Stochastic Differential Dynamic Programming

Evangelos Theodorou, Yuval Tassa & Emo Todorov

Abstract—Although there has been a significant amount of work in the area of stochastic optimal control theory towards the development of new algorithms, the problem of how to control a stochastic nonlinear system remains an open research topic. Recent iterative linear quadratic optimal control methods iLQG [1], [2] handle control and state multiplicative noise while they are derived based on first order approximation of dynamics. On the other hand, methods such as Differential Dynamic Programming expand the dynamics up to the second order but so far they can handle nonlinear systems with additive noise.

In this work we present a generalization of the classic Differential Dynamic Programming algorithm. We assume the existence of state and control multiplicative process noise, and proceed to derive the second-order expansion of the cost-to-go. We find the correction terms that arise from the stochastic assumption. Despite having quartic and cubic terms in the initial expression, we show that these vanish, leaving us with the same quadratic structure as standard DDP.

I. INTRODUCTION

Optimal Control describes the choice of actions that minimize future costs under the constraint of state space dynamics. In the continuous nonlinear case, local methods are one of the very few classes of algorithms which successfully solve general, high-dimensional Optimal Control problems. These methods are based on the observation that optimal solutions form extremal trajectories, i.e. are solutions to a calculus-of-variations problem.

Differential Dynamic Programming, or DDP, is a powerful local dynamic programming algorithm, which generates both open and closed loop control policies along a trajectory. The DDP algorithm, introduced in [3], computes a quadratic approximation of the cost-to-go and correspondingly, a local linear-feedback controller. The state space dynamics are also quadratically approximated around a trajectory. As in the Linear-Quadratic-Gaussian case, fixed additive noise has no effect on the controllers generated by DDP, and when described in the literature, the dynamics are usually assumed deterministic.

Past work on nonlinear optimal control allows the use of DDP in problems with state and control constraints [4],[5]. In this work, state and control constraints are expanded up to the first order and the KKT conditions are formulated resulting in an unconstrained quadratic optimization problem.

The application of DDP in real robotic high dimensional control tasks created the need for extensions that relax the condition for accurate models. In this vain, in [6] DDP is extended for the case of min - max formulation. The goal for this formulation is to make DDP robust against the model uncertainties and hybrid dynamics in a biped robotic locomotion task. In [7] implementation improvements regarding the evaluation of the value function allow the use of DDP in a receding horizon mode. In all the past work related to DDP, the optimal control problem is considered to be deterministic.

While the impartiality to noise can be considered a feature if the noise is indeed fixed, in many cases varying noise covariance is an important feature of the problem, as with control-multiplicative noise which is common in biological systems [1]. This latter case was addressed within the iterative-LQG framework [1], in which the optimal controller is derived based on the first order approximation of the dynamics and the second order approximation of the cost to go. However, for the iterative nonlinear optimal control algorithms such as DDP in which second order expansion of the dynamics is considered, the more general case of state and/or control multiplicative noise appears to have never been tackled.

In this paper, we derive the DDP algorithm for state and control multiplicative process noise. We find that despite the potential of cubic and quartic terms, these cancel out, allowing us to maintain the quadratic form of the approximation. Moreover we show how the new generalized formulation of Stochastic Differential Dynamic Programming (SDDP) recovers the standard DDP deterministic solution as well as the special cases in which only state multiplicative or control multiplicative noise is considered.

The remaining of this work is organized as follows: in the next section we provide the definition of the SDDP. In section III, the second order expansion of the cost to go is presented and in section IV the optimal controls are derived and the overall SDDP algorithm is presented. In addition in section IV we show how SDDP recovers the deterministic solution as well as the cases of only control multiplicative, only state multiplicative and only additive noise. Finally in sections V and VI simulation results and conclusions are discussed.

II. STOCHASTIC DIFFERENTIAL DYNAMIC PROGRAMMING

We consider the class of nonlinear stochastic optimal control problems with cost

\[ v^*(x, t) = E \left[ h(x(T)) + \int_{t_0}^{T} \ell(\tau, x(\tau), \pi(\tau, x(\tau))) d\tau \right] \]  (1)
subject to the stochastic dynamics of the form:
\[ dx = f(x, u)dt + F(x, u)d\omega \]  
where \( x \in \mathbb{R}^{n \times 1} \) is the state, \( u \in \mathbb{R}^{p \times 1} \) is the control and \( d\omega \in \mathbb{R}^{m \times 1} \) is brownian noise. The term \( h(x(T)) \) in the cost function (1), is the terminal cost while the \( \ell (\tau, x(\tau), \pi(\tau, x(\tau))) \) is the instantaneous cost rate which is a function of the state \( x \) and control policy \( \pi(\tau, x(\tau)) \). 

The cost-to-go \( v^\pi(x, t) \) is defined as the expected cost accumulated over the time horizon \((t_0, ..., T)\) starting from the initial state \( x_t \) to the final state \( x(T) \).

To enhance the readability of our derivations we write the dynamics as a function \( \Phi \in \mathbb{R}^{n \times 1} \) of the state, control and instantiation of the noise:
\[ \Phi(x, u, d\omega) \equiv f(x, u)dt + F(x, u)d\omega \]  
It will sometimes be convenient to write the matrix \( F(x, u) \in \mathbb{R}^{n \times m} \) in terms of its rows or columns:
\[ F(x, u) = \begin{bmatrix} F^1_r(x, u) \\ \vdots \\ F^p_r(x, u) \end{bmatrix} \]

Every element of the vector \( \Phi(x, u, d\omega) \in \mathbb{R}^{n \times 1} \) can now be expressed as:
\[ \Phi^j(x, u, d\omega) = f^j(x, u)dt + F^j_r(x, u)d\omega \]

Given a nominal trajectory of states and controls \((\bar{x}, \bar{u})\) we expand the dynamics around this trajectory to second order:
\[ \Phi(\bar{x} + \delta x, \bar{u} + \delta u, d\omega) = \Phi(\bar{x}, \bar{u}, d\omega) + \nabla_x \Phi \cdot \delta x + \nabla_u \Phi \cdot \delta u + O(\delta x, \delta u, d\omega) \]

where \( O(\delta x, \delta u, d\omega) \in \mathbb{R}^{n \times 1} \) contains all the second order terms in the deviations in states, controls and noise\(^1\). Writing this term element-wise:
\[ O(\delta x, \delta u, d\omega) = \begin{pmatrix} O^{(1)}(\delta x, \delta u, d\omega) \\ \vdots \\ O^{(m)}(\delta x, \delta u, d\omega) \end{pmatrix} \]

we can express the elements \( O^{(j)}(\delta x, \delta u, d\omega) \in \mathbb{R} \) as:
\[ O^{(j)}(\delta x, \delta u, d\omega) = \begin{pmatrix} \frac{1}{2} (\delta x)^T \\ 1 \end{pmatrix} \begin{pmatrix} \nabla_x \Phi^j \\ \nabla_u \Phi^j \end{pmatrix} \begin{pmatrix} \delta x \\ \delta u \end{pmatrix} . \]

We would now like to express the derivatives of \( \Phi \) in terms of the given quantities. Beginning with the first-order terms, we find that:
\[ \nabla_x \Phi = \nabla_x f(x, u)dt + \nabla_x \left( \sum_{i=1}^{m} F^i_c dw_i^{(s)} \right) \]
\[ \nabla_u \Phi = \nabla_u f(x, u)dt + \nabla_u \left( \sum_{i=1}^{m} F^i_c dw_i^{(s)} \right) \]

Next we find the second order derivatives and we have that:
\[ \nabla_x \Phi^j = \nabla_x f^j(x, u)dt + \nabla_x \left( \sum_{i=1}^{m} F^i_c^{(j)} dw_i^{(s)} \right) \]
\[ \nabla_u \Phi^j = \nabla_u f^j(x, u)dt + \nabla_u \left( \sum_{i=1}^{m} F^i_c^{(j)} dw_i^{(s)} \right) \]

After expanding the dynamics up to the second order we can transition from continuous to discrete time. More precisely the discrete-time dynamics are formulated as:
\[ \delta x_{t+\delta t} = \left( (I_n + \nabla_x f(x, u) \delta t + \nabla_x \left( \sum_{i=1}^{m} F^i_c^{(s)} \xi_i \sqrt{\delta t} \right) \right) \delta x_t \]
\[ + \left( \nabla_u f(x, u) \delta t + \nabla_u \left( \sum_{i=1}^{m} F^i_c^{(j)} \xi_i \sqrt{\delta t} \right) \right) \delta u_t \]
\[ + F(x, u) \sqrt{\delta t} \xi_t + O_{d}(\delta x, \delta u, \xi_t) \]

with \( \delta t = t_{k+1} - t_k \) corresponding to a small discretization interval. Note that the term \( O_d \) is the equivalent of \( O \) but in discrete time and therefore it is now a function of \( \delta t \). In fact, since \( O_{d} \) contains all the second order expansion terms of the dynamics it contains second order derivatives \( WRT \) state and control expressed as follows:
\[ \nabla_x \Phi^j = \nabla_x f^j(x, u) \delta t + \nabla_x \left( \sum_{i=1}^{m} F^i_c^{(j)} \xi_i \sqrt{\delta t} \right) \]
\[ \nabla_u \Phi^j = \nabla_u f^j(x, u) \delta t + \nabla_u \left( \sum_{i=1}^{m} F^i_c^{(j)} \xi_i \sqrt{\delta t} \right) \]
\[ \nabla_{ux} \Phi^j = \nabla_{ux} f^j(x, u) \delta t + \nabla_{ux} \left( \sum_{i=1}^{m} F^i_c^{(j)} \xi_i \sqrt{\delta t} \right) \]

The random variable \( \xi \in \mathbb{R}^{m \times 1} \) is zero mean and Gaussian distributed with covariance \( \Sigma = \sigma^2 I_{n \times m} \). The discretized dynamics can be written in a more compact form by grouping the state, control and noise dependent terms, and leaving the second order term separate:
\[ \delta x_{t+\delta t} = A_t \delta x_t + B_t \delta u_t + \Gamma_t \xi_t + O_d(\delta x, \delta u, \xi_t) \]

where the matrices \( A_t \in \mathbb{R}^{n \times n}, B_t \in \mathbb{R}^{n \times p} \) and \( \Gamma_t \in \mathbb{R}^{n \times m} \) are defined as:
\[ A_t = I_{n \times n} + \nabla_x f(x, u) \delta t \]
\[ B_t = \nabla_u f(x, u) \delta t \]
\[ \Gamma_t = \begin{bmatrix} \Gamma^{(1)} \\ \Gamma^{(2)} \\ \vdots \\ \Gamma^{(m)} \end{bmatrix} \]

with \( \Gamma^{(i)} \in \mathbb{R}^{n \times 1} \) defined \( \Gamma^{(i)} = \nabla_u F_c^{(i)} \delta u_t + \nabla_x F_c^{(i)} \delta x_t + F_c^{(i)} \). For the derivation of the optimal control it is useful to expresses \( \Gamma_t \) as the summation of terms that depend on variations in state and controls and terms that are independent of such variations. More precisely we will have that:

\(^1\)Not to be confused with “big-O”.

1126
\[ \Gamma_t = \Delta_t(\delta x, \delta u) + F(x, u) \] (5)

where each column vector of \( \Delta_t \) is defined as
\[ \Delta_t(i)(\delta x, \delta u) = \nabla_u F(i) \delta t + \nabla_x F(i) \delta x. \]

III. Value Function Second Order Approximation

As in classical DDP, the derivation of stochastic DDP requires the second order expansion of the cost-to-go function around a nominal trajectory \( \bar{x} \):

\[ V(\bar{x} + \delta x) = V(\bar{x}) + V_x^T \delta x + \frac{1}{2} \delta x^T V_{xx} \delta x \] (6)

Substitution of the discretized dynamics (4) in the second order Value function expansion (6) results in:

\[
\begin{align*}
V(\bar{x}_{t+\delta t} + \delta x_{t+\delta t}) &= V(\bar{x}_{t+\delta t}) \\
&+ V_x^T \left( A_t \delta x_t + B_t \delta u_t + \Gamma_t \xi + O_d \right) \\
&+ (A_t \delta x_t + B_t \delta u_t + \Gamma_t \xi + O_d)^T \\
&\times V_{xx} \left( A_t \delta x_t + B_t \delta u_t + \Gamma_t \xi + O_d \right)
\end{align*}
\] (7)

Next we will compute \( E(V(\bar{x}_{t+\delta t} + \delta x_{t+\delta t})) \) which requires the calculation of the expectation of the all the terms that appear in the equation above. This is what the rest of the analysis is dedicated to. More precisely in the next two sections we will calculate the expectation of the terms:

\[ E \left( V_x^T \delta x_{t+\delta t} \right) \] (8)

and

\[ E \left( \delta x_{t+\delta t}^T V_{xx} \delta x_{t+\delta t} \right) \] (9)

where the state deviation \( \delta x_{t+\delta t} \) at time instant \( t + \delta t \) is given by the linearized dynamics:

\[ \delta x_{t+\delta t} = A_t \delta x_t + B_t \delta u_t + \Gamma_t \xi + O_d \] (10)

The analysis that follows in section III-A consist of the computation of the expectation of the four terms which result from the substitution of the linearized dynamics (10) into (8). In section III-B we compute the expectation of the 16 terms that result from the substitution of (10) into (9).

A. Expectation of the first order term of the value function expansion \( \nabla_x V^T \delta x_{t+\delta t} \).

The expectation of the first order term results in:

\[ E \left( V_x^T \left( A_t \delta x_t + B_t \delta u_t + \Gamma_t \xi + O_d \right) \right) = V_x^T \left( A_t \delta x_t + B_t \delta u_t + E(O_d) \right) \] (11)

In order to find the expectation of \( O_d \in \mathbb{R}^{n \times 1} \) we need to find the expectation of each one of the elements of this column vector. Thus we will have that:

\[ E \left( O^{(j)}(\delta x, \delta u, \xi, \delta t) \right) = \]

\[ E \left( \frac{1}{2} \begin{pmatrix} \delta x \\ \delta u \end{pmatrix}^T \begin{pmatrix} \nabla_{xx} \Phi^{(j)} & \nabla_{xu} \Phi^{(j)} \\ \nabla_{ux} \Phi^{(j)} & \nabla_{uu} \Phi^{(j)} \end{pmatrix} \begin{pmatrix} \delta x \\ \delta u \end{pmatrix} \right) = \]

\[ = \frac{\delta t}{2} \begin{pmatrix} \delta x \\ \delta u \end{pmatrix}^T \begin{pmatrix} \nabla_{xx} f^{(j)} & \nabla_{xu} f^{(j)} \\ \nabla_{ux} f^{(j)} & \nabla_{uu} f^{(j)} \end{pmatrix} \begin{pmatrix} \delta x \\ \delta u \end{pmatrix} = \tilde{\delta}^j \]

Therefore we will have that:

\[ E \left( V_x^T \delta x_{t+\delta t} \right) = V_x^T \left( A_t \delta x_t + B_t \delta u_t + \tilde{O}_d \right) \] (13)

Where the term \( \tilde{O}_d \) is defined as:

\[ \tilde{O}_d(\delta x, \delta u, \delta t) = \left( \begin{array}{c} \hat{O}^{(1)}(\delta x, \delta u, \delta t) \\ \vdots \\ \hat{O}^{(n)}(\delta x, \delta u, \delta t) \end{array} \right) \] (14)

The term \( \nabla_x V^T \tilde{O}_d \) is quadratic in variations in the states and controls \( \delta x, \delta u \) as there thus are the symmetric matrices \( \mathcal{F} \in \mathbb{R}^{n \times n} \), \( \mathcal{Z} \in \mathbb{R}^{m \times m} \) and \( \mathcal{L} \in \mathbb{R}^{m \times n} \) such that:

\[ V_x^T \tilde{O}_d = \frac{1}{2} \delta x^T \mathcal{F} \delta x + \frac{1}{2} \delta u^T \mathcal{Z} \delta u + \delta u^T \mathcal{L} \delta x \] (15)

with

\[ \mathcal{F} = \sum_{j=1}^{n} \nabla_{xx} f^{(j)} V_{x_j} \] (16)

\[ \mathcal{Z} = \sum_{j=1}^{n} \nabla_{uu} f^{(j)} V_{x_j} \] (17)

\[ \mathcal{L} = \sum_{j=1}^{n} \nabla_{ux} f^{(j)} V_{x_j} \] (18)

From the analysis above we can see that the expectation \( \nabla_x V^T \delta x_{t+\delta t} \) is a quadratic function with respect to variations in states and controls \( \delta x, \delta u \). As we will prove in the next section the expectation of \( \delta x_{t+\delta t}^T \nabla_{xx} V^T \delta x_{t+\delta t} \) is also a quadratic function of variations in states and controls \( \delta x, \delta u \).

B. Expectation of the second order term of the value function expansion \( \delta x_{t+\delta t}^T \nabla_{xx} V^T \delta x_{t+\delta t} \).

In this section we compute all the terms that appear due to the expectation of the second approximation of the value function \( E(\delta x_{t+\delta t}^T \nabla_{xx} V \delta x_{t+\delta t}) \). The term \( \delta x_{t+\delta t} \) is given by the stochastic dynamics in (10). Substitution of (10) results in 16 terms. To make our analysis clear we classify these 16 terms above into five classes. More precisely we will have that:
\[ E(\delta x_{t+\delta t}^T V_{xx}^T \delta x_{t+\delta t}) = E_1 + E_2 + E_3 + E_4 + E_5 \] (19)

where the terms \( E_1, E_2, E_3, E_4 \) and \( E_5 \) are defined as follows:

\[
E_1 = E(\delta x_t^T A_t^T V_{xx} A_t \delta x_t) + E(\delta u_t^T B_t^T V_{xx} B_t \delta u_t) + \delta x_t^T A_t^T V_{xx} A_t^T \delta x_t^T + E(\delta u_t^T B_t^T V_{xx} A_t \delta x_t) + E(\delta u_t^T B_t^T V_{xx} A_t \delta x_t)
\] (20)

\[
E_2 = E(\xi_t^T \Gamma_t^T V_{xx} A_t \delta x_t) = E(\xi_t^T \Gamma_t^T V_{xx} A_t \delta x_t) + E(\delta u_t^T B_t^T V_{xx} \Gamma_t \xi_t) + E(\delta u_t^T B_t^T V_{xx} \Gamma_t \xi_t) + E(\xi_t^T \Gamma_t^T V_{xx} \Gamma_t \xi_t)
\] (21)

\[
E_3 = E(O_d^T V_{xx} \Gamma_t \xi_t) + E(\xi_t^T \Gamma_t^T V_{xx} O_d)
\] (22)

\[
E_4 = E(\delta x_t^T A_t^T V_{xx} O_d) + E(\delta u_t^T B_t^T V_{xx} O_d) + E(\xi_t^T \Gamma_t^T V_{xx} O_d) + E(O_d^T V_{xx} A_t \delta x_t)
\] (23)

\[
E_5 = E(O_d^T V_{xx} O_d)
\] (24)

In the first category we have all these terms that depend neither on \( \xi_t \) nor on \( O_d(\delta x, \delta u, \xi_t, \delta t) \). The second category \( E_2 \) includes terms that depend on \( \xi_t \) but not on \( O_d(\delta x, \delta u, \xi_t, \delta t) \). In the third class \( E_3 \), there are terms that depend both on \( O_d(\delta x, \delta u, \xi_t, \delta t) \) and \( \xi_t \). In the fourth category \( E_4 \) we have terms that depend on \( O_d(\delta x, \delta u, \xi_t, \delta t) \). Finally, in the fifth category \( E_5 \), we have all these terms that depend on \( O_d(\delta x, \delta u, \xi_t, \delta t) \) quadratically. The expectation operator will cancel all the terms that include noise up the first order. Moreover, the mean operator for terms that depend on the noise quadratically will result in covariance.

We compute the expectations of all the terms in the \( E_1 \) class. More precisely we will have that:

\[
E(\delta x_t^T A_t^T V_{xx} A_t \delta x_t) = \delta x_t^T A_t^T V_{xx} A_t^T \delta x_t
\]

\[
E(\delta u_t^T B_t^T V_{xx} B_t \delta u_t) = \delta u_t^T B_t^T V_{xx} B_t^T \delta u_t
\]

\[
E(\delta x_t^T A_t^T V_{xx} B_t \delta u_t) = \delta x_t^T A_t^T V_{xx} B_t \delta u_t
\]

\[
E(\delta u_t^T B_t^T V_{xx} A_t \delta x_t) = \delta u_t^T B_t^T V_{xx} A_t \delta x_t
\]

We continue our analysis by calculating all the terms in the class \( E_2 \). More precisely we will have:

\[
E(\xi_t^T \Gamma_t^T V_{xx} A_t \delta x_t) = 0
\]

\[
E(\xi_t^T \Gamma_t^T V_{xx} B_t \delta u_t) = 0
\]

\[
E(\xi_t^T \Gamma_t^T V_{xx} A_t \delta x_t) = 0
\]

\[
E(\xi_t^T \Gamma_t^T V_{xx} B_t \delta u_t) = 0
\]

\[
E(\xi_t^T \Gamma_t^T V_{xx} A_t \delta x_t) = 0
\]

\[
E(\xi_t^T \Gamma_t^T V_{xx} B_t \delta u_t) = 0
\]

The terms above are equal to zero since the brownian noise is zero mean. The expectation of the term that does not depend on \( O_d(\delta x, \delta u, \xi_t, \delta t) \) and it is quadratic with respect to the noise is given as follows:

\[
E(\xi_t^T \Gamma_t^T V_{xx} \Gamma_t \xi_t) = \text{trace} (\Gamma_t^T V_{xx} \Gamma_t \Sigma_o) = \text{trace} \left( \Gamma_t^T V_{xx} \Gamma_t \xi_t \right)
\] (27)

Since matrix \( \Gamma \) depends on variations in states and controls \( \delta x, \delta u \) we can further massage the expressions above so that it can be expressed as quadratic functions in \( \delta x, \delta u \).

\[
\text{trace} (\Gamma_t^T V_{xx} \Gamma_t \Sigma_o) = \sigma_{d, \delta t}^2 \delta t
\] (28)

\[
\text{trace} \left( \left( \begin{array}{cccc}
\Gamma(1)^T & & & \\
& \Gamma(2)^T & & \\
& & \ddots & & \\
& & & \Gamma(\delta t)^T &
\end{array} \right) \right) V_{xx} \left( \begin{array}{cccc}
\Gamma(1) & & & \\
& \Gamma(2) & & \\
& & \ddots & & \\
& & & \Gamma(\delta t)
\end{array} \right)
\] (29)

\[
= \sigma_{d, \delta t}^2 \delta t \sum_{i=1}^{m} \Gamma(i)^T V_{xx} \Gamma(i)
\] (30)

The last equation is written in the form:

\[
\text{trace} (\Gamma_t^T V_{xx} \Gamma_t \Sigma_o) = \delta x^T F \delta x + 2 \delta x^T \tilde{L} \delta u + \delta u^T \tilde{Z} \delta u + 2 \delta u^T \tilde{U} + 2 \delta x^T \tilde{S} + \gamma
\] (31)

Where the terms \( F \in \mathbb{R}^{n \times m}, \tilde{L} \in \mathbb{R}^{m \times p}, \tilde{Z} \in \mathbb{R}^{p \times p}, \tilde{U} \in \mathbb{R}^{m \times 1}, \tilde{S} \in \mathbb{R}^{m \times 1} \) and \( \gamma \in \mathbb{R} \) are defined as follows:

\[
\tilde{F} = \sigma_{d, \delta t}^2 \sum_{i=1}^{m} \nabla x_i^T F_{c}^{(i)} V_{xx} \nabla x_i F_{c}^{(i)}
\] (32)

\[
\tilde{L} = \sigma_{d, \delta t}^2 \sum_{i=1}^{m} \nabla x_i^T F_{c}^{(i)} V_{xx} \nabla u_i F_{c}^{(i)}
\] (33)

\[
\tilde{Z} = \sigma_{d, \delta t}^2 \sum_{i=1}^{m} \nabla u_i^T F_{c}^{(i)} V_{xx} \nabla u_i F_{c}^{(i)}
\] (34)

\[
\tilde{U} = \sigma_{d, \delta t}^2 \sum_{i=1}^{m} \nabla u_i^T F_{c}^{(i)} V_{xx} F_{c}^{(i)}
\] (35)

\[
\tilde{S} = \sigma_{d, \delta t}^2 \sum_{i=1}^{m} \nabla x_i^T F_{c}^{(i)} V_{xx} F_{c}^{(i)}
\] (36)

\[
\gamma = \sigma_{d, \delta t}^2 \sum_{i=1}^{m} F_{c}^{(i)} V_{xx} F_{c}^{(i)}
\] (37)

For those terms that depend both on \( O_d(\delta x, \delta u, \xi_t, \delta t) \) and on the noise class \( E_3 \) we will have:

\[
E(O_d^T V_{xx} \Gamma_t \xi_t) = E \left( \text{trace} \left( V_{xx} \Gamma_t E(\xi_t O_d^T) \right) \right) = \text{trace} \left( V_{xx} \Gamma_t E(\xi_t O_d^T) \right)
\] (38)

By writing the term \( O_d(\delta x, \delta u, \xi_t, \delta t) \) in a matrix form and putting the noise vector insight the this matrix we have:

\[
E(O_d^T V_{xx} \Gamma_t \xi_t) = \text{trace} \left( V_{xx} \Gamma_t E \left[ \begin{array}{cccc}
\xi_t O_{(1)} & & & \\
& \xi_t O_{(2)} & & \\
& & \ddots & & \\
& & & \xi_t O_{(m)}
\end{array} \right] \right)
\] (39)
Calculation of the expectation above requires to find the terms \( E(\sqrt{\delta t}\xi_t O^{(j)}) \) more precisely we will have:

\[
E\left(\sqrt{\delta t}\xi_t O^{(j)}\right) = \frac{1}{2} E\left(\sqrt{\delta t}\xi_t \delta x^T \Phi^{(i)} \delta x\right) + \frac{1}{2} E\left(\sqrt{\delta t}\xi_t \delta u^T \Phi^{(i)} \delta u\right) + E\left(\sqrt{\delta t}\xi_t \delta u^T \Phi^{(i)} \delta x\right)
\]  

(40)

We first calculate the term:

\[
E\left(\sqrt{\delta t}\xi_t \delta x^T \nabla_{xx} \Phi^{(i)} \delta x\right) = \sum_{j=1}^{m} \xi^{(j)} \nabla_{xx} F^{(i)}(\xi^{(j)}) \delta x
\]  

(41)

The term \( E\left(\sqrt{\delta t}\xi_t \delta x^T (\nabla_{xx} f^{(i)}(\delta t) \delta x\right) = 0 \) since it depends linearly on the noise and \( E(\xi_t) = 0 \). The second term \( E\left(\sqrt{\delta t}\xi_t \delta x^T (\nabla_{xx} F^{(i)} \xi_t \sqrt{\delta t} \delta x\right) \) depends quadratically in the noise and thus the expectation operator will result in the variance on the noise. We follow the analysis:

\[
E\left(\sqrt{\delta t}\xi_t \delta x^T \nabla_{xx} \Phi^{(i)} \delta x\right) = \sum_{j=1}^{m} \xi^{(j)} \nabla_{xx} F^{(i)}(\xi^{(j)}) \delta x
\]  

(42)

Since the \( \xi_t = (\xi^{(1)}, ..., \xi^{(m)})^T \) and \( F^{(i)} = (F^{(1)}, ..., F^{(m)}) \) we will have that:

\[
E\left(\sqrt{\delta t}\xi_t \delta x^T \nabla_{xx} \Phi^{(i)} \delta x\right) = \sum_{j=1}^{m} \xi^{(j)} \nabla_{xx} F^{(i)}(\xi^{(j)}) \delta x
\]  

(43)

By writing \( \xi_t \) in vector form we have that:

\[
E\left(\sqrt{\delta t}\xi_t \delta x^T \nabla_{xx} \Phi^{(i)} \delta x\right) = \delta \xi_t \delta x^T \left(\sum_{j=1}^{m} \xi^{(j)} \nabla_{xx} F^{(i)}(\xi^{(j)}) \right) \delta x
\]  

(44)

The term \( \delta x^T \left(\sum_{j=1}^{m} \xi^{(j)} \nabla_{xx} F^{(i)}(\xi^{(j)}) \right) \delta x \) is scalar and it can multiply each one of the elements of the noise vector.
\[ G = \sigma^2 \delta t \left[ \begin{array}{c} \delta u^T \nabla_{uu} F_r^{(1,1)} \delta u \\ \vdots \\ \delta u^T \nabla_{uu} F_r^{(m,1)} \delta u \\ \delta u^T \nabla_{uu} F_r^{(1,n)} \delta u \\ \vdots \\ \delta u^T \nabla_{uu} F_r^{(m,n)} \delta u \end{array} \right] \] (52)

Based on (5) the term \( \Gamma_i \) depends on \( \Delta \) which is a function of the variations in states and controls up to the 1st order. In addition the matrices \( \mathbf{M}, \mathbf{N} \) and \( \mathbf{G} \) are also functions of the deviations in state and controls up to the 2nd order. The product of \( \Delta \) with each one of the matrices \( \mathbf{M}, \mathbf{N} \) and \( \mathbf{G} \) will result into 3rd order terms that can be neglected. By neglecting these terms we can show that:

\[ E \left( O_d^T V_{xx} \Gamma_i \xi_i \right) = \text{trace} \left( V_{xx} \left( \Delta + F \right) \left( \mathbf{M} + \mathbf{N} + \mathbf{G} \right) \right) \]

\[ = \text{trace} \left( V_{xx} F \left( \mathbf{M} + \mathbf{N} + \mathbf{G} \right) \right) \]

Each element \((i,j)\) of the product \( C = V_{xx} F \) can be expressed as \( C^{(i,j)} = \sum_{r=1}^{n} V_{xx}^{(r,j)} F^{(r,i)} \) where \( C \in \mathbb{R}^{n \times p} \). Furthermore the element \((\mu, \nu)\) of the product \( \mathcal{H} = \mathcal{CM} \) is formulated \( \mathcal{H}^{(\mu, \nu)} = \sum_{k=1}^{n} C^{(\mu,k)} M^{(k,\nu)} \) with \( \mathcal{H} \in \mathbb{R}^{n \times n} \). Thus, the term \( \text{trace} \left( V_{xx} F \mathbf{M} \right) \) can be now expressed as:

\[ \text{trace} \left( V_{xx} F \mathbf{M} \right) = \sum_{\ell=1}^{n} \mathcal{H}^{(\ell, \ell)} \]

\[ = \sum_{\ell=1}^{n} \sum_{k=1}^{m} C^{(\ell,k)} \mathcal{M}^{(k,\ell)} \]

\[ = \sum_{\ell=1}^{n} \sum_{k=1}^{m} \left( \sum_{r=1}^{n} V_{xx}^{(k,r)} F^{(r,\ell)} \right) \mathcal{M}^{(k,\ell)} \] (54)

Since \( \mathcal{M}^{(k,\ell)} = \delta \sigma^2 \delta t \delta x^T \nabla_{xx} F^{(k,l)} \delta x \) the vectors \( \delta t \sigma^2 \delta x^T \nabla_{xx} F^{(k,l)} \delta x \) and \( \delta \dot{x} \) do not depend on \( k, \ell, r \) and they can be taken outside the sum. Thus we can show that:

\[ \text{trace} \left( V_{xx} F \mathbf{M} \right) = \sum_{\ell=1}^{n} \sum_{k=1}^{m} \left( \sum_{r=1}^{n} V_{xx}^{(k,r)} F^{(r,\ell)} \right) \sigma^2 \delta t \delta x^T \nabla_{xx} F^{(k,l)} \delta x \]

\[ = \delta x^T \sigma^2 \delta t \sum_{\ell=1}^{n} \sum_{k=1}^{m} \left( \sum_{r=1}^{n} V_{xx}^{(k,r)} F^{(r,\ell)} \right) \nabla_{xx} F^{(k,l)} \delta x \]

\[ = \delta x^T \tilde{\mathbf{M}} \delta x \] (55)

where \( \tilde{\mathbf{M}} \) is a matrix of dimensionality \( \tilde{\mathbf{M}} \in \mathbb{R}^{n \times n} \) and it is defined as:

\[ \tilde{\mathbf{M}} = \sigma^2 \delta t \sum_{\ell=1}^{n} \sum_{k=1}^{m} \left( \sum_{r=1}^{n} V_{xx}^{(k,r)} F^{(r,\ell)} \right) \nabla_{xx} F^{(k,l)} \] (56)

By following the same algebraic steps it can be shown that:

\[ \text{trace} \left( V_{xx} F \mathbf{N} \right) = \delta x^T \tilde{\mathbf{N}} \delta x \] (57)

with \( \tilde{\mathbf{N}} \) matrix of dimensionality \( \tilde{\mathbf{N}} \in \mathbb{R}^{n \times p} \) defined as:

\[ \tilde{\mathbf{N}} = \sigma^2 \delta t \sum_{\ell=1}^{n} \sum_{k=1}^{m} \left( \sum_{r=1}^{n} V_{xx}^{(k,r)} F^{(r,\ell)} \right) \nabla_{xx} F^{(k,l)} \] (58)

and

\[ \text{trace} \left( V_{xx} F \mathbf{G} \right) = \delta u^T \tilde{\mathbf{G}} \delta u \] (59)

with \( \tilde{\mathbf{G}} \) matrix of dimensionality \( \tilde{\mathbf{G}} \in \mathbb{R}^{p \times p} \) defined as:

\[ \tilde{\mathbf{G}} = \sigma^2 \delta t \sum_{\ell=1}^{n} \sum_{k=1}^{m} \left( \sum_{r=1}^{n} V_{xx}^{(k,r)} F^{(r,\ell)} \right) \nabla_{uu} F^{(k,l)} \] (60)

Thus the term \( E \left( O_d^T V_{xx} \Gamma_i \xi_i \right) \) is formulated as:

\[ E \left( O_d^T V_{xx} \Gamma_i \xi_i \right) = \frac{1}{2} \delta x^T \tilde{\mathbf{M}} \delta x + \frac{1}{2} \delta u^T \tilde{\mathbf{G}} \delta u + \delta x^T \tilde{\mathbf{N}} \delta u \] (61)

Similarly we can show that:

\[ E \left( \xi_i^T \Gamma_i^T V_{xx} O_d \right) = \frac{1}{2} \delta x^T \tilde{\mathbf{M}} \delta x + \frac{1}{2} \delta u^T \tilde{\mathbf{G}} \delta u + \delta x^T \tilde{\mathbf{N}} \delta u \] (62)

Next we will find the expectation for all terms that depend on \( O_d (\delta x, \delta u, \omega, \delta t) \) and not on the noise. Consequently, we will have that:

\[ E \left( \delta x^T A_i^T V_{xx} O_d \right) = \delta x^T A_i^T V_{xx} \tilde{O}_d = 0 \]

\[ E \left( \delta u^T B_i^T V_{xx} O_d \right) = \delta u^T B_i^T V_{xx} \tilde{O}_d = 0 \]

\[ E \left( O_d^T V_{xx} A_i \delta x_i \right) = \tilde{O}_d^T V_{xx} A_i \delta x_i = 0 \]

\[ E \left( O_d^T V_{xx} B_i \delta u_i \right) = \tilde{O}_d^T V_{xx} B_i \delta u_i = 0 \] (63)

where the quantity \( \tilde{O}_d \) has been defined in (14). All the 4 terms above are equal to zero since they have variations in state and control of the order higher than 2 and therefore they can be neglected.

Finally we compute the terms of the 5th class and therefore we have the expression

\[ \mathcal{E}_5 = E \left( O_d^T V_{xx} O_d \right) \]

\[ = E \left( \text{trace} \left( V_{xx} O_d O_d^T \right) \right) \]

\[ = \text{trace} \left( V_{xx} E \left( O_d O_d^T \right) \right) \]

\[ = \text{trace} \left( V_{xx} E \left( \begin{array}{c} O^{(1)} \\ \vdots \\ O^{(n)} \end{array} \right) \left( \begin{array}{c} O^{(1)} \\ \vdots \\ O^{(n)} \end{array} \right)^T \right) \] (64)

The product \( O^{(i)} O^{(j)} \) is a function of variation in state and control of order 4 since each term \( O^{(i)} \) is a function of
variation in states and control of order 2. Consequently, the term \( E^5 = E \left( O_d^T V_{xx} O_d \right) \) is equal to zero.

With the computation of the expectation of term that is quadratic WRT \( O_d \) we have calculated all the terms of the second order expansion of the cost to go function. In the next section we derive the optimal controls and we present the SDDP algorithm. Furthermore we show how SDDP recover the deterministic solution as well as the cases of only control multiplicative, only state multiplicative and only additive noise.

IV. OPTIMAL CONTROLS

In this section we provide the form of the optimal controls and we show how previous results are special cases of our generalized stochastic DDP formulation. Furthermore after we computed all the terms of the expansion of the cost to go function \( V(x_t) \) at state \( x_t \) we show that its form remains quadratic WRT variations in state \( \delta x_t \) under the constraint of the nonlinear stochastic dynamics in (2). More precisely we have that:

\[
\begin{align*}
V(x_{t+\delta t} + \delta x_{t+\delta t}) &= V(x_{t+\delta t}) + \nabla_x V_{xx} A_t \delta x_t + \nabla_x V_{uu} B_t \delta u_t + \\
&+ \frac{1}{2} \delta x^T \mathcal{F} \delta x + \frac{1}{2} \delta u^T \mathcal{Z} \delta u + \delta u^T \mathcal{L} \delta x \\
&+ \frac{1}{2} \delta x^T A_t^T V_{xx} A_t \delta x_t + \frac{1}{2} \delta u^T B_t^T V_{xx} B_t \delta u_t + \\
&+ \frac{1}{2} \delta x^T A_t^T V_{xx} B_t \delta u_t + \frac{1}{2} \delta u^T B_t^T V_{xx} A_t \delta x_t + \\
&+ \frac{1}{2} \delta x^T \mathcal{F} \delta x + \frac{1}{2} \delta x^T \tilde{\mathcal{L}} \delta x + \frac{1}{2} \delta u^T \tilde{\mathcal{Z}} \delta u + \\
&+ \frac{1}{2} \delta x^T \tilde{\mathcal{M}} \delta x + \frac{1}{2} \delta u^T \tilde{\mathcal{G}} \delta u + \delta x^T \tilde{\mathcal{N}} \delta u
\end{align*}
\]

The unmaximized state action value function is defined as follows:

\[
Q(x_t, u_t) = \ell(x_t, u_t) + V(x_{t+1})
\]

Given a trajectory in states and controls \( \bar{x}, \bar{u} \) we can approximate the state action value function as follows:

\[
Q(\bar{x} + \delta x, \bar{u} + \delta u) = Q_0 + \delta u^T Q_u + \delta x^T Q_x
\]

\[
\frac{1}{2} \begin{bmatrix} \delta x^T & \delta u^T \end{bmatrix} \begin{bmatrix} Q_{xx} & Q_{xu} \\ Q_{ux} & Q_{uu} \end{bmatrix} \begin{bmatrix} \delta x \\ \delta u \end{bmatrix}
\]

By equating the coefficients with similar powers between the state action value function \( Q(x_t, u_t) \) and the immediate reward and cost to go \( \ell(x_t, u_t) \) and \( V(x_{t+1}) \) respectively we can show that:

\[
\begin{align*}
Q_x &= \ell_x + A_t V_x + \tilde{\mathcal{S}} \\
Q_u &= \ell_u + A_t V_x + \tilde{\mathcal{U}} \\
Q_{xx} &= \ell_{xx} + A_t^2 V_{xx} A_t + \mathcal{F} + \tilde{\mathcal{F}} + \tilde{\mathcal{M}} \\
Q_{xu} &= \ell_{xu} + A_t^2 V_{xu} B_t + \mathcal{L} + \tilde{\mathcal{G}} + \tilde{\mathcal{N}} \\
Q_{uu} &= \ell_{uu} + B_t^2 V_{uu} B_t + \mathcal{Z} + \tilde{\mathcal{G}} + \tilde{\mathcal{Z}}
\end{align*}
\]

where we have assume a local quadratic approximation of the immediate reward \( \ell(x_t, u_t) \) according to the equation:

\[
\ell(\bar{x} + \delta x, \bar{u} + \delta u) = \ell_0 + \delta u^T \ell_u + \delta x^T \ell_x + \frac{1}{2} \delta x^T \delta u
\]

\[
\left[ \begin{bmatrix} \ell_x \\ \ell_u \end{bmatrix} \right] = \frac{1}{2} \begin{bmatrix} \ell_{xx} & \ell_{xu} \\ \ell_{ux} & \ell_{uu} \end{bmatrix} \begin{bmatrix} \delta x \\ \delta u \end{bmatrix}
\]

with \( \ell_x = \frac{\partial \ell}{\partial x}, \ell_u = \frac{\partial \ell}{\partial u}, \ell_{xx} = \frac{\partial^2 \ell}{\partial x^2}, \ell_{uu} = \frac{\partial^2 \ell}{\partial u^2} \) and \( \ell_{ux} = \frac{\partial^2 \ell}{\partial x \partial u} \). The local variations in control \( \delta u^* \) that maximize the state action value function are expressed by the equation that follows:

\[
\delta u^* = \arg \max_u \left( Q(x + \delta x, \bar{u} + \delta u) \right)
\]

\[
= -Q_{uu}^{-1} (Q_u + Q_{xu} \delta x)
\]

The optimal control variations have the form \( \delta u^* = 1 + L \delta x \) where \( L = -Q_{uu}^{-1} Q_u \) is the open loop gain and \( L = -Q_{uu}^{-1} Q_{xu} \) is the closed loop - feedback gain. The SDDP algorithm is provided in a pseudocode form on Table I.

<table>
<thead>
<tr>
<th>TABLE I</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSEUDOCODE OF THE SDDP ALGORITHM</td>
</tr>
</tbody>
</table>

- **Given:**
  - An immediate cost function \( \ell(x, u) \)
  - A terminal cost term \( \phi_N \)
  - The stochastic dynamics \( dx = f(x, u)dt + F(x, u)dw \)

- **Repeat** until convergence:
  - Given a trajectory in states and controls \( \bar{x}, \bar{u} \) we find the quadratic approximations of the stochastic dynamics \( A_t, B_t, \Gamma_t, O_t \) and the quadratic approximation of the immediate cost function \( \ell_0, \ell_x, \ell_{xx}, \ell_{uu}, \ell_{ux} \) around these trajectories.
  - Compute all the terms \( Q_x, Q_u, Q_{xu} \) and \( Q_{uu} \) according to equation (68).
  - Back-propagate the quadratic approximation of the value function based on the equations:
    - \( V_0^{(k+1)} = V_0^{(k+1)} - Q_u V_{uu}^{-1} Q_u \)
    - \( V_x^{(k+1)} = Q_u V_{uu}^{-1} Q_{xu} \)
    - \( V_{xx}^{(k+1)} = Q_{uu} - Q_x V_{uu}^{-1} Q_{xu} \)
  - Compute \( \delta u^* = -Q_{uu}^{-1} (Q_u + Q_{xu} \delta x) \)
  - Update controls \( x^* = u^* + \delta u^* \)
  - Get the new optimal trajectory \( x^* \) by propagating the nonlinear dynamics \( dx = f(x, u^*)dt + F(x, u^*)dw \).
  - Set \( \bar{x} = x^* \) and \( \bar{u} = u^* \) and repeat.
state depended noise and therefore $x$ is to bring the state $\alpha$ and $A$.

The time horizon the control is activated and the state reaches the target state. In figure 1, the left plot illustrates the state trajectory as it approaches the zero and thus there is only control cost. In figure 2, the right plot illustrates the convergence behavior of SDDP. An important observation is that curved trajectories correspond to cases with high variance noise. Furthermore, as the variance of the noise decreases the optimal trajectories become more and more straight.

VI. DISCUSSION

In this paper we explicitly derived the equations describing the second order expansion of the cost-to-go, given state and control dependent noise. Our main result is that the expressions remain quadratic wrt $\delta x$ and $\delta u$, so the basic structure of the algorithm, a quadratic cost-to-go approximation with a linear policy, remains unchanged. In addition we have shown how the cases of deterministic and stochastic DDP with additive noise are sub-cases of our generalized formulation of Stochastic DDP.

Current and future research includes further testing and evaluation of our generalized Stochastic DDP algorithm on multidimensional stochastic systems which are highly nonlinear and have noise that is control and state dependent. Biomechanical models belong in this class due to highly nonlinear and noisy nature of muscle dynamics. Moreover we are aiming to incorporate a second order Extended Kalman Filter that can handle observation as well as process noise. The resulting optimal controller-estimator scheme will be a version of iterative Quadratic Gaussian regulator that can handle stochastic dynamics expanded up to the second order with state and control dependent noise.

REFERENCES


