\delta\theta(t_i)} = \frac{1}{\rho J(\mathbf{x}(t_0), \vec{\mathbf{\Theta}})} \mathfrak{M}(\mathbf{x}(t_0), \vec{\mathbf{\Theta}}) \quad (4)$$

where the term $\mathfrak{M}(\mathbf{x}, \vec{\mathbf{\Theta}})$ is defined as follows:

$$\mathfrak{M}(\mathbf{x}(t_0), \vec{\mathbf{\Theta}}) = \lim_{\delta\theta(t_i)\to 0} \left(\mathbb{E}_{\mathbb{P}_0}\left[\frac{\delta S(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}})}{\delta\theta_j(t_i)} \frac{d\mathbb{P}_1}{d\mathbb{P}_0} \right] \right)$$
$$+ \lim_{\delta\theta(t_i\to 0} \left(\mathbb{E}_{\mathbb{P}_0}\left[\frac{S(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}})}{\delta\theta_j(t_i)} \left(\frac{d\mathbb{P}_1}{d\mathbb{P}_0} - 1 \right) \right] \right) \quad (5)$$

To find the limit in both terms in the expression above as $\delta\theta_j(t_i) \to 0$ we will make use of the Radon Nikodým derivative $\frac{d\mathbb{P}_1}{d\mathbb{P}_0}$ as applied to nonlinear diffusion processes. More precisely based on our analysis in section (VII) we have the expression:

$$\frac{d\mathbb{P}_1}{d\mathbb{P}_0} = \exp\left[\delta\mathbf{F}(t_i)^T \mathbf{C}(\mathbf{x}(t_i), t_i)^{-T} d\mathbf{w}_\theta(t_i) \right]$$
$$\times \exp\left[-\frac{1}{2} \delta\mathbf{F}(t_i)^T \mathbf{\Sigma}_\mathbf{C}^{-1} \delta\mathbf{F}(t_i) dt \right] \quad (6)$$

The term $\delta\mathbf{F}(t_i)$ above is defined as the difference $\delta\mathbf{F}(t) = \mathbf{F}(\mathbf{x}, \mathbf{u}(\mathbf{x}, \boldsymbol{\theta}(t) + \delta\theta\mathbf{e}_j), t) - \mathbf{F}(\mathbf{x}, \mathbf{u}(\mathbf{x}, \boldsymbol{\theta}(t)), t)$ and we assume that $\lim_{\delta\boldsymbol{\theta}\to 0} \delta\mathbf{F} = 0$. In addition the term $\mathbf{\Sigma}_\mathbf{C}$ is defined as:

$$\mathbf{\Sigma}_C(\mathbf{x}, t) = \mathbf{C}(\mathbf{x}, t)\mathbf{C}(\mathbf{x}, t)^T \quad (7)$$

Next we find the terms $\lim_{\delta\theta_j(t_i)\to 0} \left(\frac{d\mathbb{P}_1}{d\mathbb{P}_0} \right)$ and $\lim_{\delta\theta_j(t_i)\to 0} \left(\frac{1}{\delta\theta_j(t_i)} \left(\frac{d\mathbb{P}_1}{d\mathbb{P}_0} - 1 \right) \right)$. Since $e^x - 1 = x + \frac{x^2}{2!} + \frac{x^2}{3!}$we will have that:

$$\frac{1}{\delta\theta_j(t_i)} \left(\frac{d\mathbb{P}_1}{d\mathbb{P}_0} - 1 \right) = T_1 + T_2 + T_3 + \text{High Order Terms} \quad (8)$$

where the terms T_1, T_2 and T_3 are defined as follows:

$$T_1 = \frac{\left(\delta\mathbf{F}(t_i)^T \mathbf{C}(\mathbf{x}(t_i), t_i)^{-T} d\mathbf{w}_\theta(t_i) - \frac{1}{2} \delta\mathbf{F}(t_i)^T \mathbf{\Sigma}_\mathbf{C}^{-1} \delta\mathbf{F}(t_i) dt \right)}{\delta\theta_j(t_i)}$$

$$T_2 = \frac{\left(\delta \mathbf{F}(t_i)^T \mathbf{C}(\mathbf{x}(t_i,t_i)^{-T} d\mathbf{w}_\theta(t_i) - \frac{1}{2}\delta\mathbf{F}(t_i)^T \mathbf{\Sigma}_{\mathbf{C}}^{-1} \delta \mathbf{F}(t_i) dt\right)^2}{2\delta\theta_j(t_i)}$$

and $T_3 = \frac{1}{\delta\theta_j(t_i)} O(\delta \mathbf{F}^3)$. We take the limit [12] of the expression (8) which results in the following expression:

$$\lim_{\delta\theta_j(t_i)\to 0} \frac{1}{\delta\theta_j(t_i)} \left(\frac{d\mathbb{P}_1}{d\mathbb{P}_0} - 1\right) = \\ = \left(\frac{\delta \mathbf{F}(t_i)}{\delta\theta_j(t_i)}\right)^T \mathbf{C}(\mathbf{x}(t_i),t_i)^{-T} d\mathbf{w}_\theta(t_i)$$

Also $\lim_{\delta\theta(t_i)\to 0}\left(\frac{d\mathbb{P}_1}{d\mathbb{P}_0}\right) = 1$ and therefore the final result is:

$$\lim_{\delta\theta_j(t_i)\to 0} \frac{\delta \xi(\mathbf{x}(t_0), \vec{\mathbf{\Theta}}))}{\delta\theta(t_i)} = \\ \frac{\left(\mathbb{E}_{\mathbb{P}_0}\left[\frac{\delta S(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}})}{\delta\theta(t_i)}\right] + \mathbb{E}_{\mathbb{P}_0}\left[S(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}})\delta\pi(t_i)\right]\right)}{\rho J(\mathbf{x}, \vec{\mathbf{\Theta}}))} \tag{9}$$

where the term $\delta\pi_j(t_i)$ is defined as:

$$\delta\pi_j(t_i) = \left(\frac{\delta \mathbf{F}(t_i)}{\delta\theta_j(t_i)}\right)^T \mathbf{C}(\mathbf{x}(t_i),t_i)^{-T} d\mathbf{w}_\theta(t_i)$$

The gradient of the objective function with respect to $\boldsymbol{\theta}(t_i) = (\theta_1(t_i), \theta_2(t_i), ..., \theta_\nu(t_i))^T$ is given as follows:

$$\nabla_{\boldsymbol{\theta}(t_i)} \xi(\mathbf{x}(t_0), \vec{\mathbf{\Theta}})) = \frac{1}{\rho J(\mathbf{x}(t_0), \vec{\mathbf{\Theta}}))} \times \\ \left(\mathbb{E}_{\mathbb{P}_0}\left[\nabla_{\boldsymbol{\theta}(t_i)} S(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}})\right] + \mathbb{E}_{\mathbb{P}_0}\left[S(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}})\delta\boldsymbol{\pi}(t_i)\right]\right) \tag{10}$$

where the vector $\delta\boldsymbol{\pi}(t_i) \in \Re^{\nu\times 1}$ is defined as:

$$\delta\boldsymbol{\pi}(t_i) = \left(\frac{\delta \mathbf{F}(t_i)}{\delta\boldsymbol{\theta}(t_i)}\right)^T \mathbf{C}(\mathbf{x}(t_i),t_i)^{-T} d\mathbf{w}_\theta(t_i) \tag{11}$$

or as:

$$\delta\boldsymbol{\pi}(t_i) = \left(\mathbf{J}_{\boldsymbol{\theta}(t_i)}\mathbf{F}\right)^T \mathbf{C}(\mathbf{x},t_i)^{-T} d\mathbf{w}_\theta(t_i) \tag{12}$$

where $\mathbf{J}_{\boldsymbol{\theta}(t_i)}\mathbf{F}$ is the Jacobian of the drift term $\mathbf{F}(\mathbf{x}, \mathbf{u}(\mathbf{x}, \boldsymbol{\theta}(t), t)$ of the stochastic dynamics with respect to parameters $\boldsymbol{\theta}(t_i)$. Note that this result is different from the case in which the policy is parameterized as $\mathbf{u} = \mathbf{\Phi}(\mathbf{x}(t), \boldsymbol{\theta})$ and the parameter $\boldsymbol{\theta}$ time independent [9].

We summarize the analysis above regarding the time varying nonlinear policy gradients with the following theorem:

Theorem 1: (Nonlinear Policy Gradient) *Consider the objective function:*

$$\xi(\mathbf{x}(t_0), \vec{\mathbf{\Theta}}) = \frac{1}{\rho}\log \int S(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}}) d\mathbb{P}(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}}) \tag{13}$$

subject to:

i) the nonlinear stochastic dynamics affine in noise expressed as in (1) and,

ii) the nonlinear and time varying policy parameterization of the form $\mathbf{u} = \mathbf{\Phi}(\mathbf{x}, \boldsymbol{\theta}(t))$.

The policy gradient for this objective function is given by equations (10) and (12).

For the cases where the functional $S(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}})$ is defined as $S(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}}) = S_{t_o}^{t_N}(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}}) = \int_{t_0}^{t_N} q(\mathbf{x}(t), \boldsymbol{\theta}(t))dt = S_{t_o}^{t_i}(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}}) + S_{t_i}^{t_N}(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}})$ the gradient in (10) is expressed as follows:

$$\nabla_{\boldsymbol{\theta}(t_i)} \xi(\mathbf{x}(t_0), \vec{\mathbf{\Theta}})) = \frac{1}{\rho J(\mathbf{x}(t_0), \vec{\mathbf{\Theta}}))} \times \\ \left(\mathbb{E}_{\mathbb{P}_0}\left[\nabla_{\boldsymbol{\theta}(t_i)} q(\mathbf{x}(t_i), \boldsymbol{\theta}(t_i))dt\right] + \mathbb{E}_{\mathbb{P}_0}\left[S_{t_i}^{t_N}(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}})\delta\boldsymbol{\pi}(t_i)\right]\right)$$

where $S_{t_i}^{t_N}(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}}) = S(\vec{\mathbf{x}}(t_i \to t_N), \vec{\mathbf{\Theta}}(t_i \to t_{N-1}))$ is the cost accumulated starting from time t_i to t_N while $\vec{\mathbf{x}}(t_i \to t_N)$ and $\vec{\mathbf{\Theta}}(t_i \to t_{N-1})$ are the states and parameters starting from time t_i to t_N. To get the result above we make use of the fact that:

$$\mathbb{E}_{\mathbb{P}_0}\left[S_{t_0}^{t_i}(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}})\delta\boldsymbol{\pi}(t_i)\right] = \mathbb{E}_{\mathbb{P}_0}\left[S_{t_0}^{t_i}(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}})\right]\mathbb{E}_{\mathbb{P}_0}\left[\delta\boldsymbol{\pi}(t_i)\right]$$

and

$$\mathbb{E}_{\mathbb{P}_0}\left[\delta\boldsymbol{\pi}(t_i)\right] = \mathbb{E}_{\mathbb{P}_0}\left[\left(\frac{\delta \mathbf{F}}{\delta\boldsymbol{\theta}(t_i)}\right)^T \mathbf{C}(\mathbf{x}(t_i),t_i)^{-T}\right] \\ \times \mathbb{E}_{\mathbb{P}_0}\left[d\mathbf{w}_\theta(t_i)\right]$$

with the last term $\mathbb{E}_{\mathbb{P}_0}\left[d\mathbf{w}_\theta(t_i)\right] = 0$.

A. Nonlinear Risk Seeking/Sensitive Policy Gradients

We consider functionals $S(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}})$ of the form $S(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}}) = \exp\left(\rho\mathcal{L}(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}})\right)$. The gradient of the objective function can be formulated as follows:

$$\nabla_{\boldsymbol{\theta}(t_i)} \xi(\mathbf{x}(t_0), \vec{\mathbf{\Theta}}) = \frac{1}{\rho J(\mathbf{x}(t_0), \vec{\mathbf{\Theta}})} \times \\ \left(\mathbb{E}_{\mathbb{P}_0}\left[S(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}})\delta\boldsymbol{\pi}(t_i)\right] + \mathbb{E}_{\mathbb{P}_0}\left[\nabla_{\boldsymbol{\theta}(t_i)} S(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}})\right]\right) \\ = \frac{1}{J(\mathbf{x}(t_0), \vec{\mathbf{\Theta}})} \times \\ \left(\frac{1}{\rho}\mathbb{E}_{\mathbb{P}_0}\left[S(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}})\delta\boldsymbol{\pi}(t_i)\right] + \mathbb{E}_{\mathbb{P}_0}\left[\nabla_{\boldsymbol{\theta}(t_i)}\mathcal{L}(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}})S(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}})\right]\right) \\ = \frac{1}{\rho}\mathbb{E}_{\mathbb{Q}(\vec{\mathbf{x}};\boldsymbol{\theta})}\left[\delta\boldsymbol{\pi}(t_i)\right] + \mathbb{E}_{\mathbb{Q}(\vec{\mathbf{x}};\boldsymbol{\theta})}\left[\nabla_{\boldsymbol{\theta}(t_i)}\mathcal{L}(\vec{\mathbf{x}}, \vec{\mathbf{\Theta}})\right]$$

The term $\mathbb{Q}(\vec{\mathbf{x}}; \boldsymbol{\theta})$ is defined as follows:

$$d\mathbb{Q}(\vec{\mathbf{x}};\vec{\boldsymbol{\Theta}}) = \frac{S(\vec{\mathbf{x}},\vec{\boldsymbol{\Theta}})d\mathbb{P}_0(\vec{\mathbf{x}};\vec{\boldsymbol{\Theta}})}{\int S(\vec{\mathbf{x}},\vec{\boldsymbol{\Theta}})d\mathbb{P}_0(\vec{\mathbf{x}};\vec{\boldsymbol{\Theta}})} = \frac{\exp\Big(\rho\mathcal{L}(\vec{\mathbf{x}},\vec{\boldsymbol{\Theta}})\Big)d\mathbb{P}_0(\vec{\mathbf{x}};\vec{\boldsymbol{\Theta}})}{\int \exp\Big(\rho\mathcal{L}(\vec{\mathbf{x}},\vec{\boldsymbol{\Theta}})\Big)d\mathbb{P}_0(\vec{\mathbf{x}};\vec{\boldsymbol{\Theta}})} \tag{14}$$

The analysis above is summarized by the following proposition:

Proposition 1: *Consider the objective function:*

$$\xi(\mathbf{x}(t_0),\boldsymbol{\theta}) = \frac{1}{\rho}\log\int \exp\Big(\rho\mathcal{L}(\vec{\mathbf{x}},\vec{\boldsymbol{\Theta}})\Big)d\mathbb{P}(\vec{\mathbf{x}},\vec{\boldsymbol{\Theta}}) \tag{15}$$

subject to:

i) the nonlinear stochastic dynamics affine in noise expressed as in (1) and,

ii) the nonlinear and time varying policy parameterization of the form $\mathbf{u} = \boldsymbol{\Phi}(\mathbf{x},\boldsymbol{\theta}(t))$.

The policy gradient for this objective function is given as:

$$\nabla_{\boldsymbol{\theta}_{(t_i)}}\xi(\mathbf{x}(t_0),\boldsymbol{\theta}) = = \frac{1}{\rho}\mathbb{E}_{\mathbb{Q}(\vec{\mathbf{x}};\boldsymbol{\theta})}\Big[\delta\boldsymbol{\pi}(t_i)\Big] + \mathbb{E}_{\mathbb{Q}(\vec{\mathbf{x}};\boldsymbol{\theta})}\Big[\nabla_{\boldsymbol{\theta}_{(t_i)}}\mathcal{L}(\mathbf{x},\boldsymbol{\theta})\Big]$$

The term $\delta\boldsymbol{\pi}(t_i)$ *is defined as in (11) while the expectation* $\mathbb{E}_{\mathbb{Q}(\vec{\mathbf{x}};\boldsymbol{\theta})}$ *is under* $d\mathbb{Q}(\vec{\mathbf{x}};\boldsymbol{\theta})$ *that is defined in (14).*

IV. Linear Time Varying Policy Gradients.

For linear and time varying policy linear parameterizations we will have that:

$$\mathbf{u}(\mathbf{x}(t),\boldsymbol{\theta}(t)) = \Psi(\mathbf{x})\boldsymbol{\theta}(t) \tag{16}$$

In this case the policy gradient has the same expression as in (10) but now the term $\delta\boldsymbol{\pi}(t_i)$ is defined as:

$$\delta\boldsymbol{\pi}(t_i) = \left(\frac{\delta\mathbf{F}(t_i)}{\delta\mathbf{u}(t_i)}\frac{\delta\mathbf{u}}{\delta\boldsymbol{\theta}(t_i)}\right)^T\mathbf{C}(\mathbf{x}(t_i),t_i))^{-T}d\mathbf{w}_\theta(t_i) = \left(\frac{\delta\mathbf{F}(t_i)}{\delta\mathbf{u}(t_i)}\Psi(\mathbf{x}(t_i))\right)^T\mathbf{C}(\mathbf{x}(t_i),t_i)^{-T}d\mathbf{w}_\theta(t_i) \tag{17}$$

A special class of systems of this form in (1) may include dynamics affine in controls. Such dynamics are formulated as follows:

$$d\mathbf{x}(t) = \mathbf{f}(\mathbf{x},t)dt + \boldsymbol{\mathcal{B}}(\mathbf{x},t)\mathbf{u}dt + \frac{1}{\sqrt{\lambda}}\mathbf{L}(\mathbf{x},t)d\mathbf{w}(\mathbf{x},t) \tag{18}$$

For stochastic systems as in (18) we will have that:

$$\delta\boldsymbol{\pi}(t_i) = \left(\frac{\delta\mathbf{F}(t_i)}{\delta\mathbf{u}(t_i)}\frac{\delta\mathbf{u}}{\delta\boldsymbol{\theta}(t_i)}\right)^T\mathbf{C}(\mathbf{x}(t_i),t_i))^{-T}d\mathbf{w}_\theta(t_i) = \sqrt{\lambda}\Big(\boldsymbol{\mathcal{B}}(\mathbf{x},t)\Psi(\mathbf{x}(t_i))\Big)^T\mathbf{L}(\mathbf{x}(t_i),t_i)^{-T}d\mathbf{w}_\theta(t_i)$$

These results are summarized by the proposition that follows.

Proposition 2: *The gradient of the objective function in (13), subject to*

i) the nonlinear stochastic dynamics affine in control and noise expressed as in (18),

ii) the linear and time varying policy parameterization in (16),

is expressed as:

$$\nabla_{\boldsymbol{\theta}_{(t_i)}}\xi(\mathbf{x}(t_0),\vec{\boldsymbol{\Theta}})) = \frac{1}{\rho J(\mathbf{x}(t_0),\vec{\boldsymbol{\Theta}}))}\times \Big(\mathbb{E}_{\mathbb{P}_0}\Big[\nabla_{\boldsymbol{\theta}_{(t_i)}}S(\vec{\mathbf{x}},\vec{\boldsymbol{\Theta}})\Big] + \sqrt{\lambda}\mathbb{E}_{\mathbb{P}_0}\Big[S(\vec{\mathbf{x}},\vec{\boldsymbol{\Theta}})\delta\boldsymbol{\pi}^*(t_i)\Big]\Big) \tag{19}$$

The term $\delta\boldsymbol{\pi}^*(t_i)$ *above is defined as :*

$$\delta\boldsymbol{\pi}^*(t_i) = \Psi(\mathbf{x}(t_i))^T\boldsymbol{\mathcal{B}}(\mathbf{x}(t_i),t_i)^T\mathbf{L}(\mathbf{x}(t_i),t_i)^{-T}d\mathbf{w}_\theta(t_i) \tag{20}$$

Similarly to the previous section when $S(\vec{\mathbf{x}},\vec{\boldsymbol{\Theta}})$ is defined as $S(\vec{\mathbf{x}},\vec{\boldsymbol{\Theta}}) = S_{t_o}^{t_N}(\vec{\mathbf{x}},\vec{\boldsymbol{\Theta}}) = \int_{t_0}^{t_N} q(\mathbf{x}(t),\boldsymbol{\theta}(t))dt = S_{t_o}^{t_i}(\vec{\mathbf{x}},\vec{\boldsymbol{\Theta}}) + S_{t_i}^{t_N}(\vec{\mathbf{x}},\vec{\boldsymbol{\Theta}})$ the PG in proposition 2 will take the form:

$$\nabla_{\boldsymbol{\theta}_{(t_i)}}\xi(\mathbf{x}(t_0),\vec{\boldsymbol{\Theta}})) = \frac{1}{\rho J(\mathbf{x}(t_0),\vec{\boldsymbol{\Theta}}))}\times \Big(\mathbb{E}_{\mathbb{P}_0}\Big[\nabla_{\boldsymbol{\theta}_{(t_i)}}q(\mathbf{x},\boldsymbol{\theta}(t))dt\Big] + \sqrt{\lambda}\mathbb{E}_{\mathbb{P}_0}\Big[S_{t_i}^{t_N}(\vec{\mathbf{x}},\vec{\boldsymbol{\Theta}})\delta\boldsymbol{\pi}^*(t_i)\Big]\Big) \tag{21}$$

For the cases where the functional $S(\vec{\mathbf{x}},\vec{\boldsymbol{\Theta}})$ takes the form $S(\vec{\mathbf{x}},\vec{\boldsymbol{\Theta}}) = \exp\Big(\rho\mathcal{L}(\vec{\mathbf{x}},\vec{\boldsymbol{\Theta}})\Big)$ we have the following proposition which provides risk seeking and risk sensitive PGs depending on the sign of the parameter ρ. More precisely:

Proposition 3: *The gradient of the objective function in (15), subject to*

i) the nonlinear stochastic dynamics affine in control and noise expressed as in (18),

ii) the linear and time varying policy parameterization in (16),

is expressed as:

$$\nabla_{\boldsymbol{\theta}}\xi(\mathbf{x}(t_0),\boldsymbol{\theta}) = = \frac{\sqrt{\lambda}}{\rho}\mathbb{E}_{\mathbb{Q}(\vec{\mathbf{x}};\boldsymbol{\theta})}\Big[\delta\boldsymbol{\pi}^*(t_i)\Big] + \mathbb{E}_{\mathbb{Q}(\vec{\mathbf{x}};\boldsymbol{\theta})}\Big[\nabla_{\boldsymbol{\theta}}\mathcal{L}(\mathbf{x},\boldsymbol{\theta})\Big]$$

The term $\delta\boldsymbol{\pi}^*(t_i)$ *is defined as in (20) while the expectation* $\mathbb{E}_{\mathbb{Q}(\vec{\mathbf{x}};\boldsymbol{\theta})}$ *is under* $d\mathbb{Q}(\vec{\mathbf{x}};\boldsymbol{\theta})$ *defined as in (14). More over for* $\rho > 0$ *and* $\rho < 0$ *the gradient above corresponds to a risk sensitive and risk seeking version of the objective function in (15).*

Further simplifications of proposition 3 can be found regarding risk seeking and risk sensitive PGs when $\rho = \pm\sqrt{\lambda}$. Furthermore in the cases where the term $\mathcal{L}$ is independent of the policy parameters $\vec{\boldsymbol{\Theta}}(t)$ meaning that $\mathcal{L}(\vec{\mathbf{x}}(t),\vec{\boldsymbol{\Theta}}) = \mathcal{L}(\mathbf{x})$ then the PG in proposition 3 simplifies to:

$$\nabla_{\boldsymbol{\theta}}\xi(\mathbf{x}(t_0),\boldsymbol{\theta}) = \pm\, \mathbb{E}_{\mathbb{Q}(\vec{\mathbf{x}};\boldsymbol{\theta})}\Big[\delta\boldsymbol{\pi}^*(t_i)\Big]$$

In the next section we provide numerical examples for nonlinear stochastic control based on sampling.

V. Examples

We apply the PG of proposition 2 in a iterative form. The parameter updates are expressed as follows:

$$\boldsymbol{\theta}_{k+1}(t_i) = \boldsymbol{\theta}_k(t_i) - \gamma\frac{\partial\xi(\mathbf{x},\vec{\boldsymbol{\Theta}})}{\partial\boldsymbol{\theta}(t_i)}, \quad \forall t_i \in [t_0, t_{t_{N-1}}]$$

with $\gamma > 0$ playing the role of learning rate. We also consider the system with state multiplicative noise expressed as follows:

$$dx = Axdt + u(t,x)dt + \frac{1}{\sqrt{\lambda}}xdw(t)$$

The control is parameterized linearly with respect to the state and parameters. Thus it has the form $u(t,x) = K(t)x(t)$ with the policy parameter is $\boldsymbol{\theta}(t) = K(t)$. We define a cost for minimization as:

$$\xi = -\frac{1}{|\rho|}\log E\Big[\exp\Big(-|\rho|\int_{t_i}^{t_N}(x(t)-p(t)^*)^2dt\Big)\Big]$$

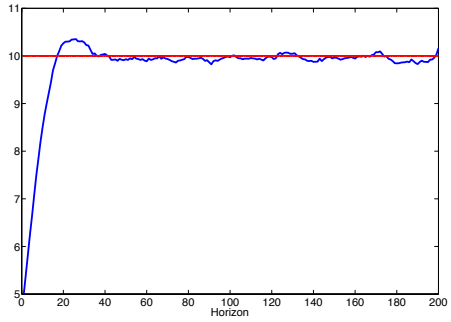
For our examples we used the values $\rho = -\sqrt{\lambda} = -30$, $A = 1$ and learning rate $\gamma = 1$. The task is to steer the state $x(t)$ to the target state trajectories $p(t)^* = 10$ and $p(t)^* = 2,\ \forall t \in [t_0, t_N]$, starting from $x(t_0) = 5$. Results of the average trajectories are shown in the Figure 1.

Next we increase the instability of the dynamics by increasing A to $A = 4$, while the desired state trajectory is a periodic movement $p(t) = 4 + cos(t)$. The resulting controlled trajectory is illustrated in Figure 2. Note that to track the sinusoidal target trajectory we did not have to make use of a special purpose nonlinear limit cycle attractor as in [9]. This is because the time varying characterization of the underlying control policy increases the capability of the policy to steer the actual trajectory to a target trajectory.

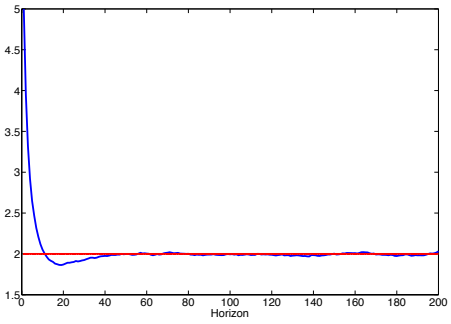
VI. Conclusion

In this paper we derive time varying PGs in continuous time for linear and nonlinear parameterizations. Future work should include evaluations of the proposed PGs on learning control application for systems with many dimensions and degrees of freedom and comparisons with other PG methods.

Another research direction is the use of the proposed PGs for training of nonlinear functions approximators such as Stochastic Neural Networks (SNN). PGs derived in this work could be applied to this setting for as long as the mathematical form of SNN could be represented by the stochastic differential equation in (1).



(a)



(b)

Fig. 1. Reaching target state, with the blue is the average trajectory, and with red is the desired state trajectory. In subfigure (a) $p^* = 10$ In (b) $p(t)^* = 2$.

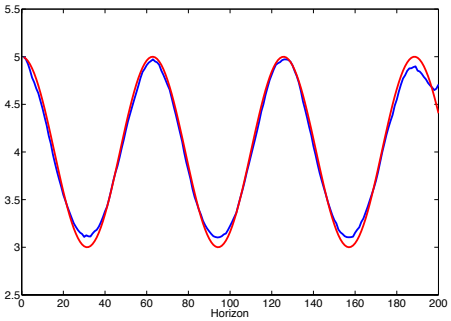


Fig. 2. Tracking a sinusoidal trajectory.

VII. Appendix

We will consider the nonlinear diffusions: $d\mathbf{x}(t) = \mathbf{F}_0\delta t + \mathbf{C}(\mathbf{x},t)d\mathbf{w}_\theta(t)$ and $d\mathbf{x}(t) = \mathbf{F}_1\delta t + \mathbf{C}(\mathbf{x},t)d\mathbf{w}_{\theta+\delta\theta}(t)$, with the terms defined as $\mathbf{F}_0 = \mathbf{F}(\mathbf{x},\mathbf{u}(\mathbf{x},\boldsymbol{\theta}(t)),t)$ and $\mathbf{F}_1 = \mathbf{F}(\mathbf{x},\mathbf{u}(\mathbf{x},\boldsymbol{\theta}+\delta\theta\mathbf{e}_j\delta(t-t_i)),t)$. The term $\mathbf{e}_j$ is a vector of zeros besides the j^{th} element that is equal to 1. We also define $\delta\mathbf{F} = \mathbf{F}_2 - \mathbf{F}_1$ and we expressed the corresponding probability measures as $\mathbb{P}_0 = \mathbb{P}\Big(\mathbf{x}_N, t_N|\mathbf{x}_0,t_0;\vec{\boldsymbol{\Theta}}\Big)$ and $\mathbb{P}_1 = \mathbb{P}\Big(\mathbf{x}_N, t_N|\mathbf{x}_0,t_0;\vec{\boldsymbol{\Theta}} + \delta\vec{\boldsymbol{\Theta}}^{(j)}(t_i)\Big)$. We consider the changes in the probability measure from $\mathbb{P}_0$ to $\mathbb{P}_1$ via the Girsanov transformation. These changes of probability mea-

sure correspond to the changes in the drift of the diffusion processes from $\mathbf{F}_0$ to $\mathbf{F}_1$. Using Girsanov's theorem [12] we have that:

$$\frac{d\mathbb{P}_0}{d\mathbb{P}_1} = \exp\left[\int_{t_0}^{t_N}\left(-\delta\mathbf{F}^T\mathbf{C}(\mathbf{x},t)^{-T}d\mathbf{w}_\theta(t)\right)\right] \times \exp\left[\int_{t_0}^{t_N}\left(+\frac{1}{2}\delta\mathbf{F}^T\boldsymbol{\Sigma}_\mathbf{C}^{-1}\delta\mathbf{F}\delta t\right)\right]$$

Since $\mathbf{F}_1 = \mathbf{F}\Big(\mathbf{x}, \mathbf{u}(\mathbf{x}, \boldsymbol{\theta} + \delta\theta\mathbf{e}_j\delta(t-t_i)), t\Big)$ we have that:

$$\begin{aligned}\delta\mathbf{F}(t) &= \Big(\mathbf{F}\Big(\mathbf{x}, \mathbf{u}(\mathbf{x}, \boldsymbol{\theta} + \delta\theta\mathbf{e}_j), t\Big) - \mathbf{F}\Big(\mathbf{x}, \mathbf{u}(\mathbf{x}, \boldsymbol{\theta}), t\Big)\Big) \\ &\quad \times \delta(t-t_i) \\ &= \delta\mathbf{F}(t)\delta(t-t_i)\end{aligned}$$

The Girsanov transformation will take the form:

$$\frac{d\mathbb{P}_0}{d\mathbb{P}_1} = \exp\left[\int_{t_0}^{t_N}\left(-\delta\mathbf{F}^T\mathbf{C}(\mathbf{x},t)^{-T}\delta(t-t_i)d\mathbf{w}_\theta(t)\right)\right] \times \exp\left[\int_{t_0}^{t_N}\left(+\frac{1}{2}\delta(t-t_i)\delta\mathbf{F}^T\boldsymbol{\Sigma}_\mathbf{C}^{-1}\delta\mathbf{F}\delta(t-t_i)\delta t\right)\right]$$

Thus the final results is:

$$\frac{d\mathbb{P}_0}{d\mathbb{P}_1} = \exp\left(-\delta\mathbf{F}(t_i)^T\mathbf{C}(\mathbf{x},t_i)^{-T}d\mathbf{w}_\theta(t_i)\right) \times \exp\left(+\frac{1}{2}\delta\mathbf{F}(t_i)^T\boldsymbol{\Sigma}_\mathbf{C}^{-1}\delta\mathbf{F}(t_i)\delta t\right)$$

REFERENCES

[1] R. Bellman. *Dynamic Programming*. Dover Publications, March 2003.

[2] M. Ghavamzadeh and Y Engel. Bayesian actor-critic algorithms. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 297–304, 2007.

[3] J. Kober and J. Peters. Policy search for motor primitives. In D. Schuurmans, J. Benigio, and D. Koller, editors, *Advances in Neural Information Processing Systems 21 (NIPS 2008)*. Cambridge, MA: MIT Press, 2008.

[4] J. Peters and S. Schaal. Learning to control in operational space. *International Journal of Robotics Research*, 27:197–212, 2008.

[5] J. Peters and S. Schaal. Natural actor critic. *Neurocomputing*, 71(7-9):1180–1190, 2008.

[6] J. Peters and S. Schaal. Reinforcement learning of motor skills with policy gradients. *Neural Networks*, 21(4):682–97, 2008.

[7] R. F. Stengel. *Optimal control and estimation*. Dover books on advanced mathematics. Dover Publications, New York, 1994.

[8] R. S. Sutton and A. G. Barto. *Reinforcement learning : An introduction*. Adaptive computation and machine learning. MIT Press, Cambridge, 1998.

[9] E. Theodorou, J. Najemnik, and E. Todorov. Free energy based policy gradients. In *Adaptive Dynamic Programming and Reinforcement Learning(To appear)*, 2013.

[10] M. Toussaint and A. Storkey. Probabilistic inference for solving discrete and continuous state markov decision processes, 2006.

[11] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

[12] J. Yang and J. H. Kushner. A monte carlo method for sensitivity analysis and parametric optimization of nonlinear stochastic systems. *SIAM Journal in Control and Optimization*, 29(5):1216–1249, 1991.