Policy gradient in continuous time

Rémi Munos

Centre de Mathématiques Appliquées,
Ecole Polytechnique, 91128 Palaiseau, France.
remi.munos@polytechnique.fr
www.cmap.polytechnique.fr/∼munos

Abstract : We consider the approach of solving approximately a deterministic optimal control problem by searching a good controller in a given class of parameterized policies.

When the dynamics of the system is known from the decision maker, an explicit representation of the sensitivity of the performance measure with respect to the control parameters is easily derived using pathwise derivation, which enables to use gradient methods for solving the parametric optimization problem.

This paper is concerned with the case of an *unknown state dynamics* (such as in reinforcement learning). It introduces a method for computing the policy gradient only from the observable. The underlying idea consists in approximating the continuous deterministic process by a stochastic discrete one and using stochastic policies to estimate the unknown coefficients by quantities that depend solely on the state and the policy. Almost sure convergence to the policy gradient is proved. The method is illustrated on a (6 dimensional) target problem.

Résumé : Nous considérons la résolution approchée d'un problème de contrôle optimal par un problème d'optimisation dans un espace de politiques paramétrées.

Lorsque la dynamique d'état du système est connue de l'agent décisionnel, une représentation explicite de la sensibilité (le gradient) de la mesure de performance par rapport aux paramètres de contrôle se déduit facilement par dérivation trajectorielle, ce qui permet d'utiliser une méthode de gradient pour résoudre le problème d'optimisation paramétrique.

Dans cet article, nous considérons le cas d'une *dynamique d'état inconnue* (cadre de l'apprentissage par renforcement). Nous décrivons une méthode pour calculer le gradient uniquement à partir des observables. L'idée sous-jacente consiste à approcher le processus déterministe continu par un processus stochastique discret et d'utiliser une politique stochastique pour remplacer les coefficients inconnus par des grandeurs qui ne dépendent que de l'état et de la politique. La convergence presque-sure vers le gradient est montrée. L'approche est illustrée sur un problème de cible (en dimension 6).

1 Introduction and statement of the problem

We consider an optimal control problem with continuous state $(x_t \in I\!R^d)_{t\geq 0}$ whose dynamics is defined by the differential equation:

$$\frac{dx_t}{dt} = f(x_t, u_t), \tag{1}$$

where the control $(u_t)_{t\geq 0}$ is a Lebesgue measurable function with values in a control space U, assumed to be finite (extension to a continuous control space is straightforward). The functional J to be maximized has finite-time horizon T. For simplicity, in the following, we illustrate the case of a terminal reward only:

$$J(x; (u_t)_{t\geq 0}) = r(x_T), \tag{2}$$

where $r : I\!R^d \to I\!R$ is the reward function. Extension to the general case of a functional of the form

$$J(x; (u_t)_{t\geq 0}) = \int_0^T r(t, x_t)dt + R(x_T), \tag{3}$$

with r and R being current and terminal reward functions, easily follows from linearity, as indicated in remark 1.

The optimal control problem of finding a control $(u_t)_{t\geq 0}$ that maximizes the functional is replaced by a parametric optimization problem in which we search for a good feed-back control law in a given class of parameterized policies $\{\pi_\alpha\}_\alpha$, where $\alpha \in I\!R^m$ represents the parameter. The control (or action) may be written $u_t = \pi_\alpha(t, x_t)$, and the dynamics of the resulting feed-back system is

$$\frac{dx_t}{dt} = f_\alpha(x_t), \tag{4}$$

where $f_\alpha(x_t) = f(x, \pi_\alpha(t, x))$. We assume that f_α is $\mathcal{C}^2$ with bounded derivatives. Let us write the performance measure

$$V(\alpha) = J(x; \pi_\alpha(t, x_t)_{t\geq 0})$$

to emphasize the dependency with respect to (w.r.t.) the parameter α. One may also consider an average performance measure defined by some distribution μ for the initial state: $V(\alpha) = \mathbb{E}[J(x; \pi_\alpha(t, x_t)_{t\geq 0})|x \sim \mu]$.

In order to find a local maximum of $V(\alpha)$, one may perform a gradient ascent method:

$$\alpha \leftarrow \alpha + \eta \nabla_\alpha V(\alpha), \tag{5}$$

with η being an adequate step (see for example (Polyak, 1987; Kushner & Yin, 1997)). Of course, many much powerful variants of gradient-based methods exists (Bonnans *et al.*, 2003).

The computation of the gradient $\nabla_\alpha V(\alpha)$ is the object of this paper.

Pathwise estimation of the gradient.

Define the gradient of the state with respect to the parameter: $z_t = \nabla_\alpha x_t$, which solves

$$\frac{dz_t}{dt} = \nabla_\alpha f_\alpha(x_t) + \nabla_x f_\alpha(x_t) z_t, \tag{6}$$

with initial condition $z_0 = 0$. Here, z_t is an $d \times m$-matrix whose (i,j)-component is the derivative of the ith component of x_t w.r.t. α_j. Similarly $\nabla_\alpha f_\alpha$ and $\nabla_x f_\alpha$ are, respectively, the derivatives of f w.r.t. the parameter (matrix of size $d \times m$) and the state (matrix of size $d \times d$).

When the reward function r is smooth, one may apply a pathwise differentiation to derive a gradient formula:

Proposition 1
(Yang & Kushner, 1991). If r is continuously differentiable then

$$\nabla_\alpha V(\alpha) = \nabla_x r(x_T) z_T. \tag{7}$$

Remark 1
In the more general setting of a functional (3), the gradient is:

$$\nabla_\alpha V(\alpha) = \int_0^T \nabla_x r(t, x_t) z_t dt + \nabla_x R(x_T) z_T.$$

What is known from the agent?

The decision maker (call it the agent) that intends to design a good controller for the dynamical system may or may not know a model of the state dynamics f. In case the dynamics is known, the state gradient $z_t = \nabla_\alpha x_t$ may be computed from (6) along the trajectory and the gradient of the performance measure w.r.t. the parameter α may be deduced at time T from (7), which allows to perform the gradient ascent step (5).

In this paper we investigate the case of an a priori unknown dynamics: the agent only observes the response of the system to its control. This specific framework is often referred to as *model-free reinforcement learning* (Sutton & Barto, 1998).

The available information to the agent at time t are its own control policy π_α and the trajectory $(x_s)_{0 \leq s \leq t}$ up to time t. At time T, it observes the reward $r(x_T)$ and, in this paper, we also assume that the gradient $\nabla r(x_T)$ is known.

From this point of view, it seems impossible to derive the state gradient z_t from (6), since $\nabla_\alpha f$ and $\nabla_x f$ are unknown. The term $\nabla_x f(x_t)$ may be approximated by least squares methods from the observation of past states $(x_s)_{s \leq t}$, as this will be explained later on in subsection 3.2. However the term $\nabla_\alpha f(x_t)$ cannot be computed analogously.

In this paper, we introduce the idea of using stochastic policies to approximate the state x_t and the state gradient z_t by discrete-time stochastic processes X_t^Δ and Z_t^Δ (with Δ being some time-discretization step). We show how Z_t^Δ may be computed without the knowledge of $\nabla_\alpha f$, but instead with likelihood ratios (such as $\nabla_\alpha \log \pi_\alpha$ and $\nabla_x \log \pi_\alpha$) of the policy.

We prove the convergence (with probability one) of the gradient estimate $\nabla_x r(X_T^\Delta) Z_T^\Delta$ derived from the stochastic processes to $\nabla_\alpha V(\alpha)$ when $\Delta \to 0$.

Remark 2
It is worth mentioning that this strong convergence result contrasts with usual likelihood ratio methods in discrete time (Williams, 1992; Baxter & Bartlett, 2001; Sutton et al., *2000; Marbach & Tsitsiklis, 2003) for which the policy gradient estimate would be subject to variance explosion when the number of discretization steps (thus the number of decisions before getting the reward) goes to infinity (i.e. $\Delta \to 0$).*

The paper is organized as follows. In Section 2, we state a general result for discretization of continuous deterministic dynamics by stochastic discrete processes and apply it to prove the convergence of the approximate state and state gradient. In Section 3, we state the convergence of the policy gradient estimate and describe the model-free reinforcement learning algorithm. In the last Section, we illustrate the method on a (6 dimensional) target problem. Appendices A and B provide all proofs.

2 Discretized stochastic processes

We first state a general result for approximating a deterministic continuous process by a stochastic discrete one. Then, we apply this result to the convergence analysis of processes (the state X_t^Δ and the state gradient Z_t^Δ) related to the introduction of stochastic policies.

2.1 A general convergence result

Let $(x_t)_{0\leq t\leq T}$ be a deterministic continuous process defined by some dynamics

$$\frac{dx_t}{dt} = f(x_t)$$

with initial condition $x_0 = x$. We assume that f is of class $\mathcal{C}^2$ with bounded derivatives. Let $\Delta = T/N$ be a time-discretization step (with N being the number of time-steps) and denote $\{t_n = n\Delta\}_{0\leq n\leq N}$ the discrete times.

Let $(X_{t_n}^\Delta)_{0\leq n<N}$ be a discrete stochastic process, such that $X_0^\Delta = x$, and which satisfies the consistency property:

$$\mathbb{E}[X_{t+\Delta}^\Delta - X_t^\Delta | X_t^\Delta = x] = f(x)\Delta + o(\Delta) \tag{8}$$

and the following bound on the jumps:

$$X_{t+\Delta}^\Delta - X_t^\Delta = O(\Delta). \tag{9}$$

The following theorem establishes the convergence of (X_t^Δ) to (x_t) when $\Delta \to 0$.

Theorem 1
We have

$$\lim_{\Delta\to 0} X_T^\Delta = x_T, \text{ with probability } 1.$$

Appendix A gives a proof of this result. Note that a weaker convergence result (i.e. convergence in probability) may be obtained from general results in approximation of diffusion processes by Markov chains (Kloeden & Platen, 1995). Here, almost sure convergence is obtained using the *concentration of measure phenomenon* (Talagrand, 1996; Ledoux, 2001), detailed in Appendix A.

Remark 3
If we assume a slightly better consistency error of $O(\Delta^2)$ instead of $o(\Delta)$ in (8), then we may prove (straightforwardly from the Appendix) that $\mathbb{E}[X_T^\Delta] = x_T + O(\Delta)$ and $\mathbb{E}[||X_T^\Delta - x_T||^2] = O(\Delta)$.

2.2 Discretization of the state

Let us go back to our initial control problem (1). We define a **stochastic policy** π_α as a random choice of an action (or control) according to some probabilities: $\pi_\alpha(u|t,x)$ denotes the probability of choosing action $u \in U$ at time t in state x. We write:

$$u_t \sim \pi_\alpha(\cdot|t, x_t)$$

a choice of an action u_t at time t and state x_t, according to such a stochastic policy.

Now, we define the stochastic discrete process $(X_{t_n}^\Delta)_{0\le n\le N}$ (using the same notations for the time-steps (t_n) as in the previous subsection) according to:

- Initialize $X_0^\Delta = x$.
- At time $t \in \{(t_n)_{0\le n<N}\}$, we choose an action $u_t \sim \pi_\alpha(\cdot|t, X_t^\Delta)$. Then, $X_{t+\Delta}^\Delta$ is the state of the system at time $t+\Delta$ resulting from keeping the action u_t constant for a period of time Δ. We write:

$$\left\{ \begin{array}{rcl} u_t & \sim & \pi_\alpha(\cdot|t, X_t^\Delta) \\ X_{t+\Delta}^\Delta & = & y(t, X_t^\Delta; u_t, \Delta) \end{array} \right. \tag{10}$$

  where $y(t,x;u,s)$ represents the state resulting from the state dynamics (1) with initial condition $x_t = x$ and using a constant control u for a period of time s.

When Δ is small, this process is close to a deterministic process $(x_t)_{0\le t\le T}$ defined by the dynamics (4) with

$$f_\alpha(x) := \sum_{u\in U} \pi_\alpha(u|t,x) f(x,u).$$

and initial condition $x_0 = x$. Indeed, the discrete stochastic process (X_t^Δ) converges to (x_t) when $\Delta \to 0$ as an immediate consequence of Theorem 1. To see that, we use Taylor formula,

$$X_{t+\Delta}^\Delta = X_t^\Delta + f(X_t^\Delta, u_t)\Delta + O(\Delta^2),$$

to derive the property on the average jumps:

$$\mathbb{E}[X_{t+\Delta}^\Delta - X_t^\Delta | X_t^\Delta = x] = \sum_{u\in U} \pi_\alpha(u|t,x) f(x,u)\Delta + O(\Delta^2) = f_\alpha(x)\Delta + O(\Delta^2),$$

and the consistency conditions (8) holds, as well as the bound on the jumps (9).

2.3 Discretization of the state gradient

Now, we discretize the state gradient $z_t = \nabla_\alpha x_t$. We build the discrete stochastic process $(Z^\Delta_{t_n})_{0\leq n\leq N}$ according to:

- Initialize $Z^\Delta_0 = 0$.
- At time $t \in \{(t_n)_{0\leq n<N}\}$, let (u_t) and (X^Δ_t) be defined according to (10). Then

$$\begin{aligned} Z^\Delta_{t+\Delta} &= Z^\Delta_t + f(X^\Delta_t, u_t)\big[l_\alpha(t, X^\Delta_t, u_t)' + l_x(t, X^\Delta_t, u_t)' Z^\Delta_t\big]\Delta \\ &\quad + \nabla_x f(X^\Delta_t, u_t) Z^\Delta_t \Delta, \end{aligned} \tag{11}$$

  where

$$l_\alpha(t,x,u) := \frac{\nabla_\alpha \pi_\alpha(u|t,x)}{\pi_\alpha(u|t,x)} \text{ and } l_x(t,x,u) := \frac{\nabla_x \pi_\alpha(u|t,x)}{\pi_\alpha(u|t,x)}$$

  are the likelihood ratios of π_α w.r.t. α and x (defined as vectors of size m and d respectively), and $'$ denotes the transpose operator.

Here again, as a consequence of Theorem 1, the process (Z^Δ_t) converges almost surely to (z_t) when $\Delta \to 0$. Indeed, from the property

$$\begin{aligned} \mathbb{E}[Z^\Delta_{t+\Delta} - Z^\Delta_t | X^\Delta_t = x, Z^\Delta_t = z] &= \sum_{u\in U} \pi_\alpha(u|t,x)\big(f(x,u)[l_\alpha(t,x,u)' + l_x(t,x,u)'z] \\ &\quad + \nabla_x f(x,u) z\big)\Delta \\ &= \big[\nabla_\alpha f_\alpha(x) + \nabla_x f_\alpha(x) z\big]\Delta, \end{aligned}$$

we deduce that the coupled process (X^Δ_t, Z^Δ_t) is consistent with (x_t, z_t) in the sense of (8):

$$\mathbb{E}\left[\begin{pmatrix} X^\Delta_{t+\Delta} \\ Z^\Delta_{t+\Delta}\end{pmatrix} - \begin{pmatrix} X^\Delta_t \\ Z^\Delta_t \end{pmatrix} \,\middle|\, \begin{pmatrix} X^\Delta_t \\ Z^\Delta_t \end{pmatrix} = \begin{pmatrix} x \\ z \end{pmatrix}\right] = \begin{pmatrix} f_\alpha(x) \\ \nabla_\alpha f_\alpha(x) + \nabla_x f_\alpha(x) z \end{pmatrix} \Delta + o(\Delta) \tag{12}$$

and $X^\Delta_{t+\Delta} - X^\Delta_t = O(\Delta)$ and $Z^\Delta_{t+\Delta} - Z^\Delta_t = O(\Delta)$.

3 Model-free reinforcement learning algorithm

We show how to use the approximation results of the previous section to design a model-free reinforcement learning algorithm for estimating the policy gradient $\nabla_\alpha V(\alpha)$ using one trajectory. First, we state the convergence of the policy gradient estimate computed from the discrete process, then show how to approximate the unknown coefficient $\nabla_x f$ using least-squares regression from the observed trajectory, and finally describe the reinforcement learning algorithm.

3.1 Convergence of the policy gradient estimate

One may use formula (7) to define a gradient estimate of the performance measure w.r.t. the parameter α based on the discrete process (X_t^Δ, Z_t^Δ):

$$g(\Delta) := \nabla_x r(X_T^\Delta) Z_T^\Delta. \tag{13}$$

Proposition 2
Assume that r is continuously differentiable. Then

$$\lim_{\Delta \to 0} g(\Delta) = \nabla_\alpha V(\alpha) \text{ with probability } 1.$$

Proof. This is a direct consequence of the almost sure convergence of (X_T^Δ, Z_T^Δ) to (x_T, z_T) and the continuity of $\nabla_x r$. $\Box$

Now, let us illustrate how Z_t^Δ may be approximated with quantities available to the agent. The definition (11) of Z_t^Δ requires the term $\nabla_x f(X_t^\Delta, u)$. We now explain how to built a consistent approximation $\widehat{\nabla_x f}(X_t^\Delta, u)$ of this term based on the past of the trajectory $(X_s^\Delta)_{0 \le s \le t}$.

3.2 Least-squares approximation of $\nabla_x f(X_t^\Delta, u)$

For clarity, we omit references to Δ, for example writing X_s instead of X_s^Δ. Let us write $\Delta X_t = X_{t+\Delta} - X_t$ the jumps of the state. Define $S(t) := \{s \le t, u_s = u_t, X_s - X_t = O(\Delta)\}$ the set of past times $s \le t$ when action u_t have been chosen, and such that the distance between the states X_s and X_t is bounded by a constant times Δ. From Taylor formula, we have for all $s \in S(t)$,

$$\Delta X_s = X_{s+\Delta} - X_s = f(X_s, u_t)\Delta + \nabla_x f(X_s, u_t) f(X_s, u_t) \frac{\Delta^2}{2} + O(\Delta^3). \tag{14}$$

Now, since

$$f(X_s, u_t) = f(X_t, u_t) + \nabla_x f(X_t, u_t)(X_s - X_t) + O(\Delta^2),$$

we deduce (using the fact that $\nabla_x f(X_s, u_t) = \nabla_x f(X_t, u_t) + O(\Delta)$) that

$$\begin{aligned}
\Delta X_s &= \Delta X_t + \left[\nabla_x f(X_s, u_t) f(X_s, u_t) - \nabla_x f(X_t, u_t) f(X_t, u_t)\right] \frac{\Delta^2}{2} \\
&\quad + \nabla_x f(X_t, u_t)(X_s - X_t)\Delta + O(\Delta^3) \\
&= \Delta X_t + \nabla_x f(X_t, u_t)[X_s - X_t + \frac{1}{2}(\Delta X_s - \Delta X_t)]\Delta + O(\Delta^3) \qquad (15) \\
&= b + A(X_s + \frac{1}{2}\Delta X_s)\Delta + O(\Delta^3)
\end{aligned}$$

with $b := \Delta X_t - \nabla_x f(X_t, u_t)(X_t + \frac{1}{2}\Delta X_t)\Delta$ and $A := \nabla_x f(X_t, u_t)$. Based on the observation of several jumps $\{\Delta X_s\}_{s \in S(t)}$, one may derive an approximation of

$\nabla_x f(X_t, u_t)$ by solving the least-squares problem:

$$\min_{A,b} \frac{1}{n_t} \sum_{s \in S(t)} \left\| \Delta X_s - b - A\big(X_s + \frac{1}{2}\Delta X_s\big)\Delta \right\|^2, \tag{16}$$

where n_t is the cardinality of $S(t)$. Write $X_s^+ := X_s + \frac{1}{2}\Delta X_s = \frac{1}{2}(X_s + X_{s+\Delta})$ and use the simplified notations: $\overline{X}$, $\overline{X\,X'}$, $\overline{\Delta X}$, and $\overline{\Delta X\,X'}$, to denote the average values, when $s \in S(t)$, of X_s^+, $X_s^+(X_s^+)'$, ΔX_s, and $\Delta X_s(X_s^+)'$, respectively. For example,

$$\overline{X} := \frac{1}{n_t} \sum_{s \in S(t)} X_s^+.$$

The optimality condition of (16) holds when the matrix $Q_t := \overline{X\,X'} - \overline{X}\,\overline{X}'$ is invertible, and in that case, the least squares solution provides the approximation $\widehat{\nabla_x f}(X_t, u_t)$ of $\nabla_x f(X_t, u_t)$:

$$\widehat{\nabla_x f}(X_t, u_t) = \frac{1}{\Delta}\big(\overline{\Delta X\,X'} - \overline{\Delta X}\,\overline{X}'\big)\big(\overline{X\,X'} - \overline{X}\,\overline{X}'\big)^{-1}. \tag{17}$$

This optimality condition does not hold when the set of points $(X_s^+)_{s \in S(t)}$ lies in a vector space of dimension $< d$ (then Q_t is degenerate). In order to circumvent this problem, we assume that the eigenvalues of the matrix Q_t are bounded away from 0, in the sense given in the following proposition (whose proof in given in Appendix B).

Proposition 3
The matrix $Q_t = \overline{X\,X'} - \overline{X}\,\overline{X}'$ is symmetric non-negative. Let $\nu(\Delta) \geq 0$ be the smallest eigenvalue of Q_t, for all $0 \leq t \leq T$. Then, if $\nu(\Delta) > 0$ and $\nu(\Delta)$ satisfies

$$\frac{1}{\nu(\Delta)} = o(\Delta^4), \tag{18}$$

then, for all $0 \leq t \leq T$, the least squares estimate $\widehat{\nabla_x f}(X_t, u_t)$ defined by (17) is consistent with the gradient $\nabla_x f(X_t, u_t)$, that is:

$$\lim_{\Delta \to 0} \widehat{\nabla_x f}(X_t, u_t) = \nabla_x f(X_t, u_t).$$

The condition (18) is not easy to check, since it depends on the state dynamics and the policy. However, it seems (at least in our experiments) that its holds in general, specially since we used stochastic policies. Future work should focus on its well-foundedness.

3.3 The reinforcement learning algorithm

Here, we derive an policy gradient estimate from (13) in which all information required to build the state gradient Z_t is the past trajectory $(X_s)_{0 \leq s \leq t}$.

Choose a time step Δ. For a given stochastic policy π_α, the algorithm proceeds as follows:

- At time $t = 0$, initialise $X_0 = x$ and $Z_0 = 0$.

- At time $t \in \{(t_n)_{0 \leq n < N}\}$, choose a random action u_t according to the stochastic policy π_α and keep this action for a period of time Δ, which moves the system from X_t to $X_{t+\Delta}$ (summarized by the dynamics (10)). Update Z_t according to

$$\begin{aligned} Z_{t+\Delta} &= Z_t + \Delta X_t \big[l_\alpha(t, X_t, u_t)' + l_x(t, X_t, u_t)' Z_t \big] \Delta \\ &\quad + \widehat{\nabla_x f}(t, X_t, u_t) Z_t \Delta \end{aligned} \tag{19}$$

  where $\widehat{\nabla_x f}(t, X_t, u_t)$ is processed from (17).

- At time T, return the policy gradient estimate $\nabla_x r(X_T) Z_T$.

This algorithm returns a consistent approximation of the policy gradient $\nabla_\alpha V(\alpha)$, as stated in the next Proposition.

Proposition 4
Assume that the property (18) of Proposition 3 holds, and that r is continuously differentiable. Then the estimate $\nabla_x r(X_T) Z_T$ returned by the algorithm converges to $\nabla_\alpha V(\alpha)$ with probability 1, when $\Delta \to 0$.

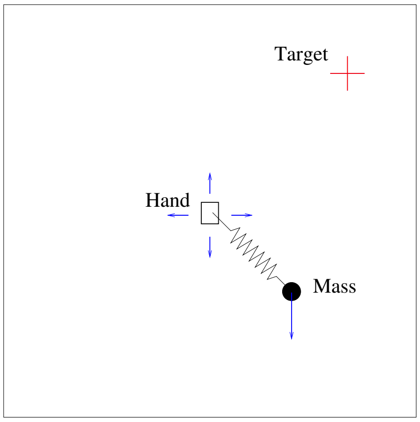
Proof. From Proposition 3, $\widehat{\nabla_x f}$ is a consistent approximation of $\nabla_x f$, thus the process (Z_t) built from (19) also satisfies the consistency condition (12), and the proof follows like in Proposition 2. $\square$

Notice that a simple on-line way for approximating $\nabla_x f$ is to use an exponentially decreasing trace with coefficient $\lambda \in (0, 1)$. For that purpose, we may define a table with values $\overline{Y}(u)$ (where Y means X, XX', ΔX, and $\Delta X\, X'$) for all $u \in U$. The values are initialized (at the first time t each action u is chosen) by Y_t, where Y_t means X_t^+, $X_t^+(X_t^+)'$, ΔX_t, and $\Delta X_t (X_t^+)'$, respectively. Then, the values are updated at time t, according to

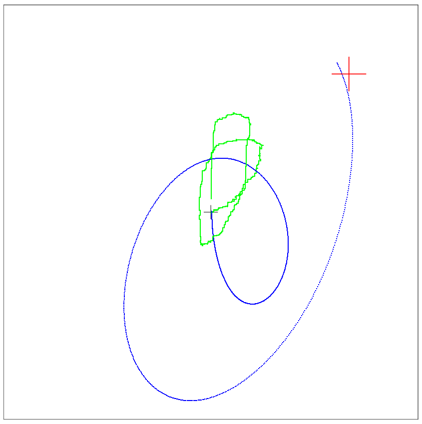
$$\begin{array}{lll} \overline{Y}(u) \leftarrow \lambda \overline{Y}(u) + (1-\lambda) Y_t & \text{for} & u = u_t, \\ \overline{Y}(u) \text{ stays unchanged} & \text{for} & u \neq u_t. \end{array}$$

Thus, the quantities $\overline{X}$, $\overline{X\, X'}$, $\overline{\Delta X}$, and $\overline{\Delta X\, X'}$ are easily updated and the term $\widehat{\nabla_x f}$ may be advantageously computed from (17) by an iterative matrix inversion, such as with Sherman-Morrison formula (see for example (Golub & Loan, 1996)).

Remark 4
Notice that for the initial discrete times t, the matrix $\overline{X\, X'} - \overline{X}\, \overline{X}'$ may not be invertible, simply because there are not enough points $(X_s)_{s<t}$ to define a subspace of dimension d. Then, we may simply set $\widehat{\nabla_x f}$ to 0, which has no impact on the general convergence result.



(a) The physical system



(b) A trajectory (the mass and the hand) starting from the origin

Figure 1: (a) the physical system. (b) A trajectory obtained after 1000 gradient steps.

4 Numerical experiment

We illustrate the algorithm described in the previous section on a 6 dimensional system $(x_0, y_0, x, y, v_x, v_y)$ that represents a hand ((x_0, y_0) position) holding a spring to which is attached a mass (defined by its position (x, y) and velocity (v_x, v_y)) subject to gravitation. The control is the movement of the hand, in any 4 possible directions (up, down, left, right). The goal is to reach a target (x_G, y_G) with the mass at a specific time T (see Figure 1a), while keeping the hand close to the origin. For that purpose, the terminal reward function is defined by

$$r = -x_0^2 - y_0^2 - (x - x_G)^2 - (y - y_G)^2.$$

The dynamics of the system is:

$$\begin{array}{lll} \dot{x_0} = u_x, & \dot{x} = v_x, & \dot{v_x} = -\frac{k}{m}(x - x_0), \\ \dot{y_0} = u_y, & \dot{y} = v_y, & \dot{v_y} = -\frac{k}{m}(y - y_0) - g, \end{array}$$

with k being the spring constant, m the mass, g the gravitational constant, and $(u_x, u_y) = u \in U := \{(1,0), (0,1), (-1,0), (0,-1)\}$ the control. We consider a stochastic policy of the form

$$\pi_\alpha(u|t, x) = \frac{\exp Q_\alpha(t, x, u)}{\sum_{u' \in U} \exp Q_\alpha(t, x, u')}$$

with a linear parameterization of the Q_α values: $Q_\alpha(t, x, u) = \alpha_0^u + \alpha_1^u t + \alpha_2^u x_0 + \alpha_3^u y_0 + \alpha_4^u x + \alpha_5^u y + \alpha_6^u v_x + \alpha_7^u v_y$, for each 4 possible actions $u \in U$. Thus the parameter $\alpha \in I\!R^{32}$. We initialized α with uniform random values in the range $[-0.01, 0.01]$. In

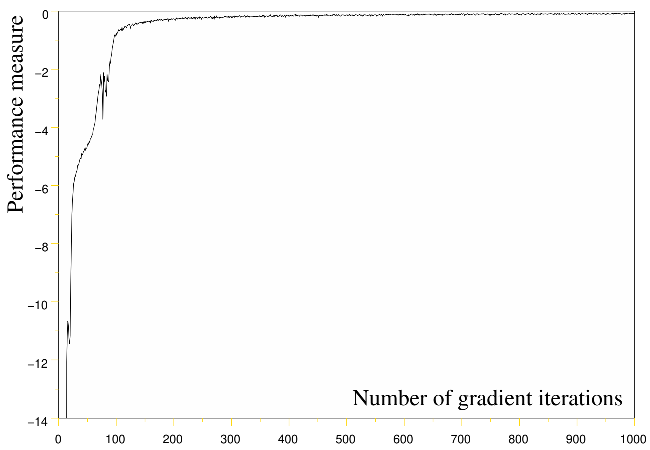


Figure 2: Performance of successive parameterized controllers (performance for the first iterations are below -14).

our experiments we chose $k = 1$, $m = 1$, $g = 1$, $x_G = y_G = 2$, $\lambda = 0.9$, $\Delta = 0.01$, $T = 10$.

At each iteration, we run one (stochastic) trajectory $(X_t)_{0 \leq t \leq T}$ using the stochastic policy, and calculate the policy gradient estimate according to the algorithm described in Section 3.3. We then perform a gradient ascent step (5) (with a fixed $\eta = 0.01$). Figure 2 shows the performance of the parameterized controller as a function of the number of gradient iterations.

For that problem, we chose initial states uniformly sampled from the domain $[-0.1, 0.1]^6$. We found that the randomness introduced in the choice of the initial state helped in not getting stuck in local minima. Here, convergence of the gradient method to an optimal controller (for which $r = 0$) occurs. We illustrate in Figure 1b the trajectory (where only the hand and the mass positions are showed) obtained after 1000 gradient steps, starting from the initial state $(x_0, y_0, x, y, v_x, v_y)_{t=0} = 0$.

5 Conclusion

We described a method for approximating the gradient of the performance measure of a continuous-time deterministic problem, with respect to the control parameters. This was obtained by discretizing the continuous dynamics by a consistent stochastic process built from using a stochastic policy. We showed how a consistent estimation of the gradient may be computed only from the observable.

In future work, it would be useful to extend this idea to stochastic dynamics, and to non-smooth reward, or when its gradient is unknown (maybe using integration-by-part formula for the gradient estimate, such as in the *likelihood ratio method* of (Yang & Kushner, 1991) or the *martingale approach* of (Gobet & Munos, 2005)).

A Appendix: proof of Theorem 1

For convenience, we write x_n for x_{t_n}, X_n for $X^{\Delta}_{t_n}$, and u_n for u_{t_n}, $0 \leq n \leq N$. Let us define the average approximation errors $m^{\Delta}_n = \mathbb{E}[||X_n - x_n||]$ and the squared errors $v^{\Delta}_n = \mathbb{E}[||X_n - x_n||^2]$.

A.1 Convergence of the squared error $\mathbb{E}[||X^{\Delta}_T - x_T||^2]$:

We use the decomposition:

$$\begin{aligned} v^{\Delta}_{n+1} \quad = \quad & \mathbb{E}[||X_{n+1} - X_n||^2] + \mathbb{E}[||X_n - x_n||^2] + \mathbb{E}[||x_n - x_{n+1}||^2] \\ & +2\mathbb{E}[(X_n - x_n)'(X_{n+1} - X_n + x_n - x_{n+1})] \\ & +2\mathbb{E}[(X_{n+1} - X_n)'(x_n - x_{n+1})]. \end{aligned} \tag{20}$$

From the bounded jumps property (9), $\mathbb{E}[||X_{n+1} - X_n||^2] = O(\Delta^2)$. From Taylor formula,

$$x_{n+1} - x_n = f(x_n)\Delta + O(\Delta^2), \tag{21}$$

thus $\mathbb{E}[||x_n - x_{n+1}||^2] = O(\Delta^2)$ (since f is Lipschitz, and x_t and $f(x_t)$ are uniformly bounded on $[0, T]$) and from Cauchy-Schwarz inequality, $|\mathbb{E}[(X_{n+1} - X_n)'(x_n - x_{n+1})]| = O(\Delta^2)$. From (8) and (21),

$$\mathbb{E}[X_{n+1} - X_n + x_n - x_{n+1}|X_n] = [f(X_n) - f(x_n)]\Delta + o(\Delta). \tag{22}$$

Now, from (9) we deduced that $||X_n - x|| = O(1)$ thus X_n is bounded (for all n and N), as well as x_n. Let B a constant such that $||X_n|| \leq B$ and $||x_n|| \leq B$ for all $n \leq N$, $N \geq 0$. Since f is $\mathcal{C}^2$, from Taylor formula, there exists a constant k, such that, for all $n \leq N$,

$$||f(X_n) - f(x_n) - \nabla_x f(x_n)(X_n - x_n)|| \leq k||X_n - x_n||^2. \tag{23}$$

We deduce that

$$\begin{aligned} & |\mathbb{E}[(X_n - x_n)'(X_{n+1} - X_n + x_n - x_{n+1})]| \\ & = \big|\mathbb{E}\big[(X_n - x_n)'(f(X_n) - f(x_n))\big]\big|\,\Delta + o(\Delta) \\ & \leq \big|\mathbb{E}\big[(X_n - x_n)'\nabla_x f(x_n)(X_n - x_n)\big]\big|\,\Delta + 2kBv_n\Delta + o(\Delta) \\ & \leq Mv_n\Delta + o(\Delta) \end{aligned}$$

with $M = \sup_{||x|| \leq B} ||\nabla_x f(x)|| + 2kB$. Thus, (20) leads to the recurrent bound

$$v^{\Delta}_{n+1} \leq (1 + M\Delta)v^{\Delta}_n + o(\Delta).$$

This actually means that there exists a function $e(\Delta) \to 0$ when $\Delta \to 0$, such that $v_{n+1}^{\Delta} \leq (1 + M\Delta)v_n^{\Delta} + e(\Delta)\Delta$. Thus,

$$v_N^{\Delta} \leq \frac{(1+M\Delta)^N - 1}{(1+M\Delta) - 1} e(\Delta)\Delta \leq (e^{NM\Delta} - 1)\frac{1}{M} e(\Delta)$$

thus $v_N^{\Delta} = o(1)$, that is $\mathbb{E}[||X_T^{\Delta} - x_T||^2] \stackrel{\Delta \to 0}{\longrightarrow} 0$.

A.2 Convergence of the mean $\mathbb{E}[||X_T^{\Delta} - x_T||]$:

From (22), we have

$$\mathbb{E}[X_{n+1} - x_{n+1} | X_n] = X_n - x_n + [f(X_n) - f(x_n)]\Delta + o(\Delta).$$

Thus from (23),

$$\begin{aligned} m_{n+1}^{\Delta} = \mathbb{E}[||X_{n+1} - x_{n+1}||] &\leq (1 + ||\nabla_x f(x_n)||\Delta)\mathbb{E}[||X_n - x_n||] + kv_n^{\Delta}\Delta + o(\Delta) \\ &\leq (1 + M'\Delta)m_n^{\Delta} + o(\Delta), \end{aligned}$$

since $v_n^{\Delta} = o(1)$ (with $M' = \sup_{||x|| \leq B} ||\nabla_x f(x)||$). Using the same deduction as above, we obtain that $m_N^{\Delta} = o(1)$, that is $\mathbb{E}[||X_T^{\Delta} - x_T||] \stackrel{\Delta \to 0}{\longrightarrow} 0$.

A.3 Almost sure convergence

Here, we use the *concentration-of-measure phenomenon* (Talagrand, 1996; Ledoux, 2001), which states that under mild conditions, a function (say Lipschitz or with bounded differences) of many independent random variables concentrates around its mean, in the sense that the tail probability decreases exponentially fast.

One may write the state X_N as a function h of some independent random variables $(U_n)_{0 \leq n < N}$:

$$X_N = h(U_0, \dots, U_{N-1}) := \sum_{n=0}^{N-1} X_{n+1} - X_n. \tag{24}$$

Observe that $h - \mathbb{E}[h] = \sum_{n=0}^{N-1} d_n$ with

$$\begin{aligned} d_n &= \mathbb{E}[h(U_0, \dots, U_{N-1}) | U_0, \dots, U_n] - \mathbb{E}[h(U_0, \dots, U_{N-1}) | U_0, \dots, U_{n-1}] \\ &= X_{n+1} - X_n - \mathbb{E}[X_{n+1} - X_n] \end{aligned}$$

being a martingale difference sequence (that is $\mathbb{E}[d_n | U_0, \dots, U_{n-1}] = 0$). Now, from (Ledoux, 2001, lemma 4.1), one has:

$$\mathbb{P}(|g - \mathbb{E}[g]| \geq \varepsilon) \leq 2e^{-\varepsilon^2/(2D^2)} \tag{25}$$

for any $D^2 \geq \sum_{n=0}^{N-1} ||d_n||_{\infty}^2$. Thus, from (9), and since X_n is bounded, as well as $f(X_n)$ (for all $n < N$ and all $N > 0$), there exists a constant C that does not depend on N such that $d_n \leq C/N$. Thus we may take $D^2 = C^2/N$.

Now, from the previous paragraph, $|\mathbb{E}[X_N] - x_N| \leq e(N)$, with $e(N) \to 0$ when $N \to \infty$. This means that $|h - \mathbb{E}[h]| + e(N) \geq |X_N - x_N|$, thus

$$\mathbb{P}(|h - \mathbb{E}[h]| \geq \varepsilon + e(N)) \geq \mathbb{P}(|X_N - x_N| \geq \varepsilon),$$

and we deduce from (25) that

$$\mathbb{P}(|X_N - x_N| \geq \varepsilon) \leq 2e^{-N(\varepsilon + e(N))^2/(2C^2)}.$$

Thus, for all $\varepsilon > 0$, the series $\sum_{N \geq 0} \mathbb{P}(|X_N - x_N| \geq \varepsilon)$ converges. Now, from Borel-Cantelli lemma, we deduce that for all $\varepsilon > 0$, there exists N_ε such that for all $N \geq N_\varepsilon$, $|X_N - x_N| < \varepsilon$, which proves the convergence $X_N \to x_N$ as $N \to \infty$, with probability 1.

B Proof of proposition 3

First, note that $Q_t = \overline{X\,X'} - \overline{X}\,\overline{X}'$ is a symmetric, non-negative matrix, since it may be rewritten as

$$\frac{1}{n_t} \sum_{s \in S(t)} (X_s^+ - \overline{X})(X_s^+ - \overline{X})'.$$

In solving the least squares problem (16), we deduce $b = \overline{\Delta X} + A\overline{X}\Delta$, thus

$$\begin{aligned}
&\min_{A,b} \frac{1}{n_t} \sum_{s \in S(t)} ||\Delta X_s - b - A(X_s + \frac{1}{2}\Delta X_s)\Delta||^2 \\
&= \min_{A} \frac{1}{n_t} \sum_{s \in S(t)} ||\Delta X_s - \overline{\Delta X} - A(X_s^+ - \overline{X})\Delta||^2 \\
&\leq \frac{1}{n_t} \sum_{s \in S(t)} ||\Delta X_s - \overline{\Delta X} - \nabla_x f(\overline{X}, u_t)(X_s^+ - \overline{X})\Delta||^2. \qquad (26)
\end{aligned}$$

Now, since $X_s = \overline{X} + O(\Delta)$ one may obtain like in (14) and (15) (by replacing X_t by $\overline{X}$) that:

$$\Delta X_s - \overline{\Delta X} - \nabla_x f(\overline{X}, u_t)(X_s^+ - \overline{X})\Delta = O(\Delta^3). \qquad (27)$$

We deduce from (26) and (27) that

$$\frac{1}{n_t} \sum_{s \in S(t)} \left\| \left[\widehat{\nabla_x f}(X_t, u_t) - \nabla_x f(\overline{X}, u_t) \right] (X_s^+ - \overline{X})\Delta \right\|^2 = O(\Delta^6).$$

By developping each component,

$$\sum_{i=1}^{d} \left[\widehat{\nabla_x f}(X_t, u_t) - \nabla_x f(\overline{X}, u_t) \right]_{row\, i} Q_t \left[\widehat{\nabla_x f}(X_t, u_t) - \nabla_x f(\overline{X}, u_t) \right]'_{row\, i} = O(\Delta^4).$$

Now, from the definition of $\nu(\Delta)$, for all vector $u \in I\!R^d$, $u'Q_t u \geq \nu(\Delta)||u||^2$, thus

$$\nu(\Delta) \sum_{i=1}^{d} ||\widehat{\nabla_x f}(X_t, u_t) - \nabla_x f(\overline{X}, u_t)||^2 = O(\Delta^4).$$

Condition (18) yields $\widehat{\nabla_x f}(X_t, u_t) = \nabla_x f(\overline{X}, u_t) + o(1)$, and since $\nabla_x f(X_t, u_t) = \nabla_x f(\overline{X}, u_t) + O(\Delta)$, we deduce

$$\lim_{\Delta \to 0} \widehat{\nabla_x f}(X_t, u_t) = \nabla_x f(X_t, u_t).$$

References

BAXTER J. & BARTLETT P. (2001). Infinite-horizon gradient-based policy search. *Journal of Artificial Intelligence Research*, **15**, 319–350.

BONNANS F., GILBERT J., LEMARÉCHAL C. & SAGASTIZABAL C. (2003). *Numerical Optimization. Theoretical and Practical Aspects*. Springer-Verlag.

GOBET E. & MUNOS R. (2005). Sensitivity analysis using itô-malliavin calculus and martingales. application to stochastic optimal control. *To appear in SIAM journal on Control and Optimization*.

GOLUB G. H. & LOAN C. F. V. (1996). *Matrix Computations, 3rd ed.* Baltimore, MD: Johns Hopkins.

KLOEDEN P. E. & PLATEN E. (1995). *Numerical Solutions of Stochastic Differential Equations*. Springer-Verlag.

KUSHNER H. J. & YIN G. (1997). *Stochastic Approximation Algorithms and Applications*. Springer-Verlag, Berlin and New York.

LEDOUX M. (2001). *The concentration of measure phenomenon*. American Mathematical Society, Providence, RI.

MARBACH P. & TSITSIKLIS J. N. (2003). Approximate gradient methods in policy-space optimization of markov reward processes. *Journal of Discrete Event Dynamical Systems*, **13**, 111–148.

POLYAK B. (1987). *Introduction to Optimization*. Optimization Software Inc., New York.

SUTTON R., MCALLESTER D., SINGH S. & MANSOUR Y. (2000). Policy gradient methods for reinforcement learning with function approximation. *Neural Information Processing Systems. MIT Press*, p. 1057–1063.

SUTTON R. S. & BARTO A. G. (1998). Reinforcement learning: An introduction. *Bradford Book*.

TALAGRAND M. (1996). A new look at independence. *Annals of Probability*, **24**, 1–34.

WILLIAMS R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, **8**, 229–256.

YANG J. & KUSHNER H. (1991). A Monte Carlo method for sensitivity analysis and parametric optimization of nonlinear stochastic systems. *SIAM J. Control Optim.*, **29**(5), 1216–1249.