Bias Correction in Reinforcement Learning via the Deterministic Policy Gradient Method for MPC-Based Policies

Sébastien Gros, Mario Zanon

Abstract— **In this paper, we discuss the implementation of the Deterministic Policy Gradient using the Actor-Critic technique based on linear compatible advantage function approximations in the context of constrained policies. We focus on MPC-based policies, though the discussion is general. We show that in that context, the classic linear compatible advantage function approximation fails to deliver a correct policy gradient due to the exploration becoming distorted by the constraints, and we propose a generalized linear compatible advantage function approximation that corrects the problem. We show that this correction requires an estimation of the mean and covariance of the constrained exploration. The validity of that generalization is formally established and demonstrated on a simple example.**

I. Introduction

Reinforcement Learning (RL) offers useful tools for tackling Markov Decision Processes (MDP) without relying on a detailed model of the probability distributions underlying the state transitions [16], [3]. RL has drawn an increasingly large attention thanks to its accomplishments, such as, e.g., making it possible for robots to learn to walk or fly without supervision [18], [1]. In the recent RL literature, unstructured function approximation techniques, e.g., Deep Neural Networks (DNN) are often used to carry the policy approximation. Structured function approximations, often based on formal control methods, have recently gained the attention of the research community [2], [5], [7], [8], [9], [10], [11], [13], [14], [17], since they allow one to provide formal guarantees on the closed-loop behavior of the system, and to use prior knowledge about the system more directly.

Model Predictive Control (MPC) can naturally carry the policy: provided that a (possibly inaccurate) model of the real system is available, MPC delivers a suboptimal but typically reasonable candidate policy for the real system. Moreover, MPC can explicitly treat hard constraints, which are typically used to restrict the predicted state and inputs to evolve in a feasible region of the state-input space. Furthermore, because it seeks to minimize a given cost and respect constraints, the behavior of the MPC policy is often easier to interpret than the one of a generic policy approximation. The use of MPC within RL is formally justified in [5]: under some stability assumptions, MPC schemes can generate jointly the optimal value function, action value function and policy underlying an MDP even though the MPC model does not capture the real system perfectly. This can be achieved via modifications of the MPC cost and constraints. This approach has been further used in [6], [7], [12], [19], [20], [21] to improve the MPC closed-loop performance based on data.

Among the well-established RL methods, policy gradient methods based on Actor-Critic (AC) techniques offer an attractive approach because, unlike Q-learning, they are based on genuine conditions of optimality of the closed-loop policy. Moreover, AC techniques tend to deliver less noisy policy gradients than direct policy search. In this paper, we focus on the Deterministic Policy Gradient (DPG) approach and the use of compatible advantage function approximations (CAFA) [15] in the context of constrained policies. Exploration is required in order to implement the DPG approach, such that when the policy is subject to constraints, the exploration can become restricted. We show that the CAFA of [15] can then deliver an incorrect policy gradient, and propose a generalized CAFA to tackle this problem. The generalized CAFA requires an estimation of the mean and covariance of the exploration, which is unfortunately not always straightforward to obtain. The issue detailed in this paper is illustrated on a very small, constructed example of an MPC-based policy where the classical DPG approach delivers an incorrect gradient, while the proposed CAFA delivers a correct one.

The paper is organized as follows. Section II provides some background material. Section III provides the generalized CAFA and establishes formally that it delivers the correct policy gradient. Section IV further discusses the application of the generalized CAFA in the context of MPC-based policies. Section V proposes a very simple example demonstrating the problem, and showing that the proposed CAFA corrects it. Section VI delivers conclusions.

II. Background

We consider real systems described as stochastic processes on continuous state-action spaces, and denote deterministic policies parametrized in $\boldsymbol{\theta}$ and delivering action $\mathbf{a}$ as a function of the state $\mathbf{s}$ as:

$$\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}) : \mathbb{R}^n \rightarrow \mathbb{R}^m. \tag{1}$$

For a given stage cost function $L(\mathbf{s},\mathbf{a}) \in \mathbb{R}$ and a discount factor $\gamma \in [0,1]$, the performance of a policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ is assessed via the total expected cost:

$$J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\varrho_{\mathbf{s}}}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k,\mathbf{a}_k) \,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}_k)\right]. \tag{2}$$

where $\mathbb{E}_{\varrho_{\mathbf{s}}}[\cdot]$ is the expected value associated to the measure $\varrho_{\mathbf{s}}$ underlying the distribution of the Markov Chain resulting

Sébastien Gros is with the Department of Cybernetic, NTNU, Norway. Mario Zanon is with the IMT School for Advanced Studies Lucca, Italy.

from the real-system operating in closed loop with policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}$. The optimal policy parameters are then given by:

$$\boldsymbol{\theta}_\star = \arg\min_{\boldsymbol{\theta}} \; J(\boldsymbol{\pi}_{\boldsymbol{\theta}}). \tag{3}$$

The policy gradient

$$\nabla_{\boldsymbol{\theta}} \, J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\mathbf{s}}\left[\nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}} \, \nabla_{\mathbf{a}} A_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\right] \tag{4}$$

is then instrumental in finding the optimal policy parameters $\boldsymbol{\theta}_\star$, where $A_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ is the advantage function associated to $\boldsymbol{\pi}_{\boldsymbol{\theta}}$. Here the expected value $\mathbb{E}_{\mathbf{s}}\left[\cdot\right]$ can be understood as roll-outs of the Markov Chain where the contributions $\nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}} \, \nabla_{\mathbf{a}} A_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ are discounted over time, or in the sense of the Markov Chain steady-state distribution where they are averaged in time.

In this paper, we consider policies $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ delivered by an MPC scheme parametrized by $\boldsymbol{\theta}$. The use of MPC as a policy approximation in RL has been investigated and justified in [5], [6], [7], [12], [19], [20], [21], and offers some advantages over the more generic function approximations often used in RL. Indeed, MPC-based policies allow one to 1. ensure that the system trajectories resulting from operating the real system in closed-loop with the MPC policy $\boldsymbol{\pi}_{\theta}$ are feasible with respect to some hard constraints; 2. directly exploit existing knowledge of the system (i.e., models); 3. easily use predictive information; and 4. exploit the vast set of theoretical tools available for MPC to enforce key requirements in the resulting closed-loop behavior, such as stability and recursive feasibility.

For a given state $\mathbf{s}$ of the real system, the MPC policy is

$$\boldsymbol{\pi}_{\boldsymbol{\theta}}\left(\mathbf{s}\right) = \mathbf{u}_0^\star\left(\mathbf{s}, \boldsymbol{\theta}\right) \, \in \mathbb{R}^m, \tag{5}$$

where $\mathbf{u}_0^\star$ is the first element of the input sequence $\mathbf{u}^\star = \left\{\mathbf{u}_0^\star, \dots, \mathbf{u}_{N-1}^\star\right\}$ resulting from solving the MPC scheme:

$$\begin{aligned} \mathbf{u}^\star\left(\mathbf{s}, \boldsymbol{\theta}\right) = \arg\min_{\mathbf{u}} \;\; & T_{\boldsymbol{\theta}}(\mathbf{x}_N) + \sum_{k=0}^{N-1} \ell_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) && \text{(6a)} \\ \text{s.t.} \;\; & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right), \;\; \mathbf{x}_0 = \mathbf{s}, && \text{(6b)} \\ & \mathbf{h}_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \leq 0, \quad \mathbf{h}^{\mathrm{f}}_{\boldsymbol{\theta}}\left(\mathbf{x}_N\right) \leq 0, && \text{(6c)} \end{aligned}$$

where $\mathbf{u}_{0,\dots,N-1}$ is the MPC input profile and $\mathbf{x}_{0,\dots,N}$ the corresponding predicted state trajectory; $T_{\boldsymbol{\theta}}$, $\ell_{\boldsymbol{\theta}}$ are the terminal and stage costs, respectively, and $\mathbf{f}_{\boldsymbol{\theta}}$, the system model. Function $\mathbf{h}_{\boldsymbol{\theta}}$ is used to impose the state and input constraints limiting the system behavior. Function $\mathbf{h}^{\mathrm{f}}_{\boldsymbol{\theta}}$ is a constraint on the terminal state often used to enforce the stability and recursive feasibility of the MPC scheme [4]. Parametrizing the MPC constraints and adapting them via RL is formally motivated in [5], and finds a practical use in, e.g., [19].

It will be useful in the following to cast (6) as a more generic parametric Nonlinear Program (NLP):

$$\begin{aligned} \boldsymbol{\pi}_{\theta}\left(\mathbf{s}\right) = \arg\min_{\mathbf{u}_0} \quad & \Phi(\mathbf{s}, \mathbf{u}_0, \boldsymbol{\theta}) && \text{(7a)} \\ \text{s.t.} \quad & \mathbf{H}\left(\mathbf{s}, \mathbf{u}_0, \boldsymbol{\theta}\right) \leq 0, && \text{(7b)} \end{aligned}$$

resulting from a nonlinear condensing of (6), such that the only remaining decision variable is the first control input. NLP (7) will be used for performing theoretical analysis in a simple and compact way, but it is impractical to be solved numerically, such that one should rather solve (6).

In order to estimate the gradient of the advantage function $\nabla_{\mathbf{a}} A_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ required in (4), exploration must be introduced, i.e., the input $\mathbf{a}$ applied to the real system must differ from the actual policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}\left(\mathbf{s}\right)$. It is common in RL to produce the exploration by adding a random disturbance $\mathbf{d}$ to the policy

$$\mathbf{a}^{\mathrm{d}} := \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}) + \mathbf{d}, \tag{8}$$

However, since the policy is subject to the hard constraints (7b), it can be desirable that the exploration does not produce constraints violation. As a result, the exploration ought to be restricted such that it satisfies (7b). In order to address this issue, we consider a projection of $\mathbf{a}^{\mathrm{d}}$ on the feasible set of the MPC scheme. More specifically, let us consider the NLP

$$\begin{aligned} \mathbf{a}^{\mathrm{e}} = \arg\min_{\mathbf{u}_0} \quad & \frac{1}{2}\left\|\mathbf{u}_0 - \mathbf{a}^{\mathrm{d}}\right\|^2 && \text{(9a)} \\ \text{s.t.} \quad & \mathbf{H}\left(\mathbf{s}, \mathbf{u}_0, \boldsymbol{\theta}\right) \leq 0, && \text{(9b)} \end{aligned}$$

which generates, by construction, inputs $\mathbf{a}$ that are feasible for (7) and, therefore, also for the original MPC scheme (6). Note that (9) is a projection of the perturbed input $\mathbf{a}^{\mathrm{d}}$ onto the set of feasible initial inputs for the MPC scheme (6). In the following, we will label as exploration the difference

$$\mathbf{e} = \mathbf{a}^{\mathrm{e}} - \boldsymbol{\pi}_{\boldsymbol{\theta}}\left(\mathbf{s}\right), \tag{10}$$

and denote its covariance and mean as

$$\Sigma_{\mathbf{e}}\left(\mathbf{s}, \boldsymbol{\theta}\right) = \mathrm{Cov}\left(\mathbf{e}\,|\,\mathbf{s}\right), \quad \boldsymbol{\mu}_{\mathbf{e}}\left(\mathbf{s}, \boldsymbol{\theta}\right) = \mathbb{E}\left[\mathbf{e}\,|\,\mathbf{s}\right]. \tag{11}$$

In [7], this projection approach was analyzed for DPG, but it assumed for brevity that the true advantage function was available. In practice, the advantage function must be estimated through exploration. We recall next the DPG method based on the classical linear CAFA.

A. Compatible Advantage Function Approximation

A difficulty arises when forming estimations of $\nabla_{\mathbf{a}} A_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ required in the policy gradient (4). It is well known that building a parametric (in $\mathbf{w}$) estimation $\widehat{\nabla_{\mathbf{a}} A^{\mathbf{w}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \approx \nabla_{\mathbf{a}} A_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ directly is very difficult, hence one typically considers estimating the advantage function $\hat{A}^{\mathbf{w}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \approx A_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ instead, from which the gradient $\nabla_{\mathbf{a}} \hat{A}^{\mathbf{w}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ is evaluated. The policy gradient is then approximated as:

$$\widehat{\nabla_{\boldsymbol{\theta}} \, J}(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\mathbf{s}}\left[\nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}} \, \nabla_{\mathbf{a}} \hat{A}^{\mathbf{w}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\right]. \tag{12}$$

In order for this policy gradient approximation to be correct, the advantage function must be compatible, i.e., it must satisfy the conditions of the following theorem.

Theorem 1 ([15]): A function approximator $\hat{A}^{\mathbf{w}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ is compatible if

(i) $\nabla_{\mathbf{a}} \hat{A}^{\mathbf{w}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} = \nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}}^\top \mathbf{w}$;

(ii) $\mathbf{w}$ minimizes the following mean-squared error

$$\mathbf{w} = \arg\min_{\bar{\mathbf{w}}} \, \mathbb{E}_{\mathbf{s}}\left[\left\|\nabla_{\mathbf{a}} A_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} - \nabla_{\mathbf{a}} \hat{A}^{\bar{\mathbf{w}}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\right\|^2\right], \tag{13}$$

where the gradients are evaluated at $\mathbf{a} = \boldsymbol{\pi}_{\boldsymbol{\theta}}$.

This implies $\widehat{\nabla_{\boldsymbol{\theta}} J^{\mathbf{w}}}(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$. □

The importance of this theorem is that it relaxes the need to know $\nabla_{\mathbf{a}} A_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ exactly. Furthermore, the estimate of the advantage function gradient needs to be correct only in expected value and the function approximation can be linear. Nevertheless, as also pointed out in [15], minimizing (13) directly is very difficult in practice.

The practical solution consists in estimating the advantage function by solving the least-squares problem

$$\mathbf{w} = \mathrm{a}\min_{\mathbf{w}} \mathbb{E}_{\mathbf{s},\mathbf{e}} \left[\left(Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left(\mathbf{s}, \mathbf{a}^{\mathrm{e}}\right) - \hat{V}^{\mathbf{v}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left(\mathbf{s}\right) - \hat{A}^{\mathbf{w}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left(\mathbf{s}, \mathbf{a}^{\mathrm{e}}\right) \right)^2 \right], \tag{14}$$

where we label $\mathbb{E}_{\mathbf{s},\mathbf{e}}\left[\cdot\right] = \mathbb{E}_{\mathbf{s}}\left[\mathbb{E}_{\mathbf{e}}\left[\cdot|\mathbf{s}\right]\right]$ and where $\mathbf{a} = \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}) + \mathbf{e}$, with $\|\mathbf{e}\| > 0$ in order to introduce exploration. The value function estimation $\hat{V}^{\mathbf{v}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \approx V_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ is a baseline supporting the evaluation of $\mathbf{w}$. We ought to underline that we formulate (14) using the real action-value function $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$, which is not directly available. Nevertheless, there exist RL approaches that solve (14), though some might have convergence issues or be very slow at converging [16]. The CAFA proposed in [15] is

$$\hat{A}^{\mathbf{w}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left(\mathbf{s}, \mathbf{a}\right) = \mathbf{w}^\top \nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}) \left(\mathbf{a}^{\mathrm{e}} - \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s})\right). \tag{15}$$

The main argument used to support solving (14) rather than (13) is that, if the functions are regular enough, then $\hat{A}^{\mathbf{w}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \approx A_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ implies $\widehat{\nabla_{\mathbf{a}} A^{\mathbf{w}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \approx \nabla_{\mathbf{a}} A_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$. In the next section, we show that this argument formally requires that

$$\Sigma_{\mathbf{e}}\left(\mathbf{s}, \boldsymbol{\theta}\right) = \sigma I, \qquad \boldsymbol{\mu}_{\mathbf{e}}\left(\mathbf{s}, \boldsymbol{\theta}\right) = 0, \tag{16}$$

holds for almost every $\mathbf{s}$ for some scalar $\sigma > 0$. We further show that if (16) does not hold then the classic advantage function approximation (15) must be modified by accounting for $\Sigma_{\mathbf{e}}$, $\boldsymbol{\mu}_{\mathbf{e}}$. This modification constitutes a generalization of (15). This result will apply in general, and, in particular, to the projection (9)-(10) considered here, for which (16) does not necessarily hold.

III. Constrained exploration

In this section, we will investigate how an exploration $\mathbf{e}$ that fails (16) should be treated in the context of the DPG method. We will show that corrections are required in the advantage function approximation (15) to obtain a correct policy gradient estimation, and that these corrections require the exploration mean and covariance $\boldsymbol{\mu}_{\mathbf{e}}$, $\Sigma_{\mathbf{e}}$ to be known. To that end, we propose a generalization of (15), given by

$$\hat{A}^{\mathbf{w}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left(\mathbf{s}, \mathbf{a}\right) = \mathbf{w}^\top \nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}} M \left(\mathbf{a}^{\mathrm{e}} - \boldsymbol{\pi}_{\boldsymbol{\theta}} - \boldsymbol{\eta}\right), \tag{17}$$

where matrix $M\left(\mathbf{s}\right)$ and vector $\boldsymbol{\eta}\left(\mathbf{s}\right)$ are given by:

$$M\left(\mathbf{s}\right) = \sigma \Sigma_{\mathbf{e}}\left(\mathbf{s}, \boldsymbol{\theta}\right)^{-1}, \quad \boldsymbol{\eta}\left(\mathbf{s}\right) = \boldsymbol{\mu}_{\mathbf{e}}\left(\mathbf{s}\right). \tag{18}$$

From now on we will use $\sigma > 0$ to label the covariance of $\mathbf{d}$ assumed centered and isotropic, i.e., we assume that $\Sigma_{\mathbf{d}} = \sigma I$, $\boldsymbol{\mu}_{\mathbf{d}} = 0$. One can readily observe that (17) is a generalization of (15), as the case $\Sigma_{\mathbf{e}} = \sigma I$ and $\boldsymbol{\mu}_{\mathbf{e}} = 0$ makes them identical. We will show next that (17) is required to yield a correct policy gradient estimation when the exploration fails (16).

Assumption 1: We assume the following:

a. $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left(\mathbf{s}, \mathbf{a}\right)$ is analytic and smooth almost everywhere w.r.t. $\mathbf{a} = \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s})$ on the set of feasible $\mathbf{s}$, such that it admits a Taylor expansion almost everywhere on the domain of the state space where the policy is defined.
b. The probability density underlying the measure $\varrho_{\mathbf{s}}$ is bounded.
c. The following limits hold:

$$\mathbb{E}_{\mathbf{e}}\left[\frac{1}{\sigma}\left(\mathbf{e} - \boldsymbol{\mu}_{\mathbf{e}}\right)^{\boldsymbol{\alpha}}\right] = 0, \quad \forall \left|\boldsymbol{\alpha}\right| > 2, \tag{19a}$$

$$\lim_{\sigma \to 0} \boldsymbol{\mu}_{\mathbf{e}} = 0. \tag{19b}$$

Note that we used the multi-index notation in (19a), and it will be used again later.

We ought to briefly discuss these assumptions. Assumption 1a. is usually satisfied in practice, as the $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ tends to be piecewise smooth for may problems based on continuous state-input spaces. Assumption 1b. in principle excludes deterministic real systems. Both assumption can arguably be relaxed, albeit that would make the following developments significantly more technical. The limit (19a) requires that the moments of order higher than 2 of the projected exploration vanish faster than the second moment of the disturbance $\mathbf{d}$, while (19b) requires that the exploration mean decays with σ. To our best understanding, these assumptions are straightforward to satisfy if a reasonable distribution for the disturbance $\mathbf{d}$ is selected (bounded density and support, and vanishing moments) and if the feasible set of (9b) is well behaved (connected, non-empty interior). However, further investigations are required to establish these conjectures formally.

The Deterministic Policy Gradient method is meant to be deployed with "small" exploration, because all results are valid in the sense of $\sigma \to 0$. In order to construct a formally correct argument in that limit case, we will consider the fitting of the advantage function (17) in the sense of:

$$\mathbf{w} = \arg\min_{\bar{\mathbf{w}}} \mathbb{E}_{\mathbf{s},\mathbf{e}}\left[\frac{1}{\sigma}\left(Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a}) - \hat{V}^{\mathbf{v}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{s}) - \hat{A}^{\bar{\mathbf{w}}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a})\right)^2\right]. \tag{20}$$

Note that the solution of (20) must satisfy

$$\mathbb{E}_{\mathbf{s},\mathbf{e}}\left[\frac{1}{\sigma} \nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}} M \left(\mathbf{e} - \boldsymbol{\eta}\right) \left(Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} - \hat{V}^{\mathbf{v}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} - \hat{A}^{\mathbf{w}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\right)\right] = 0, \tag{21}$$

where, by (17), $\hat{A}^{\mathbf{w}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} = \left(\mathbf{e} - \boldsymbol{\eta}\right)^\top M \nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}^\top \mathbf{w}$, such that in (21) $\mathbf{w}$ is multiplied by

$$\begin{aligned} &\mathbb{E}_{\mathbf{s},\mathbf{e}}\left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}} M \left(\mathbf{e} - \boldsymbol{\eta}\right) \left(\mathbf{e} - \boldsymbol{\eta}\right)^\top M \nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}^\top\right] \\ &\qquad = \mathbb{E}_{\mathbf{s}}\left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}} \sigma M \nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}^\top\right]. \end{aligned} \tag{22}$$

Consequently, without the factor $\frac{1}{\sigma}$, any $\mathbf{w}$ would solve the problem in the limit $\sigma \to 0$. Therefore, $\frac{1}{\sigma}$ ensures that the least-squares problem (20) remains well posed when $\sigma \to 0$.

We can now turn to establishing our main result.

Theorem 2: Under Assumption 1, the deterministic policy gradient estimation (12) is asymptotically exact, i.e.,

$$\lim_{\sigma\rightarrow 0} \widehat{\nabla_{\boldsymbol{\theta}}\, J(\boldsymbol{\pi}_{\boldsymbol{\theta}})} = \nabla_{\boldsymbol{\theta}}\, J(\boldsymbol{\pi}_{\boldsymbol{\theta}}), \tag{23}$$

if using the approximator (17) which solves (20).

Proof: Since $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ is analytic and at least twice differentiable almost everywhere, its second-order expansion in $\mathbf{a}$ at $\mathbf{e} = 0$ is valid almost everywhere:

$$\begin{aligned} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{s},\mathbf{a}) &= V_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{s}) + \nabla_{\mathbf{a}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{s},\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}))^\top (\mathbf{a} - \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s})) + \xi \\ &= V_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{s}) + \nabla_{\mathbf{a}} A_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{s},\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}))^\top \mathbf{e} + \xi, \end{aligned} \tag{24}$$

where ξ is the second-order remainder of the Taylor expansion of $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ at $\mathbf{e} = 0$, and we used the identity $\nabla_{\mathbf{a}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} = \nabla_{\mathbf{a}} A_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$. We rewrite (17) as

$$\hat{A}^{\mathbf{w}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{s},\mathbf{a}) = \nabla_{\mathbf{a}} \hat{A}^{\mathbf{w}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{s},\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}))^\top \mathbf{e} - \boldsymbol{\eta}^\top M \nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}^\top \mathbf{w}. \tag{25}$$

Using (24)-(25), the optimality condition (21) becomes:

$$\begin{aligned} &\mathbb{E}_{\mathbf{s},\mathbf{e}}\left[\frac{1}{\sigma}\nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}} M(\mathbf{e}-\boldsymbol{\eta})\mathbf{e}^\top \left(\nabla_{\mathbf{a}} A_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} - \nabla_{\mathbf{a}} \hat{A}^{\mathbf{w}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\right)\right] \\ &\quad + \mathbb{E}_{\mathbf{s},\mathbf{e}}\left[\frac{1}{\sigma}\nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}} M(\mathbf{e}-\boldsymbol{\eta})\xi\right] \\ &\qquad + \mathbb{E}_{\mathbf{s},\mathbf{e}}\left[\frac{1}{\sigma}\nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}} M(\mathbf{e}-\boldsymbol{\eta})\boldsymbol{\eta}^\top M \nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}}^\top\right]\mathbf{w} \\ &\qquad\quad + \mathbb{E}_{\mathbf{s},\mathbf{e}}\left[\frac{1}{\sigma}\nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}} M(\mathbf{e}-\boldsymbol{\eta})\left(V_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} - \hat{V}^{\mathbf{v}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\right)\right] = 0. \end{aligned} \tag{26}$$

Since $\mathbb{E}_{\mathbf{e}}[\mathbf{e}-\boldsymbol{\eta}] = 0$, one can readily observe that

$$\begin{aligned} &\mathbb{E}_{\mathbf{s},\mathbf{e}}\left[\frac{1}{\sigma}\nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}} M(\mathbf{e}-\boldsymbol{\eta})\boldsymbol{\eta}^\top M\nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}}^\top\right] && \text{(27a)} \\ &\quad = \mathbb{E}_{\mathbf{s}}\left[\frac{1}{\sigma}\nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}} M\,\mathbb{E}_{\mathbf{e}}[\mathbf{e}-\boldsymbol{\eta}]\,\boldsymbol{\eta}^\top M\nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}}^\top\right] = 0, \\ &\mathbb{E}_{\mathbf{s},\mathbf{e}}\left[\frac{1}{\sigma}\nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}} M(\mathbf{e}-\boldsymbol{\eta})\left(V_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} - \hat{V}^{\mathbf{v}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\right)\right] && \text{(27b)} \\ &\quad = \mathbb{E}_{\mathbf{s}}\left[\frac{1}{\sigma}\nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}} M\,\mathbb{E}_{\mathbf{e}}[\mathbf{e}-\boldsymbol{\eta}]\left(V_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} - \hat{V}^{\mathbf{v}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\right)\right] = 0. \end{aligned}$$

Furthermore, we observe that

$$\begin{aligned} &\mathbb{E}_{\mathbf{s},\mathbf{e}}\left[\frac{1}{\sigma}\nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}} M(\mathbf{e}-\boldsymbol{\eta})\mathbf{e}^\top\left(\nabla_{\mathbf{a}} A_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} - \nabla_{\mathbf{a}}\hat{A}^{\mathbf{w}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\right)\right] = \\ &\mathbb{E}_{\mathbf{s}}\left[\nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}}\,\mathbb{E}_{\mathbf{e}}\left[\frac{1}{\sigma}M(\mathbf{e}-\boldsymbol{\eta})\mathbf{e}^\top\right]\left(\nabla_{\mathbf{a}} A_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} - \nabla_{\mathbf{a}}\hat{A}^{\mathbf{w}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\right)\right]. \end{aligned} \tag{28}$$

We then focus on the term inside the expectation on $\mathbf{e}$:

$$\begin{aligned} &\mathbb{E}_{\mathbf{e}}\left[\frac{1}{\sigma}M(\mathbf{e}-\boldsymbol{\eta})\mathbf{e}^\top\right] \\ &= \mathbb{E}_{\mathbf{e}}\left[\frac{1}{\sigma}M(\mathbf{e}-\boldsymbol{\eta})(\mathbf{e}-\boldsymbol{\eta})^\top\right] + \mathbb{E}_{\mathbf{e}}\left[\frac{1}{\sigma}M(\mathbf{e}-\boldsymbol{\eta})\boldsymbol{\eta}^\top\right] \\ &= \frac{1}{\sigma}M\mathrm{Cov}[\mathbf{e}\,|\,\mathbf{s}] + \frac{1}{\sigma}M\mathbb{E}_{\mathbf{e}}[(\mathbf{e}-\boldsymbol{\eta})]\boldsymbol{\eta}^\top = I. \end{aligned} \tag{29}$$

Let us finally focus on the second term in (26):

$$\begin{aligned} \mathbb{E}_{\mathbf{e}}\left[\frac{1}{\sigma}M(\mathbf{e}-\boldsymbol{\eta})\xi(\mathbf{e})\right] &= \mathbb{E}_{\mathbf{e}}\left[\frac{1}{\sigma}M(\mathbf{e}-\boldsymbol{\eta})\right]\xi(\boldsymbol{\eta}) \\ &+ \mathbb{E}_{\mathbf{e}}\left[\frac{1}{\sigma}M(\mathbf{e}-\boldsymbol{\eta})(\mathbf{e}-\boldsymbol{\eta})^\top\right]\nabla_{\mathbf{e}}\xi(\boldsymbol{\eta}) \\ &+ \sum_{k=2}^{\infty}\sum_{|\boldsymbol{\alpha}|=k}\frac{1}{\boldsymbol{\alpha}!}D^{\boldsymbol{\alpha}}\xi(\boldsymbol{\eta})\,M\mathbb{E}_{\mathbf{e}}\left[\frac{1}{\sigma}(\mathbf{e}-\boldsymbol{\eta})(\mathbf{e}-\boldsymbol{\eta})^{\boldsymbol{\alpha}}\right], \end{aligned} \tag{30}$$

where we used the multivariate Taylor expansion of ξ at $\boldsymbol{\eta}$. Using (19a), (18), and $\mathbb{E}_{\mathbf{e}}[\mathbf{e}-\boldsymbol{\eta}] = 0$, we observe that:

$$\lim_{\sigma\rightarrow 0}\mathbb{E}_{\mathbf{e}}\left[\frac{1}{\sigma}M(\mathbf{e}-\boldsymbol{\eta})\xi(\mathbf{e})\right] = \lim_{\sigma\rightarrow 0}\nabla_{\mathbf{e}}\xi(\boldsymbol{\eta}) = 0, \tag{31}$$

where the second equality follows from (19b) and observing that the Taylor theorem ensures that

$$\nabla_{\mathbf{e}}\xi(\boldsymbol{\eta}) = \left(\nabla_{\mathbf{a}} A_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{s},\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s})) + \varsigma(\boldsymbol{\eta})\right)\boldsymbol{\eta} \tag{32}$$

for some function ς such that $\lim_{\boldsymbol{\eta}\rightarrow 0}\varsigma(\boldsymbol{\eta}) = 0$.

Using (27)-(31) in (26), the optimality conditions (21) read

$$\lim_{\sigma\rightarrow 0}\mathbb{E}_{\mathbf{s}}\left[\nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}}\left(\nabla_{\mathbf{a}} A_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} - \nabla_{\mathbf{a}}\hat{A}^{\mathbf{w}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\right)\right] = 0, \tag{33}$$

which delivers (23). ■

We prove next that a slight relaxation of the condition (18) is sufficient to obtain the result of Theorem 2.

Corollary 1: The results of Theorem 2 can be generalized to the case in which (18) is replaced by

$$\lim_{\sigma\rightarrow 0}\nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}} M\,\mathbb{E}_{\mathbf{e}}[\mathbf{e}-\boldsymbol{\eta}] = 0, \tag{34}$$

$$\lim_{\sigma\rightarrow 0}\nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}} M\,\frac{1}{\sigma}\Sigma_{\mathbf{e}}(\mathbf{s},\boldsymbol{\theta}) = \beta\nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}}, \tag{35}$$

for some $\beta > 0$ satisfying $\lim_{\sigma\rightarrow 0}\beta \neq 0$.

Proof: One can verify that (34) are sufficient to guarantee that (27) hold in the limit $\sigma \rightarrow 0$, and (31) holds. Additionally, by pre-multiplying (29) by $\nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}}$ we obtain

$$\nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}}\mathbb{E}_{\mathbf{e}}\left[\frac{1}{\sigma}M(\mathbf{e}-\boldsymbol{\eta})\mathbf{e}^\top\right]\nabla_{\boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}} M\,\frac{1}{\sigma}\Sigma_{\mathbf{e}}(\mathbf{s},\boldsymbol{\theta}),$$

such that (33) becomes

$$\lim_{\sigma\rightarrow 0}\mathbb{E}_{\mathbf{s}}\left[\beta\nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}}\left(\nabla_{\mathbf{a}} A_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} - \nabla_{\mathbf{a}}\hat{A}^{\mathbf{w}}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\right)\right] = 0. \tag{36}$$

■

IV. Policy gradient for MPC-based policies

Section III established that the mean and covariance of the exploration $\mathbf{e}$ must be known (at least in the range space of the gradient of the policy $\nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}}$). In this section, we further discuss the problem of obtaining this mean and covariance in the case of constrained exploration, as in the projection (9).

It can be helpful here to recast (9) in terms of the transformation of disturbance $\mathbf{d}$ and exploration $\mathbf{e}$:

$$\begin{aligned} \mathbf{e} = \arg\min_{\mathbf{e}} \quad & \frac{1}{2}\|\mathbf{e}-\mathbf{d}\|^2 && \text{(37a)} \\ \text{s.t.} \quad & \mathbf{H}(\mathbf{s},\boldsymbol{\pi}_{\boldsymbol{\theta}}+\mathbf{d},\boldsymbol{\theta}) \leq 0. && \text{(37b)} \end{aligned}$$

We recall that $\mathbf{d}$ is a random variable whose distribution can be selected and is therefore known. One can readily

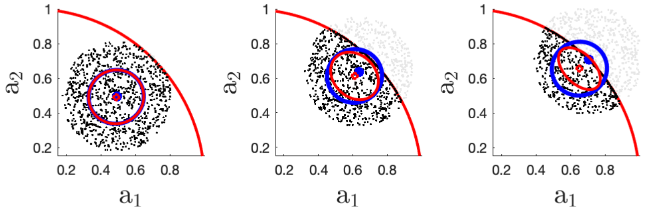

Fig. 1: Illustration of the mean and covariance of $\mathbf{a}^{\mathrm{d}}$ and $\mathbf{a}^{\mathrm{e}}$ subject to the projection on the feasible set of a constraint (solid red line). The samples $\mathbf{a}^{\mathrm{d}} = \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}) + \mathbf{d}$ are represented as light dots, and the samples of $\mathbf{a}^{\mathrm{e}} = \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}) + \mathbf{e}$ are represented as black dots. The mean of $\mathbf{a}^{\mathrm{d}}$ is represented as the blue dot, and its covariance as the blue circle. The mean of $\mathbf{a}^{\mathrm{e}}$ is depicted as the red dot and its covariance is illustrated via the red ellipsoid. When the policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s})$ nears the constraint, the mean and covariance of $\mathbf{a}^{\mathrm{e}}$ does not match the ones of $\mathbf{a}^{\mathrm{d}}$. As a result, the mean and covariance of $\mathbf{e}$ do not match the ones of $\mathbf{d}$, and are not trivial to evaluate.

observe that if $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ is sufficiently inside the feasible domain of the constraints (37b), then $\mathbf{e} = \mathbf{d}$ such that the mean and covariance of $\mathbf{e}$ are identical to those chosen for $\mathbf{d}$. However, when $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ is close to the boundary of the feasible set of the constraints (37b), then $\mathbf{e} \neq \mathbf{d}$ for some samples $\mathbf{d}$, and obtaining the mean and covariance of $\mathbf{e}$ become more intricate, as illustrated in Fig. 1.

Let us label $\mathbb{F}_{\boldsymbol{\theta}}$ the feasible set of the MPC scheme (7), i.e., the set of states $\mathbf{s}$ for which (37) has a solution. We further define the following subset of $\mathbb{F}_{\boldsymbol{\theta}}$:

$$\mathbb{D}_{\boldsymbol{\theta}} = \left\{ \, \mathbf{s} \in \mathbb{F}_{\boldsymbol{\theta}} \;\middle|\; \mathbb{P}\left[\mathbf{H}\left(\mathbf{s}, \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}) + \mathbf{d}, \boldsymbol{\theta}\right) > 0\right] > 0 \, \right\}, \tag{38}$$

describing the set of states for which the disturbance $\mathbf{d}$ has a positive probability of violating the constraints (37b), and therefore of requiring a projection. We can then observe that

$$\boldsymbol{\mu}_{\mathbf{e}} = \boldsymbol{\mu}_{\mathbf{d}}, \quad \Sigma_{\mathbf{e}} = \Sigma_{\mathbf{d}}, \qquad \forall \, \mathbf{s} \in \mathbb{F}_{\boldsymbol{\theta}} \setminus \mathbb{D}_{\boldsymbol{\theta}}, \tag{39}$$

such that $\boldsymbol{\mu}_{\mathbf{e}}\left(\mathbf{s}\right)$, $\Sigma_{\mathbf{e}}\left(\mathbf{s}\right)$ are readily known on the set $\mathbb{F}_{\boldsymbol{\theta}} \backslash \mathbb{D}_{\boldsymbol{\theta}}$. However, for $\mathbf{s} \in \mathbb{D}_{\boldsymbol{\theta}}$, (39) will in general not hold, and $\boldsymbol{\mu}_{\mathbf{e}}$, $\Sigma_{\mathbf{e}}$ must be evaluated. We observe that in general $\mathbb{D}_{\boldsymbol{\theta}}$ has a positive measure under $\varrho_{\mathbf{s}}$, such that this evaluation cannot be dismissed.

The mean $\boldsymbol{\mu}_{\mathbf{e}}$ and covariance $\Sigma_{\mathbf{e}}$ can be evaluated on $\mathbb{D}_{\boldsymbol{\theta}}$ to an arbitrary accuracy by sampling (37) whenever $\mathbf{s} \in \mathbb{D}_{\boldsymbol{\theta}}$. This is done by drawing a number of samples $\mathbf{d}_{1,\ldots,N_{\mathrm{s}}}$, performing (37) on each of those samples to obtain the corresponding samples $\mathbf{e}_{1,\ldots,N_{\mathrm{s}}}$, and use the estimations:

$$\hat{\Sigma}_{\mathbf{e}} = \frac{1}{N_{\mathrm{s}} - 1} \sum_{k=1}^{N_{\mathrm{s}}} (\mathbf{e}_k - \hat{\boldsymbol{\mu}}_{\mathbf{e}})(\mathbf{e}_k - \hat{\boldsymbol{\mu}}_{\mathbf{e}})^\top, \quad \hat{\boldsymbol{\mu}}_{\mathbf{e}} = \frac{1}{N_{\mathrm{s}}} \sum_{k=1}^{N_{\mathrm{s}}} \mathbf{e}_k. \tag{40}$$

Unfortunately, this must be performed for every encountered state $\mathbf{s}$, and is fairly expensive to carry out unless (37) is in a form that can be solved at a minimum computational cost. To our best knowledge, this observation is general, i.e., there is no simple way to evaluate $\Sigma_{\mathbf{e}}$, $\boldsymbol{\mu}_{\mathbf{e}}$ on $\mathbb{D}_{\boldsymbol{\theta}}$.

V. Example

In this section, we provide an example of the DPG error discussed in this paper, and of the proposed correction. In order to show very clear results, we will adopt a trivial and constructed problem, which is very easily reproducible. Consider the linear scalar dynamics with scalar input a:

$$s_+ = 0.97s + 0.1a + n, \tag{41}$$

where n is a scalar process noise uniformly distributed in the interval $\left[-10^{-3},\, 10^{-3}\right]$. The baseline cost will be

$$L\left(s, a\right) = 10\left(s - 0.5\right)^2 + \left(a - 0.15\right)^2. \tag{42}$$

The MPC scheme is set up as:

$$\min_{u} \; \sum_{k=0}^{50} \gamma^k \left(10\left(x_k - \frac{1}{3}\right)^2 + \left(u_k - u_{\mathrm{ref}}\left(\theta\right)\right)^2\right) \tag{43a}$$

$$\text{s.t.} \quad x_{k+1} = 0.97x_k + 0.1u_k, \quad x_0 = s, \tag{43b}$$

$$u_k \leq \theta, \tag{43c}$$

where $u_{\mathrm{ref}}\left(\theta\right) = 0.2 - \theta$. The MPC scheme receives a deterministic model matching the expected state transition of the real system, but an incorrect cost function, and includes an input upper bound. The initial RL parameter $\theta = 0.1$ is selected. The MPC policy can be improved by increasing the input bound in (43c) and increasing the input reference u_{ref}. However, raising the input bound in (43c) by increasing θ results in decreasing the input reference u_{ref}, such that these terms are in conflict for improving the policy.

The example is simple enough that $M(s)$ and $\boldsymbol{\eta}(s)$ can be explicitly evaluated. Moreover, the exact advantage function $A_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ can obtained via policy evaluation techniques, and its gradient can be computed via finite differences. Using roll-outs of the closed-loop trajectories to build the expected values $\mathbb{E}_{\mathbf{s}}[\cdot]$, we can evaluate the policy gradient $\widehat{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})}$ from (12) and the classic advantage function approximation (15), which we will label as $\widehat{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})}^{\mathrm{classic}}$ in the figures, and we can evaluate $\widehat{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})}$ from (12) and using the corrected advantage function approximation (17), which we will label as $\widehat{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})}^{\mathrm{corrected}}$ in the figures.

The disturbance $\mathbf{d}$ was chosen uniformly distributed in the interval $\left[-10^{-3},\, 10^{-3}\right]$, yielding a small σ, making the experiments close to the theoretical results. A discount of $\gamma = 0.9$ was selected. The learning was performed in an episodic fashion.

Fig. 2 depicts the evolution of the closed-loop performance and MPC parameter when using $\widehat{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})}^{\mathrm{classic}}$ and $\widehat{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})}^{\mathrm{corrected}}$ for learning. One can see that learning using $\widehat{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})}^{\mathrm{classic}}$ yields a parameter which results in worse performance than the one obtained with $\widehat{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})}^{\mathrm{corrected}}$. Fig. 3 depicts the evolution of the different policy gradients $\widehat{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})}^{\mathrm{corrected}}$, $\widehat{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})}^{\mathrm{classic}}$ and $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$ when learning is performed with both $\widehat{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})}^{\mathrm{classic}}$ and $\widehat{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})}^{\mathrm{corrected}}$. One can readily observe that learning from $\widehat{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})}^{\mathrm{classic}}$ yields a gradient error and a corresponding loss of performance, while $\widehat{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})}^{\mathrm{corrected}}$ is very close to the true policy gradient $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$, and achieves optimality.

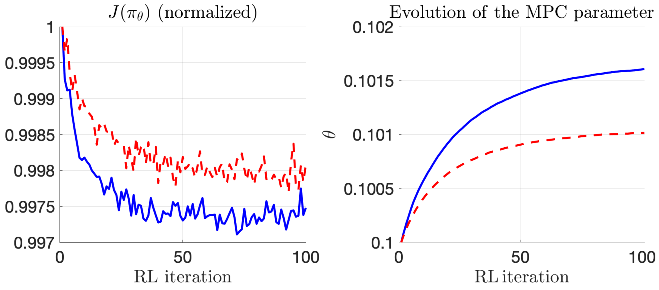

Fig. 2: The left graph shows the evolution of $J(\pi_\theta)$ over the RL iterations. The red (dashed) curve is the outcome of learning from $\widehat{\nabla_{\boldsymbol{\theta}} J}(\boldsymbol{\pi}_{\boldsymbol{\theta}})^{\text{classic}}$, while the blue (solid) curve is the one from $\widehat{\nabla_{\boldsymbol{\theta}} J}(\boldsymbol{\pi}_{\boldsymbol{\theta}})^{\text{corrected}}$. The right graph displays the corresponding evolution of the MPC parameter θ.

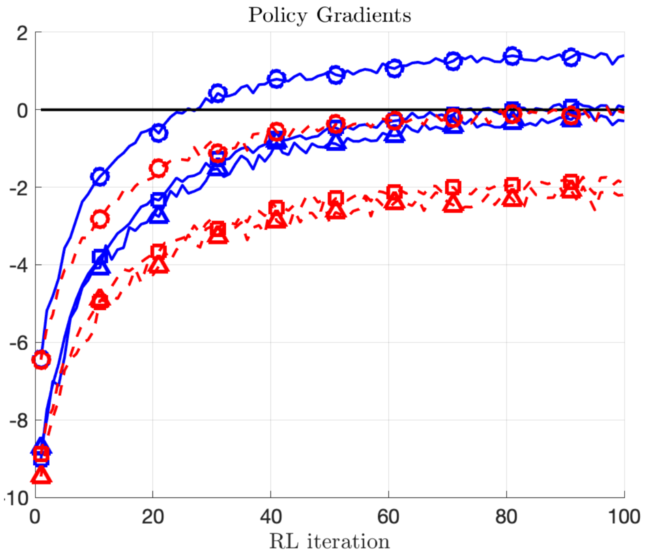

Fig. 3: Policy gradients over the RL iterations when learning from $\widehat{\nabla_{\boldsymbol{\theta}} J}(\boldsymbol{\pi}_{\boldsymbol{\theta}})^{\text{classic}}$ (red, dashed curves), and from $\widehat{\nabla_{\boldsymbol{\theta}} J}(\boldsymbol{\pi}_{\boldsymbol{\theta}})^{\text{corrected}}$ (blue, solid curves). The curves marked with $\circ$ depict $\widehat{\nabla_{\boldsymbol{\theta}} J}(\boldsymbol{\pi}_{\boldsymbol{\theta}})^{\text{classic}}$, the curves marked by $\square$ depict $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$, and the curves marked by $\triangle$ depict $\widehat{\nabla_{\boldsymbol{\theta}} J}(\boldsymbol{\pi}_{\boldsymbol{\theta}})^{\text{corrected}}$.

VI. Conclusions

This paper discussed the implementation of the Actor-Critic Deterministic Policy Gradient method using a compatible advantage function approximation when the policy is restricted by constraints. We showed that if the exploration ought to respect the constraints imposed in the policy, then the classical compatible advantage function approximation fails to deliver a correct policy gradient. We proposed a generalization of the compatible advantage function approximation correcting this problem, which requires to estimate the mean and covariance of the exploration. We further discussed the problem and solution in the context of MPC-based policies, and demonstrated them on a simple example.

References

[1] Pieter Abbeel, Adam Coates, Morgan Quigley, and Andrew Y. Ng. An application of reinforcement learning to aerobatic helicopter flight. In *In Advances in Neural Information Processing Systems 19*, page 2007. MIT Press, 2007.

[2] Brandon Amos, Ivan Dario Jimenez Rodriguez, Jacob Sacks, Byron Boots, and J. Zico Kolter. Differentiable mpc for end-to-end planning and control. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 8299–8310, USA, 2018. Curran Associates Inc.

[3] D.P. Bertsekas. *Reinforcement Learning and Optimal Control*. Athena Scientific, 2009.

[4] L. Chisci, J.A. Rossiter, and G. Zappa. Systems with persistent disturbances: predictive control with restricted constraints. *Automatica*, 37:1019–1028, 2001.

[5] S. Gros and M. Zanon. Data-Driven Economic NMPC Using Reinforcement Learning. *IEEE Transactions on Automatic Control*, 65(2):636–648, Feb 2020.

[6] S. Gros and M. Zanon. Reinforcement Learning for Mixed-Integer Problems Based on MPC. In *21st IFAC World Congress*, 2020.

[7] S. Gros, M. Zanon, and A. Bemporad. Safe Reinforcement Learning via Projection on a Safe Set: How to Achieve Optimality? In *21st IFAC World Congress*, 2020.

[8] Torsten Koller, Felix Berkenkamp, Matteo Turchetta, and Andreas Krause. Learning-based Model Predictive Control for Safe Exploration and Reinforcement Learning. Published on Arxiv, 2018.

[9] F. L. Lewis and D. Vrabie. Reinforcement learning and adaptive dynamic programming for feedback control. *IEEE Circuits and Systems Magazine*, 9(3):32–50, 2009.

[10] F. L. Lewis, D. Vrabie, and K. G. Vamvoudakis. Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers. *IEEE Control Systems*, 32(6):76–105, 2012.

[11] T. Mannucci, E. van Kampen, C. de Visser, and Q. Chu. Safe exploration algorithms for reinforcement learning controllers. *IEEE Transactions on Neural Networks and Learning Systems*, 29(4):1069–1081, 2018.

[12] A. B. Martinsen, A. M. Lekkas, and S. Gros. Combining Ssystem Identification with Reinforcement Learning-Based MPC. In *21st IFAC World Congress*, 2020. (accepted).

[13] Ryan Murray and Michele Palladino. A model for system uncertainty in reinforcement learning. *Systems & Control Letters*, 122:24 – 31, 2018.

[14] Chris J. Ostafew, Angela P. Schoellig, and Timothy D. Barfoot. Robust Constrained Learning-based NMPC enabling reliable mobile robot path tracking. *The International Journal of Robotics Research*, 35(13):1547–1563, 2016.

[15] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *Proceedings of the 31st International Conference on Machine Learning*, ICML'14, pages I–387–I–395, 2014.

[16] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, MA, USA, 2nd edition, 2018.

[17] Kim P. Wabersich and Melanie N. Zeilinger. Safe exploration of nonlinear dynamical systems: A predictive safety filter for reinforcement learning. *arXiv e-prints*, 2018.

[18] Shouyi Wang, Wanpracha Chaovalitwongse, and Robert Babuska. Machine learning algorithms in bipedal robot control. *Trans. Sys. Man Cyber Part C*, 42(5):728–743, September 2012.

[19] M. Zanon and Gros. Safe Reinforcement Learning Using Robust MPC. *Transaction on Automatic Control, (in press)*, 2021.

[20] M. Zanon, S. Gros, and A. Bemporad. Practical Reinforcement Learning of Stabilizing Economic MPC. In *Proceedings of the European Control Conference*, 2019.

[21] M. Zanon, V. Kungurtsev, and S. Gros. Reinforcement Learning Based on Real-Time Iteration NMPC. In *21st IFAC World Congress*, 2020.