# Entropy Regularized Variational Dynamic Programming for Stochastic Optimal Control

Marc Lambert, Francis Bach, Silvère Bonnabel

# Entropy Regularized Variational Dynamic Programming for Stochastic Optimal Control

Marc Lambert
INRIA - Ecole Normale Supérieure - PSL Research university
DGA - French Procurement Agency
marc.lambert@inria.fr

Francis Bach
INRIA - Ecole Normale Supérieure - PSL Research university
francis.bach@inria.fr

Silvère Bonnabel
MINES ParisTech, PSL University, Center for robotics
silvere.bonnabel@mines-paristech.fr

**Abstract**

This paper addresses stochastic optimal control, where the state-feedback control policies are probability distributions, with an additional entropy penalty on the control policy. We interpret the cost function as a Kullback-Leibler (KL) divergence between two joint distributions, enabling the use of variational inference techniques. This approach leads to a dynamic programming principle, also defined in terms of KL divergence. In the linear case, we show the entropic penalty leads to Gaussian control policies, whose mean coincides with the linear quadratic regulator (LQR). Furthermore, when the state is not directly measured, we may prove a separation principle along the lines of linear quadratic Gaussian (LQG) control. In the case of nonlinear control systems being linear in the control inputs, with quadratic costs, we utilize our variational framework to approximate the optimal solution with Gaussian distributions. This yields closed-form recursive updates that extend traditional LQR control. We demonstrate the effectiveness of this new method in simulations.

## 1 Introduction

In this article, we consider a stochastic dynamical system governed in discrete time by a known Markovian transition $p(x_{k+1}|x_k, u_k)$ where $x_k \in \mathbb{R}^d$ is the current state and $u_k \in \mathbb{R}^m$ is the control variable with $m \leq d$. The initial state $x_0$ is supposed to be known. To precisely state our positioning and our contributions, we need to introduce from now on the notation and mathematical basics. For the expectations $\int p(x)f(x)dx$, we use the notation $\mathbb{E}[f(x)]$ or $\mathbb{E}_{p(x)}[f(x)]$.

### 1.1 Basics of Stochastic Optimal Control

To control future states starting from $x_0$ and over a finite horizon $K$, we first consider the stochastic finite horizon optimal control problem in discrete time:

$$\min_{u_0,\dots,u_{K-1}} \mathbb{E}\Big[\sum_{k=0}^{K-1} \ell_k(x_k, u_k) + L_K(x_K)\Big], \tag{1}$$

where the expectation is taken under the stochastic trajectories starting from $x_0$; $\ell_k$ denote the cost functions for each step $0 \leq k \leq K-1$ and $L_K$ the final cost function. These functions are supposed to be continuous.

In this context, the goal is typically to derive causal state-feedback control policies $u_k = \varphi_k(x_0, \dots, x_{k-1})$ so as to solve (1). A key result in that regard is that of dynamic programming, which states that one may define a value function $V_K$ based on the final cost $V_K(x_K) := L_K(x_K)$, and then define $V_k$ through the backward recursion:

$$V_k(x_k) = \min_v \ \ell_k(x_k, v) + \mathbb{E}_{p(x_{k+1}|x_k,v)}\Big[V_{k+1}(x_{k+1})\Big]. \tag{2}$$

$V_k$ is a function defined over the entire state space, termed "cost-to-go," also called value function, that encapsulates the minimum cost when (deterministically) starting from $x_k$. This yields an optimal causal state-feedback policy

$$\varphi^*(x_k) = \operatorname*{argmin}_v \quad \ell_k(x_k, v) + \mathbb{E}_{p(x_{k+1}|x_k,v)}\Big[V_{k+1}(x_{k+1})\Big],$$

defining a sequence of control inputs that minimize (1).

### 1.2 Probabilistic State-Feedback Policy

A somewhat different problem arises when the control policy is taken as a *probability distribution* (a density) of the form $p(u_k|x_k)$ instead of $u_k = \varphi_k(x_k)$. Letting $z_{0:K} := (u_0, x_1, u_1, \dots, x_{K-1}, u_{K-1}, x_K)$, its density then decomposes using the Markov property as follows:

$$p(z_{0:K}|x_0) = \prod_{k=0}^{K-1} p(u_k|x_k)p(x_{k+1}|x_k, u_k). \tag{3}$$

As is common in probabilistic graphical models, we will overload notation by letting letter $p$ denote all probability densities. The associated graphical model is shown in Figure 1.

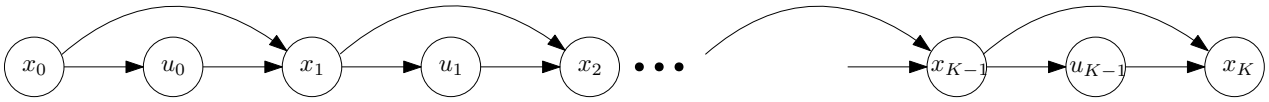


Figure 1: Graphical model associated with (3), where $u_k|x_k$ is a probability distribution.

This turns (1) into the alternative control problem

$$\min_{p(u_k|x_k), 0\leq k\leq K-1} \mathbb{E}_{p(z_{0:K}|x_0)}\Big[\sum_{k=0}^{K-1} \ell_k(x_k, u_k) + L_K(x_K)\Big]. \tag{4}$$

As is, the argmin over policies $p(u_k|x_k)$ consists of Dirac distributions $\varphi^*(x_k)$, and one recovers the optimal deterministic state-feedback policy for (1). To obtain a random policy, one can add to the cost (4) a penalty of magnitude $\varepsilon$ on the negentropy of the policy function $\int p(u_k|x_k) \log p(u_k|x_k) du_k$, as proposed in [1, 2], leading to the following regularized problem:

$$\min_{p(u_k|x_k), 0\leq k\leq K-1} \mathbb{E}_{p(z_{0:K}|x_0)}\Big[\sum_{k=0}^{K-1} \Big(\ell_k(x_k, u_k) + \varepsilon \log p(u_k|x_k)\Big) + L_K(x_K)\Big], \tag{5}$$

where $\varepsilon$ is a "temperature" parameter. Note that, as $\varepsilon \to 0$, we recover the deterministic policy (1).

It turns out the dynamic programming principle carries over to the problem above, see [3, 4]. Starting from $V_K^{(c)}(x_K) := L_K(x_K)$, we may define a cost-to-go through the backward recursion:

$$V_k^{(c)}(x_k) \quad := \quad \min_{p(u_k|x_k)} \mathbb{E}_{p(u_k|x_k)p(x_{k+1}|x_k,u_k)}\Big[\ell_k(x_k,u_k) + \varepsilon \log p(u_k|x_k) + V_{k+1}^{(c)}(x_{k+1})\Big], \tag{6}$$

where superscript $(c)$ stands for conditional entropy regularization. An appealing aspect of this conditional entropy regularization is that the policy can be expressed exactly [3]:

$$p^*(u_k|x_k) \propto \exp\Big[-\frac{1}{\varepsilon}\big(\ell_k(x_k,u_k) + \mathbb{E}[V_{k+1}^{(c)}(x_{k+1})]\big)\Big]. \tag{7}$$

### 1.3 Considered Problem

Instead of penalizing the negentropy of the state-feedback policy, as done in (5), we propose to consider the classical stochastic optimal control problem (SOC) (1) with an additional penalty $+\varepsilon \log p(z_{0:K}|x_0)$, that is, penalizing the negentropy of the *full* joint distribution $p(z_{0:K}|x_0)$. Given the decomposition (3), this means we consider the problem:

$$\min_{p(u_k|x_k), 0\le k\le K-1} \mathbb{E}_{p(z_{0:K}|x_0)}\Big[\sum_{k=0}^{K-1}\Big(\ell_k(x_k,u_k) + \varepsilon \log p(u_k|x_k) + \varepsilon \log p(x_{k+1}|x_k,u_k)\Big) + L_K(x_K)\Big]. \tag{8}$$

Starting from $V_K^{(f)}(x_K) := L_K(x_K)$, we now define a cost-to-go through the backward recursion

$$V_k^{(f)}(x_k) = \min_{p(u_k|x_k)} \mathbb{E}_{p(u_k|x_k)p(x_{k+1}|x_k,u_k)}\Big[\ell_k(x_k,u_k) + \varepsilon \log p(u_k|x_k) + \varepsilon \log p(x_{k+1}|x_k,u_k) + V_{k+1}^{(f)}(x_{k+1})\Big], \tag{9}$$

where superscript $(f)$ stands for full entropy regularization. (9) consists of (6) plus the term $\varepsilon \log p(x_{k+1}|x_k,u_k)$. Letting $H(p(x)) := -\int p(x)\log p(x)dx$ denote the entropy, the latter entropy-regularized SOC problem rewrites as

$$V_k^{(f)}(x_k) = \min_{p(u_k|x_k)}\Big\{-\varepsilon H(p(u_k|x_k)p(x_{k+1}|x_k,u_k)) + \mathbb{E}_{p(u_k|x_k)p(x_{k+1}|x_k,u_k)}\Big[\ell_k(x_k,u_k) + V_{k+1}^{(f)}(x_{k+1})\Big]\Big\}.$$

This variational dynamic programming principle minimizes indeed the total loss (8) which is exactly $V_0^{(f)}(x_0)$. Moreover, the problem (8) can be recast as a Kullback-Leibler (KL) divergence between two joint distributions. The *exact* optimal solution to this problem can be analytically expressed using variational inference methods [5], as we will show, thereby generalizing Equation (7).

We briefly discuss the rationale behind this joint entropic regularization. First, penalizing the negentropy of $p(u_k|x_k)$ fosters distributions of greater entropy, that is, "flatter" distributions. On the one hand, this results in softer controls (as opposed to, typically, bang-bang control). On the other hand, encouraging dispersion results in more unpredictable control policies, that remain sensible though, as they minimize a cost function. Our formulation moreover penalizes the entropy of the states through the term $\varepsilon \log p(x_{k+1}|x_k,u_k)$, leading to further exploration of the state space. The formulation is compatible with stochastic control or model-based reinforcement learning when we have access to $p(x_{k+1}|x_k,u_k)$.

### 1.4 Variational Approximation of the Optimal Control

In practice, determining the exact optimal policy for either conditional negentropy penalization (5) or full negentropy penalization (8) is generally difficult, as unnormalized formulas such as (7) make the computation of simple features such as the mean and covariance typically intractable. Hence, it is desirable to approximate the optimal policy $p^*(u_k|x_k)$ by a parametric family of distributions $q(u_k|x_k) \in \mathcal{P}_k^{(u)}$. Using the Kullback-Leibler (KL) divergence as a measure of discrepancy, this may be done by considering the variational problem:

$$\underset{q(u_k|x_k)\in\mathcal{P}_k^{(u)}}{\arg\min} \quad \mathrm{KL}\big(q(u_k|x_k)\|p^*(u_k|x_k)\big) \tag{10}$$

where "KL" denotes the unnormalized KL divergence defined by $\mathrm{KL}(q(y)\|p(y)) := \int q(y)\log q(y)dy - \int q(y)\log p(y)dy$. For the conditional problem (6), this was done in [2].

Since the optimal policy depends on the cost-to-go, the latter needs also to be estimated, which is difficult as it is a function over the state space. This led to the soft-actor critic algorithm [2] where the policy and cost-to-go are updated alternately. However, back to our full entropic problem, we propose instead to stick with the KL-based approximation framework, and to seek an approximation of $V_k(x_k)$, the cost-to-go (9), in the form $-\varepsilon\log\phi_k(x)$. In this way, the approximating function $\phi_k$ can be viewed as a unnormalized probability density, meant to approximate the positive function $\exp(-V_k(x_k)/\varepsilon)$. We may then tackle this approximation problem by minimizing the following discrepancy:

$$\min_{\phi_k(x_k)\in\mathcal{P}_k^{(x)}} \quad \varepsilon\mathrm{KL}\big(\phi_k(x_k)\|\exp(-V_k^{(f)}(x_k)/\varepsilon)\big), \tag{11}$$

with $\mathcal{P}_k^{(x)}$ an approximating family of distributions, e.g., Gaussians. We are then left with two families of approximating distributions, $q(u_k|x_k)$ and $\phi_k(x_k)$, at each stage.

In this paper, we consider a joint Gaussian distribution for $q(u_k, x_k) = q(u_k|x_k)\phi_k(x_k)$ as the approximating family, and derive recursive updates for its parameters. In particular, we show the precision matrix of the Gaussian marginal $\phi_k(x_k)$ follows an implicit backward equation that generalizes the Riccati equation from LQR control.

### 1.5 Related Works and Motivations

#### On Risk Sensitivity Control

Stochastic optimal control was first reformulated with a "log-sum-exp" variant of the Bellman recursion under the framework of risk-sensitive control [6, 7]. This line of work has led to a rich literature in robust control and differential games [8–12]. Whittle revisited the linear quadratic Gaussian controller in this risk-averse setting, introducing a min-max problem that aims to solve the estimation and control problems jointly [13]. This controller has been extended to the nonlinear case [14–16]. All this literature can be seen as the "dual" version of the max-entropy approach: instead of relaxing the problem with entropy, the penalty cost is enhanced with a "log-sum-exp" function.

#### On Path Integral Formulation

A path integral formulation to solve the stochastic Hamilton-Jacobi equation was proposed in [17] based on the Feynman-Kac formula. This formulation has been related to the "log-sum-exp"

Bellman recursion and used for risk-averse or risk-seeking control [18]. These ideas have been further developed and have led to efficient model predictive control algorithms known as MPPI [19]. While close to our approach, this family of algorithms doesn't explicitly search for the optimal random policy but implicitly finds it by sampling from the uncontrolled dynamics.

**On Maximum Entropy Policy**

Random policies were proposed in the context of Markov Decision Processes in [20, 21], where optimal control problems were recast as probabilistic inference problems. The maximum entropy principle for control was introduced in inverse reinforcement learning [1] to learn a policy from observations, and in reinforcement learning to enhance exploration through the soft actor-critic algorithm [2], as discussed above. Gaussian approximation of the optimal maximum entropy policy (7) is considered in the context of differential dynamic programming [4, 22]. Note that, to this aim the dynamics are linearized, and the cost function is approximated locally with a quadratic cost. Our variational approach avoids these linearizations.

**On the KL formulation of Stochastic Optimal Control**

Our variational formulation (8), (15) differs also from previous KL formulations proposed in stochastic control. In "KL control" [23] the relative entropy is introduced to penalize a discrepancy between the controlled dynamic $p(x_{k+1}|x_k, u_k)$ and the passive one $p(x_{k+1}|x_k, u_k = 0)$, and thus serves as an indirect penalty on input usage. A KL cost setting was also proposed as an extension of the Schrödinger bridge problem for stochastic control, see [24]. All these approaches do not use the regularization with the entropy of the policy and do not provide a random policy. A KL formulation close to (8), (15) with a random policy was proposed in [25] and related to a cost-to-go regularized with the entropy of the policy. In this setting, the rewards are viewed as observations, and the optimal policy is computed via variational inference. The case of control-affine inputs is discussed, but no closed-form updates have been derived. In [26], a KL divergence between two joint distributions—representing the learned dynamic and a reference dynamic—was proposed to design control policies from demonstrations, providing an explicit solution for the optimal policy. However, the reference dynamic was not explicitly linked to any cost function.

## 1.6 Main Contributions and Paper Organization

We aim to address problem (8), or equivalently (9), using the variational inference (VI) framework to derive both theoretical and practical approximate solutions. The organization and contributions of the remainder of this paper are as follows:

- In Section 2, we turn (9) into variational dynamic programming, by re-writing it as a Kullback-Leibler (KL) minimization problem. We give the exact formulation of the optimal policy and cost-to-go. We then show how one may approximate *jointly* the policy and the value function, with a joint distribution $q(x_k, u_k)$ lying in a "nice" approximating family. Letting this family be that of Gaussian distributions, solving the variational joint problem enables in turn closed-form formulas for the approximate policy $q(u_k|x_k)$ and approximate value function $q(x_k)$. In the remainder, we highlight two cases where this is feasible.

- In Section 3, we examine the case of a linear dynamical system with a quadratic cost function and entropic penality. We demonstrate that the exact optimal policy is then a Gaussian whose mean coincides with the linear quadratic regulator (LQR) controller, and whose dispersion's

magnitude is the temperature parameter $\varepsilon$. Furthermore, when only partial linear and noisy measurements of the state are available, we can prove an exact separation principle. This results in a novel maximum-entropic linear quadratic Gaussian (LQG) controller.

- In Section 4, we consider the particular case of nonlinear dynamics of the form $x_{k+1} = f(x_k) + Bu_k + \nu_k$, $\nu_k \sim \mathcal{N}(0, C)$, along with a quadratic cost function, and we show we can derive explicit formulas for the optimal parameters of the Gaussian $q(x_k, u_k)$, leading to novel variational backward Riccati equations.
- In Section 5, we compare the obtained policy and linearized LQR for the stabilization of a noisy nonlinear inverted pendulum around an equilibrium point and show how our policy increases entropy, while stabilizing. We also apply the policy to a nonlinear Dubins car, to make it follow a path while being unpredictable.

The present article extends preliminary ideas from the conference paper [27] by adding: 1) A new proof in the linear case for the Max-Entropy LQR policy, along with the computation of the probabilistic envelopes (covariance) for this policy. 2) The extension to a Max-Entropy LQG controller when the state is partially observable, including a proof of a separation principle in the Max-Entropy case. 3) Novel simulations and applications.

# 2 Variational Dynamic Programming

In this section, we introduce our variational framework to tackle the full-entropic SOC problem (9).

## 2.1 A Variational Dynamic Programming Principle

Following [28], [29], we can rewrite the cost as follows: $\ell_k(x_k, u_k) = -\varepsilon \log r(x_k, u_k)$, and $L_K(x_K) = -\varepsilon \log r(x_K)$ where $\varepsilon > 0$ is the same temperature parameter introduced in Equation (5). Here, $r(x_k, u_k)$ and $r(x_K)$ can be interpreted as reward distributions, taking the form of an unnormalized Gibbs distributions.

The dynamic programming recursion (9) can then be translated into a KL minimization problem. Using the Gibbs formulation for the loss $\ell_k(x_k, u_k) = -\varepsilon \log r(x_k, u_k)$, (9) rewrites:

$$V_k^{(f)}(x_k) = \min_{p(u_k|x_k)} \varepsilon \mathrm{KL}(p(x_{k+1}|u_k, x_k)p(u_k|x_k) \| r(x_k, u_k)\phi(x_{k+1})), \tag{12}$$

where we let:

$$\phi(x_{k+1}) := \exp\left(-V_{k+1}^{(f)}(x_{k+1})/\varepsilon\right). \tag{13}$$

This is immediately proved by writing $\mathbb{E}_p(\varepsilon \log p + (\ell_k + V_{k+1})) = \varepsilon \mathrm{KL}\big(p \| \exp(-\ell_k/\varepsilon) \exp(-V_{k+1}/\varepsilon)\big)$.

One may also associate an unnormalized joint distribution with the cost function through the following factorization:

$$r(z_{0:K}|x_0) = \Big( \prod_{k=0}^{K-1} r(x_k, u_k) \Big) r(x_K). \tag{14}$$

Problem (8), which we address, then rewrites:

$$\min_{p(u_k|x_k), 0 \le k \le K-1} \varepsilon \mathrm{KL}(p(z_{0:K}|x_0) \| r(z_{0:K}|x_0)). \tag{15}$$

The dynamic programming problem (13), hence (15), may be solved exactly using properties of KL divergences.

### 2.2 The "Exact" Optimal Policy

We now give our first main result, which generalizes (7). The proof is differed to Appendix 8.

**Proposition 1.** *The solution to Problem* (12) *is given by:*

$$p^*(u_k|x_k) = \frac{1}{\phi(x_k)} \exp\big(-Q_k^f(u_k, x_k)\big)$$
$$Q_k^f(u_k, x_k) := \mathrm{KL}(p(x_{k+1}|u_k, x_k) \| r(u_k, x_k)\phi(x_{k+1})).$$

*The optimal cost-to-go* $V_k^{(f)}(x_k)$ *depends on* $\phi(x_k)$*, the partition function of* $p^*(u_k|x_k)$ *as follows:*

$$\begin{aligned} V_k^{(f)}(x_k) &= -\varepsilon \log \phi(x_k) \\ &= -\varepsilon \log \int \exp\big(-Q_k^f(u_k, x_k)\big) du_k, \end{aligned} \tag{16}$$

*such that* $V_k^{(f)}(x_k)$ *takes a "log-sum-exp" form.* ■

The optimal policy derived here aligns with the solution (7) proposed earlier in [3] when the entropy of the transition kernel $p(x_{k+1}|x_k, u_k)$ is a constant, which is typically the case for additive Gaussian state-independent noise. Our proposition hence generalizes this result when this is not the case. Akin to Bayesian inference, such formulas prove intractable when one wants to compute even simple statistics of the randomized control law, such as its mean. Thus, we now introduce an alternative variational approximation of the exact solution, as a joint distribution on the control policy and cost-to-go.

### 2.3 Variational Approximation

As previously explained, the cost-to-go is a general function, and hence needs to be approximated using a family $\phi(x_k) \in \mathcal{P}_k^{(x)}$, see (11). To obtain a joint approximation formulation, we start indeed from the optimal approximating value distribution $\phi(x_k)$ as the solution to a variational problem:

$$\min_{\phi(x_k) \in \mathcal{P}_k^{(x)}} \varepsilon \mathrm{KL}(\phi(x_k) \| \exp(-V_k^{(f)}(x_k)/\varepsilon)) \tag{17}$$

$$:= \min_{\phi(x_k) \in \mathcal{P}_k^{(x)}} \varepsilon \int \phi(x_k) \log(\phi(x_k)) dx_k + \int \phi(x_k) V_k^{(f)}(x_k) dx_k. \tag{18}$$

Now, we can substitute (9) into the latter. Let $H(p(x)) := -\int p(x) \log p(x) dx$ denote the entropy. (18) is equal to

$$\min_{\phi(x_k) \in \mathcal{P}_k^{(x)}} \big[ -\varepsilon H(\phi(x_k)) + \int \phi(x_k) \big[ \tag{19}$$

$$\min_{p(u_k|x_k) \in \mathcal{P}_k^{(u)}} -\varepsilon H(p(u_k|x_k)p(x_{k+1}|x_k, u_k)) + \int \big(\ell_k(x_k, u_k) + V_{k+1}^{(f)}(x_{k+1})\big) p(u_k|x_k) p(x_{k+1}|x_k, u_k) dx_{k+1} du_k \big] dx_k \big]$$

$$\leq \min_{\phi(x_k) \in \mathcal{P}_k^{(x)}} \min_{p(u_k|x_k) \in \mathcal{P}_k^{(u)}} \big[ -\varepsilon H(\phi(x_k)) + \int \phi(x_k) \big[ \tag{20}$$

$$-\varepsilon H(p(u_k|x_k)p(x_{k+1}|x_k, u_k)) + \int \big(\ell_k(x_k, u_k) + V_{k+1}^{(f)}(x_{k+1})\big) p(u_k|x_k) p(x_{k+1}|x_k, u_k) dx_{k+1} du_k \big] dx_k \big], \tag{21}$$

where the inequality arises when interchanging integral and the second min, and it would be an equality if the approximation were exact (i.e., if the model were fully expressive).

Using $H(\phi(x_k))+\int \phi(x_k)H(p(u_k|x_k)p(x_{k+1}|x_k,u_k))dx_k = H(\phi(x_k)p(u_k|x_k)p(x_{k+1}|x_k,u_k))$, which comes from the fact that $p(u_k|x_k)$ and $\phi(x_{k+1}|x_k,u_k)$ are normalized, the expression (21) writes:

$$\min_{\phi(x_k)\in\mathcal{P}_k^{(x)},\ p(u_k|x_k)\in\mathcal{P}_k^{(u)}} \big[ -\varepsilon H(\phi(x_k)p(u_k|x_k)p(x_{k+1}|x_k,u_k)) \\ +\int \phi(x_k)\big(\ell_k(x_k,u_k)+V_{k+1}^{(f)}(x_{k+1})\big)p(u_k|x_k)p(x_{k+1}|x_k,u_k)dx_{k+1}du_kdx_k\big].$$

Starting from the value function (cost-to-go) approximation problem (17), we thus end up with the following variational approximation problem

$$\min_{\phi(x_k)\in\mathcal{P}_k^{(x)},\ p(u_k|x_k)\in\mathcal{P}_k^{(u)}} \varepsilon\mathrm{KL}(\underbrace{\phi(x_k)p(u_k|x_k)}_{\text{joint}}\, p(x_{k+1}|x_k,u_k)\|r(u_k,x_k)\phi(x_{k+1})), \tag{22}$$

We now have a fully defined Bellman-like recursion to attack the entropy-regularized problem (15), based on approximating probability density families. One contribution of the present paper is to have defined two distinct variational approximation problems, and proved that the minima combine nicely so as to obtain a joint approximation problem.

### 2.4 Joint Gaussian Variational Approximation

To keep (22) tractable, one possibility, that we advocate in the present paper, is to restrict the family of approximating distributions, by assuming the joint distribution $\phi(x_k)p(u_k|x_k)$ to directly belong to an approximating family $\mathcal{P}_k^{(x,u)}$.

To distinguish the problem from the previous one, and to insist on the joint distribution approach, we change notation. By once more overloading $q$, we replace $\phi(x)$ with $q(x)$, and $p(u|x)$ with $q(u|x)$. The dynamic programming problem (22) on the joint distribution then becomes at each step:

$$\min_{q(x_k)q(u_k|x_k)\in\mathcal{P}_k^{(x,u)}} \varepsilon\mathrm{KL}(q(x_k)q(u_k|x_k)p(x_{k+1}|x_k,u_k)\|r(u_k,x_k)q(x_{k+1})), \tag{23}$$

where $\phi(x_{k+1})$ has been approximated in the previous step by $q(x_{k+1})$. This joint formulation is quite practical when $q(u_k|x_k)q(x_k)$ ends up being in a simple family, so that we are faced with the usual variational approximation consisting of a (left) KL minimization of a function of $(x_k,u_k)$.

Although various approximating families can be leveraged, the simplest is arguably to use a joint Gaussian distribution $q(x_k,u_k)=q(u_k|x_k)q(x_k)$. We then intend to learn its parameters based on the obtained Bellman-like equations and immediately benefit from Gaussian conditioning formulas to recover $q(x_k)$ and $q(u_k|x_k)$. We parametrize this joint Gaussian as

$$q(x_k,u_k) := \mathcal{N}\big(\mu_k,\Sigma_k\big) \tag{24}$$
$$= \mathcal{N}\left(\begin{pmatrix}\alpha_k\\ \beta_k\end{pmatrix}, \varepsilon\begin{pmatrix} P_k^{-1} & P_k^{-1}K_k^\top \\ K_kP_k^{-1} & K_kP_k^{-1}K_k^\top + S_k^{-1}\end{pmatrix}\right)$$
$$q(x_k) = \mathcal{N}(\alpha_k,\varepsilon P_k^{-1}) \tag{25}$$
$$q(u_k|x_k) = \mathcal{N}(\beta_k+K_k(x_k-\alpha_k),\varepsilon S_k^{-1}). \tag{26}$$

The rationale for using a joint Gaussian distribution is as follows. With this choice, the feedback policy $q(u_k|x_k)$ takes the form of a distribution dispersed around its mean. More interestingly, the

fact that $q(x_k)$ be Gaussian means we use a quadratic approximation for the value function, as $q(x_k)$ is an approximation of $\phi(x_k)$. Moreover, this choice of joint distribution enforces a linear feedback $\beta_k + K_k(x_k - \alpha_k)$ in terms of the mean of $q(u_k|x_k)$. Last but not least, we will show in the linear case that this Gaussian setting actually exactly solves the initial problem, and allows for a novel separation principle.

### 2.5 Summary and Outlook

Starting from the stochastic optimal control problem (8), we defined a dynamic programming recursion (9). We then used the KL divergence as a measure of discrepancy to minimize, for approximating both the optimal control policy, and the value function. We proved the associated minimizations nicely combine, leaving us with a single variational approximation problem concerning a joint distribution. By forcing this joint distribution to belong to the class of multivariate Gaussians, we end up with a further approximation, that enables tractable formulas for the control policy and value functions, provided we can solve the associated dynamic programming recursion (23).

In the sequel, we show this is feasible at least in two cases. First, in the linear case, where the problem becomes that of an entropy-penalized linear quadratic regulation problem with noisy dynamics $x_{k+1} = Ax_k + Bu_k + \nu_k$. We prove in this case that closed-form formulas may be obtained, and that the Gaussian approximation setting in fact solves the initial problem exactly. Moreover, we may prove a separation principle in the case where only noisy and partial linear measurements of the state of the form $y_k = Hx_k + n_k$ are available. Then, we will turn in Section 4 to control systems which may contain a nonlinearity of the form $x_{k+1} = f(x_k) + Bu_k + \nu_k$. We will apply our Gaussian variational approximation framework, and show closed-form formulas may again be derived for the control policy. This will be applied in Section 5 to a nonlinear inverted pendulum example, and a Dubins car example.

## 3 Max-entropy LQR and Max-entropy LQG

In this section, we consider a noisy linear dynamical system with quadratic cost. When the state is known, we show the entropy-regularized controller resembles the well-known LQR controller, with Gaussian dispersion, providing unpredictable random control that ensures entropic dispersion. We can extend this property to the case where the state is only partially observed, resulting in generalization of LQG control.

### 3.1 Max-entropy LQR

Assume the system takes the linear form:

$$x_{k+1} = Ax_k + Bu_k + \nu_k; \quad k = 0, \dots, K-1, \tag{27}$$

where $x_k, \nu_k \in \mathbb{R}^d$ and $u_k \in \mathbb{R}^p$ with $p \leq d$. $\nu_k$ is a Gaussian white noise $\nu_k \sim \mathcal{N}(0, C)$. $A$, $B$ are real-valued matrices and $C \succ 0$ (i.e., $C$ positive definite). Moreover, we suppose that the transition cost is quadratic and the final cost is "affine quadratic":

$$\begin{aligned} \ell(x_k, u_k) &= \tfrac{1}{2} x_k^T Q x_k + \tfrac{1}{2} u_k^T R u_k, \\ L(x_K) &= \tfrac{1}{2} x_K^T P_K x_K + q_K, \end{aligned} \tag{28}$$

where $Q, P_K \in \mathcal{M}_{d\times d}(\mathbb{R})$; $Q, P_K \succ 0$.

**Synthesis of the MaxEnt LQR**

Without the entropy regularization ($\varepsilon = 0$), the optimal controller is known to be the linear quadratic regulator (LQR). We now show that for our variational problem (12), corresponding to the LQR problem with entropic regularization, the optimal solution generalizes the LQR control, by adding a Gaussian dispersion.

**Corollary 1** (Max-Entropy LQR)**.** *For the variational problem* (12)*, or equivalently* (9)*, with linear dynamics* (27) *and quadratic cost* (28)*, the theoretical optimal solution given in Proposition 1 can be computed analytically and writes:*

$$p^*(u_k|x_k) = \mathcal{N}(K_k x_k, \varepsilon {S_k}^{-1}), \tag{29}$$

*where* $K_k = -{S_k}^{-1} B^T P_{k+1} A$ *is the LQR gain and* $S_k = R + B^T P_{k+1} B$*. The value function is the same as for stochastic LQR and satisfies the recursion:*

$$V_k^*(x_k) = \tfrac{1}{2} x_k^T P_k x_k + q_k \tag{30}$$

$$= \min_{p(u_k|x_k)} \mathbb{E}\Big[\tfrac{1}{2}(x_k^T Q x_k + u_k^T R u_k) + \varepsilon \log p(u_k|x_k) + \tfrac{1}{2} x_{k+1}^T P_{k+1} x_{k+1} + q_{k+1}\Big], \tag{31}$$

*where* $q_k = q_{k+1} + \frac{1}{2}\mathrm{Tr}(P_{k+1} C)$ *and* $P_k$ *satisfies the backward Riccati equation:*

$$P_k = A^\top P_{k+1} A + Q - A^\top P_{k+1} B (R + B^T P_{k+1} B)^{-1} B^T P_{k+1} A. \tag{32}$$

The proof is provided in Appendix 9.

**Remark 1.** *The optimal distribution* $u \mid x$ *is a Gaussian. This means the entropy penalty does not destroy the Gaussianity in the problem. This is not wholly surprising, though, given that the normal law has maximum entropy properties [30].*

With the obtained "randomized" LQR controller, we can sample control inputs from the policy. On average, we recover the LQR feedback control; however, the variance has the magnitude of the temperature parameter $\varepsilon$, resulting in an unpredictable control that ensures entropic dispersion. In the next paragraph, we quantify the resulting entropy.

**Entropy along the trajectory**

The closed-loop stochastic linear system:

$$x_{k+1} = A x_k + B u_k + \nu_k; \quad \nu_k \sim \mathcal{N}(0, C) \tag{33}$$

$$u_k = K_k x_k + z_k; \quad z_k \sim \mathcal{N}(0, \varepsilon S_k^{-1}) \tag{34}$$

rewrites as:

$$x_{k+1} = (A + BK_k) x_k + B z_k + \nu_k \tag{35}$$

$$= (A + BK_k) x_k + n_k; \quad n_k \sim \mathcal{N}(0, N_k), \tag{36}$$

where the covariance of the overall noise is $N_k := \mathbb{E}[(Bz_k + \nu_k)(Bz_k + \nu_k)^T] = \varepsilon B S_k^{-1} B^T + C$.

**Marginals** Starting from a Gaussian prior $p(x_0) = \mathcal{N}(\mu_0, P_0)$, all subsequent marginals remain Gaussian, and we have $p(x_{k+1}) = \mathcal{N}(\mu_{k+1}, P_{k+1})$ where:

$$\mu_{k+1} = \mathbb{E}[x_{k+1}] = (A + BK_k)\mu_k \tag{37}$$

$$\begin{aligned} P_{k+1} &= \mathbb{E}[x_{k+1}x_{k+1}^T] - \mu_{k+1}\mu_{k+1}^T \\ &= (A + BK_k)P_k(A + BK_k)^T + N_k. \end{aligned} \tag{38}$$

The evolution of the mean and parameters of the Gaussian distribution of the state in closed loop is illustrated in Figure 2, where we have considered a simple inverted pendulum system linearized around the unstable equilibrium, that we aim to regulate around the vertical position. We observe that the distribution of the state exhibits a "turnpike" behavior between the initial time and the final time, staying close to an asymptotic distribution with given entropy. We will now characterize this asymptotic distribution.

**Asymptotic distribution** As $k \to +\infty$ we recover a steady-state control scenario. In steady-state control, we search the stationary solution $\Lambda^*$ of the backward Riccati equation (32):

$$\Lambda = A^\top \Lambda A + Q - A^\top \Lambda B(R + B^T \Lambda B)^{-1} B^T \Lambda A.$$

Under the standard controllability and detectability conditions [31], the solution $\Lambda^*$ exists and is unique. This stationary solution yields a fixed steady-state gain $K^*$ and a constant transition probability $p(u|x)$. Furthermore, the Lyapunov equation in this steady-state regime:

$$P = (A + BK^*)P(A + BK^*)^T + N^*$$

has a unique solution [31] given by: $P = \sum_{k=0}^{\infty}(A+BK^*)^k N^*(A + BK^*)^{kT}$. We can then compute the asymptotic distribution as $\mathcal{N}(0, P)$, which is centered around 0, as we have considered a regulation problem with the system operating around the equilibrium point $x^* = 0$. This asymptotic distribution is illustrated in Figure 2, and we can verify that it matches the Gaussian distribution once the asymptotic regime is reached.

### 3.2 Max-entropy LQG

We now consider the case where the state is not known but only partially observable as follows:

$$x_{k+1} = Ax_k + Bu_k + \nu_k, \tag{39}$$

$$y_k = Hx_k + w_k, \tag{40}$$

with $\nu_k \sim \mathcal{N}(0, C)$ as before and $w_k \sim \mathcal{N}(0, N)$ is the observation noise which is a Gaussian white noise independent of $\nu_k$. We still consider a quadratic cost (28).

We suppose we have $K+1$ observations $y_0, \dots, y_K := y_{0:K}$. At time $k$, the estimate of the state $x_k$ depends not only on the observations $y_0, \dots, y_k$ but also on the past controls $u_0, \dots, u_{k-1}$ through the stochastic dynamic equation (39). We hence let $\mathcal{F}_k = \{y_0, u_0, \dots, y_{k-1}, u_{k-1}, y_k\}$ represent the information available up to time $k$, including $y_k$.

#### Reminders on the separation principle

The linear quadratic Gaussian (LQG) controller seeks optimal control inputs given past information, i.e., $u_k = \phi(\mathcal{F}_k)$. For the linear system (39)-(40) with quadratic costs, one can show that all the value

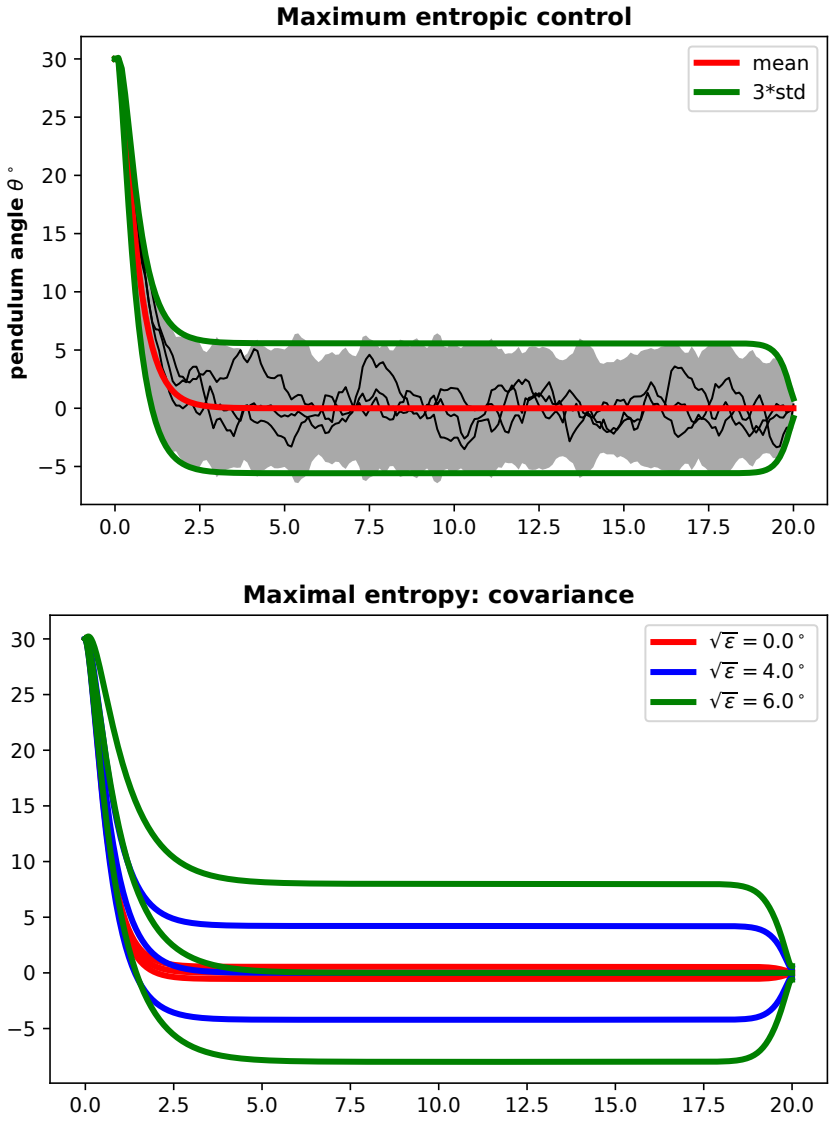


Figure 2: Top plot shows the inverted pendulum's angle over time, where the problem is wholly linearized around equilibrium, with an initial angle of 30°, and with entropic penalty $\sqrt{\varepsilon} = 4°$, over a number of Monte-Carlo simulations. The $3\sigma$ envelope is also displayed, according to our calculation (38), and we see it is accurate. Bottom plot illustrates the $3\sigma$ envelope over time for various values of $\varepsilon$. The final cost is chosen so as to enforce convergence to 0 at the last step.

functions $V_k$ are necessary quadratic, and we can compute the optimal control policy in a closed form [32], yielding the so-called linear quadratic Gaussian (LQG) controller. This controller takes exactly the same form as the LQR controller, except the feedback is based on the state estimate $\hat{x}_k$ provided as the mean of a linear Kalman filter, rather than the true state $x_k$. Please refer to Appendix 11 for a summary of the Kalman filter's equations. This remarkable property is known as the separation principle (and falls more generally into the principle of certainty equivalence) because it allows the optimal control and the optimal state estimate to be computed separately, and then merely combined to get the optimal control of the form $u_k = \phi(\mathcal{F}_k)$. For the LQG problem, the optimal policy indeed writes $u_k = K_k \mathbb{E}[x_k|\mathcal{F}_k] := K_k \hat{x}_k$ where $\hat{x}_k$ is the Kalman filter's mean state estimate. $K_k$ is the same gain as in the LQR case.

Although the separation is genereally difficult to extend, we now show that a separation principle still holds for our variational problem (12) with entropic penalty, in this linear quadratic setting with partially observable state.

**Separation principle for Max-entropy control**

In our context, where $p(u_k|\mathcal{F}_k)$ becomes a distribution, we search for controls of the form $u_k = \phi_k(\mathcal{F}_k, z_k)$, with $z_k$ a random variable independent of the noises: at all times, $u_k$ is computed with the available past information, and $z_k$ is a random term that ensures dispersion and enhances entropy. In other terms, we choose a randomized control input $u_k$ being independent of $w_k, v_{k+1}, w_{k+1}, v_{k+2}, \dots$ conditionally on $y_{0:k}, u_{0:k-1}$.

Given this conditional independence assumption, we search a stochastic policy $p(u_k|\mathcal{F}_k)$ in that form, for $k = 0, \dots, K-1$, which is optimal for the entropy-penalized problem: this matches the LQG framework, but with an added negentropy term $\varepsilon \mathbb{E}[\log p(u_k|\mathcal{F}_k)]$. Starting from time $k = 0$ and after acquiring an observation $y_0$, the optimal control problem now writes:

$$\min_{\substack{p(u_k|\mathcal{F}_k) \\ k=0\dots K-1}} \mathbb{E}\Big[ \sum_{k=0}^{K-1} \ell_k(x_k, u_k) + \varepsilon \log p(u_k|\mathcal{F}_k) + L_K(x_K) | \mathcal{F}_0 \Big], \tag{41}$$

where the expectation is on all the future controlled states, conditionally on the initial information $\mathcal{F}_0 = \{y_0\}$. As for LQG, we can rewrite this problem recursively. Starting from $V_K(x_K) := \mathbb{E}[x_K^T P_K x_K \mid \mathcal{F}_K]$, the optimal cost-to-go (or value function) is defined recursively by (see Appendix 10 for a proof):

$$V_k(\mathcal{F}_k) \quad := \min_{p(u_k|\mathcal{F}_k)} \quad \mathbb{E}\Big[ \ell_k(x_k, u_k) + \varepsilon \log p(u_k|\mathcal{F}_k) + V_{k+1}^{(p)}(\mathcal{F}_{k+1}) \mid \mathcal{F}_k \Big]. \tag{42}$$

It turns out that for the linear system (39)-(40) with quadratic cost, all the value functions $V_k^{(p)}$ hence defined are quadratic, and we can compute the control policy in a closed form. This policy is now stochastic, and comes as a Gaussian distribution, which is remarkable as no assumption is made beforehand on the Gaussianity of the solution. Besides, it satisfies a separation principle as follows.

**Theorem 1** (Max-Entropy LQG)**.** *We consider the problem* (42) *with the linear model* (39)-(40)*, and quadratic cost* (28)*. Assuming $u_k|\mathcal{F}_k$ is independent of the other noises, the optimal solution to this problem is given by:*

$$p^*(u_k|\mathcal{F}_k) = \mathcal{N}(K_k \hat{x}_k, \varepsilon {S_k}^{-1}), \tag{43}$$

*where* $\hat{x}_k := \mathbb{E}[x_k|\mathcal{F}_k]$ *is the linear Kalman filter's state estimate,* $K_k$ *is the LQR gain and* $S_k = R + B^T P_{k+1} B$ *as before. In other terms, we let* $u_k = K_k \hat{x}_k + z_k$*, with* $z_k \sim \mathcal{N}(0, \varepsilon {S_k}^{-1})$ *a dispersion variable being independent of all the other variables at play. The optimal value function is the same as for LQG:*

$$V_k^*(\mathcal{F}_k) = \frac{1}{2} \hat{x}_k^T P_k \hat{x}_k + q_k, \tag{44}$$

*where* $q_k = q_{k+1} + \frac{1}{2}\mathrm{T}r(Q\Sigma_k) + \frac{1}{2}\mathrm{T}r(P_{k+1}(\Sigma_{k+1|k} - \Sigma_k))$ *with* $\Sigma_k = \mathrm{Cov}(x_k|\mathcal{F}_k)$ *and* $\Sigma_{k+1|k} = \mathrm{Cov}(x_{k+1}|\mathcal{F}_k)$*. The matrix* $P_k$ *solves the backward Riccati equation* (32)*.*

The proof is given in Appendix 11. Note that the uncertainty in this LQG policy (the covariance term) arises only from the entropic term, not from the observation noise, which aligns with the certainty equivalence principle in LQG.

**Summary**

Considering entropy-regularized stochastic optimal control for the linear model (39)-(40), results in the following closed-loop optimally-controlled system:

$$x_{k+1} = Ax_k + Bu_k + \nu_k \tag{45}$$
$$y_k = Hx_k + w_k \tag{46}$$
$$u_k = K_k \hat{x}_k + z_k, \tag{47}$$

with $\nu_k \sim \mathcal{N}(0, C)$, $w_k \sim \mathcal{N}(0, N)$ are white noises, $z_k \sim \mathcal{N}(0, \varepsilon(R + B^T P_{k+1} B)^{-1})$ optimally ensures dispersion (i.e., entropy) and $\hat{x}_k$ the Kalman filter's estimate. Note the linear-Gaussian form of $u_k$ was not assumed a priori, and appears as a consequence of our separation principle, see Theorem 1.

# 4 Max-entropy VLQR

We have seen that the original entropy-regularized stochastic optimal control problem (8) is amenable to the dynamic programming recursion (23), when constraining the policy and value distribution to lie in some approximating families. If we opt for the Gaussian family (24), Problem (22) becomes an optimization problem over the parameters $\alpha_k$, $\beta_k$, $S_k$, $P_k$ and $K_k$. Following our previous work on recursive variational Gaussian approximation [33], we seek to derive (backward) recursive equations for those parameters. To achieve this, we focus on control-affine systems where the control inputs enter linearly into the dynamics. This includes various mechanical systems, such as the cart-pole system or the two-link robot of [34], see also applications below. Opting for quadratic cost functions, we then obtain equations that generalize the backward Riccati equation from LQR control.

## 4.1 Nonlinear Control-Affine Dynamics

We now focus on dynamics of the following form:

$$x_{k+1} = f(x_k) + Bu_k + \nu_k, \qquad \nu_k \sim \mathcal{N}(0, C), \tag{48}$$

where $B \in \mathcal{M}_{d\times m}(\mathbb{R})$ and $C \in \mathcal{M}_{d\times d}(\mathbb{R})$; $C \succ 0$. It entails that $p(x_{k+1}|x_k, u_k) = \mathcal{N}(x_{k+1}|f(x_k) + Bu_k, C)$. We also choose to work with quadratic costs with $Q, P_K \in \mathcal{M}_{d\times d}(\mathbb{R})$; $Q, P_K \succ 0$:

$$\begin{aligned} \ell(x_k, u_k) &= \tfrac{1}{2}(x_k - x_k^*)^T Q(x_k - x_k^*) + \tfrac{1}{2}u_k^T R u_k, \\ L(x_K) &= \tfrac{1}{2}(x_K - x_K^*)^T P_K(x_K - x_K^*), \end{aligned} \tag{49}$$

where $x_k^*$ for $k = 1, \dots, K$ is the reference trajectory. Our results will remain valid if the matrixes $B, C, Q, R$ depend on $k$. We start with a Gaussian centered at $\alpha_K = x_K^*$.

### 4.2 Variational Backward Riccati Equation

We now show that the solution to the problem (22) is given by a generalization of the backward Riccati equation (the proof is postponed to Appendix 12).

**Proposition 2.** *Consider the dynamic programming recursion* (22) *stemming from problem* (8)*, with dynamics* (48) *and with costs* (49)*. Suppose the "value distribution"* (13) *at previous step is in the form of a Gaussian* $\phi(x_{k+1}) = \mathcal{N}(\alpha_{k+1}, \varepsilon P_{k+1}^{-1})$ *with known parameters* $\alpha_{k+1}, P_{k+1}$*. Then, the optimal joint Gaussian* (24) *for the problem* (22) *satisfies:*

$$\begin{aligned} q(x_k) &= \mathcal{N}(\alpha_k, \varepsilon P_k^{-1}) \\ q(u_k|x_k) &= \mathcal{N}(\beta_k + K_k(x_k - \alpha_k), \varepsilon S_k^{-1}), \end{aligned}$$

*with* $S_k$*,* $\beta_k$ *and* $K_k$ *given by*

$$\begin{aligned} S_k &= R + B^T P_{k+1} B, \quad K_k = -S_k^{-1} B^T P_{k+1} \mathbb{E}_q\Big[\frac{\partial f}{\partial x}(x_k)\Big] \\ \beta_k &= -S_k^{-1} B^T P_{k+1}(\mathbb{E}_q\big[f(x_k)\big] - \alpha_{k+1}), \end{aligned} \tag{50}$$

*and where* $\alpha_k$ *and* $P_k$ *satisfy the generalized (implicit) backward Riccati equation*

$$\begin{aligned} \alpha_k &= x_k^* - Q^{-1}\mathbb{E}_q\Big[\frac{\partial f}{\partial x}(x_k)^\top P_{k+1}(f(x_k) + Bu_k - \alpha_{k+1})\Big] \\ P_k &= Q - \mathbb{E}_q\Big[\frac{\partial f}{\partial x}(x_k)\Big]^\top P_{k+1} B S_k^{-1} B^T P_{k+1} \mathbb{E}_q\Big[\frac{\partial f}{\partial x}(x_k)\Big] \\ &\qquad + \mathbb{E}_q\Big[\frac{\partial f}{\partial x}(x_k)^\top P_{k+1}\frac{\partial f}{\partial x}(x_k) + H_k\Big], \end{aligned} \tag{51}$$

*where* $H_k \in \mathcal{M}_d(\mathbb{R})$ *is given by the tensor contraction of the Hessian of* $f$*:*

$$H_k[\mu, \nu] = \sum_{ij}(P_{k+1})_{ij}(f(x_k) + Bu_k - \alpha_{k+1})_i \frac{\partial^2 f_j}{\partial x^\mu \partial x^\nu}.$$

*In all the expectancies above, subscript* $q$ *denotes the joint distribution* $q(x_k, u_k)$*.* ■

The obtained equation resembles the Riccati equation from LQR, but with the presence of expectations. These equations are implicit because the expectations are taken over the sought distribution. This is akin to our prior work in the field of probabilistic inference [33], and various techniques can allow us to get around this issue, as will be discussed shortly.

We conclude this subsection with an additional result, proving that when $f$ is odd, the problem simplifies by symmetry (see proof in Appendix 12).

**Lemma 1.** *Assume we start with a terminal cost that is centered, in the sense that* $\phi(x_K) := \exp(-L_K(x_K)/\varepsilon)$ *is (up to a normalization constant) a centered Gaussian* $\mathcal{N}(0, \varepsilon P_K^{-1})$*. Assume additionally* $f(-x) = -f(x)$ *for all* $x$*. Then for all* $k < K$ *we have* $\alpha_k = \beta_k = 0$*.* ■

### 4.3 Discussion

In the case of control-affine nonlinear dynamics, and assuming centered distributions to simplify, we see we essentially recover LQR equations where $A$ is replaced with an expectation of the form

$$\mathbb{E}_q\Big[\frac{\partial f}{\partial x}(x_k)\Big] = \int \frac{\partial f}{\partial x}(x)\tilde{C}\exp\Big(-\frac{x^T P_k x}{2\varepsilon}\Big)|P_k|^{-1/2}\frac{1}{\sqrt{\varepsilon}}dx.$$

A change of variables shows this is equal to

$$\int \frac{\partial f}{\partial x}(\sqrt{\varepsilon}y)\tilde{C}\exp\Big(-\frac{y^T P_k y}{2}\Big)|P_k|^{-1/2}dy.$$

We see the effect of entropy regularization is to perform an average of magnitude $\sqrt{\varepsilon}$ around the equilibrium (assuming 0 is the equilibrium we seek to stabilize), and as $\varepsilon \to 0$ we have $\mathbb{E}_q\Big[\frac{\partial f}{\partial x}(x_k)\Big] \to \frac{\partial f}{\partial x}(0)$, and we recover the LQR equations linearized at equilibrium.

Note that the equations are implicit. In (50), the definition of $P_k$ is based on an average over $q$, whose variance is $P_k/\varepsilon$, which reminds of our previous work on variational inference [33]. In practice, we can cycle as follows for small $\varepsilon$. We assume $\varepsilon = 0$ initially, which gives a first estimate for $P_k$ based on the linearization at equilibrium, as previously explained. Then, we may recompute, letting the obtained $P_k$ be the variance of $q$. After a few iterations, the scheme converges in practice.

**Remark 1.** *Note that the control gain of our policy* (50) *is defined by* $K_k = -{S_k}^{-1}B^T P_{k+1}\mathbb{E}_q[\frac{\partial f}{\partial x}(x)]$. *Taking an average is likely to make the policy more robust to model uncertainty; see, e.g., [3].*

**Remark 2.** *Another attractive property of our policy is that it allows for the computation of controls when the dynamics* $f$ *is nondifferentiable. Indeed, we can avoid computing the Jacobian matrix of the dynamics considering instead the Jacobian matrix of the Gaussian:* $\mathbb{E}_q[\frac{\partial f}{\partial x}(x)] = -\int \frac{\partial q}{\partial x}(x)f(x)dx$. *This equality results from integration by part on the Gaussian* $q$, *which has a support that vanishes at* $\pm\infty$. *Nondifferentiable control appears, for example, in collision detection with randomized smoothing [35].*

## 5 Numerical results

### 5.1 Variational Control of a Pendulum

To illustrate the method and to gain some insight into the obtained optimal solution, we focus on the case study of a pendulum controlled by a torque $u$ and perturbed by a noise $w$. This is a simple example but sufficiently nonlinear to showcase the differences between linearized LQR and entropy-regularized optimal control. The dynamics write

$$\ddot{\theta} + \lambda\dot{\theta} - \omega^2\sin\theta = \tfrac{1}{m\ell^2}u + \sqrt{\eta}w,$$

where $\theta$ is the angle with respect to the pendulum at the unstable equilibrium (upward position), $\omega = \sqrt{g/\ell}$ is the pulsation, $\lambda = \xi/m$ the damping parameter and $\eta > 0$ is the magnitude of the noise. In state-space form, the dynamics are discretized in time as follows:

$$\begin{aligned}\begin{pmatrix}\theta_{k+1}\\ \dot{\theta}_{k+1}\end{pmatrix} &= \begin{pmatrix}\theta_k\\ \dot{\theta}_k\end{pmatrix} + \delta t\begin{pmatrix}\dot{\theta}_k\\ -\lambda\dot{\theta}_k + \omega^2\sin\theta_k\end{pmatrix}\\ &\quad + \delta t\begin{pmatrix}0\\ \frac{1}{m\ell^2}\end{pmatrix}u_k + \sqrt{\delta t\eta}\begin{pmatrix}0\\ 1\end{pmatrix}w, \quad w\sim\mathcal{N}(0,1),\\ &:= f(x_k) + Bu_k + \nu_k.\end{aligned}$$

The discrete cost writes

$$x_K^T Q x_K + \sum_{k=0}^{K-1} x_k^T Q x_k + u_k^T R u_k.$$

Starting from $\theta_0$ we seek to stabilize the inverted pendulum while penalizing the entropy of the policy.

We will compare our variational control with the control given by LQR with dynamics linearized around the equilibrium $x^* = 0$, that is, letting $A = \begin{pmatrix} 1 & \delta t \\ \delta t g/\ell & 1 - \delta t \lambda \end{pmatrix}$.

**Solution**

Since the dynamics of the inverted pendulum satisfy the oddness condition of Lemma 1, we have $\alpha_k = 0$ and $\beta_k = 0$, and the optimal policy is given by $q(u_k|x_k) = \mathcal{N}\big(K_k x_k, \varepsilon S_k^{-1}\big)$ with $K_k$ and $S_k$ defined in Proposition 2. To compute this optimal policy, there are two hurdles: the variational Riccati equation (51) is implicit, and there are expectations to compute. As already mentioned in Section 4.3, to cope with the fact the equation is implicit, we may open the loop and iterate on the equation in an inner loop. As concerns the expectations under Gaussians, they are approximated using quadrature rules:

$$\int \mathcal{N}(\mu, P) g(x) dx \approx \sum_{i=1}^{M} w_i g(x_i),$$

where we can choose $M = 2d$ cubature points [36] defined by $w_i = \frac{1}{2d}$ and $x_i = \mu + \sqrt{d} L e_i$ where $e_i$ are basis vectors in dimension $d$, and $L$ the square root matrix of the covariance such that the points are equally spread at the edge of the Gaussian ellipsoid.

**Numerical Settings**

We take the following parameters: $g = 9.8, m = 1, \ell = 1$ and $\lambda = 1$. We start at $\theta_0 = \frac{\pi}{6}$ and $\dot{\theta} = 0$, and we want to put the pendulum at $(\theta, \dot{\theta}) = (0, 0)$ which corresponds by convention to the unstable equilibrium (upward position) such that the stable equilibrium (downward position) is at $\theta = \pi$. We consider a backward pass with 1000 iterations with stepsize $\delta t = 0.01$ such that the temporal horizon is $T = 10s$. We simulate the Brownian motion with a Gaussian increment of covariance $\delta t \eta$ where $\eta = 0.02\ rd/s^2$ in the first experiment and $\eta = 0.2\ rd/s^2$ in the second one. The forward trajectory is simulated with a semi-implicit Euler-Maruyama scheme to better conserve the system's energy. For the "variational control," the implicit Riccati backward equation is iterated 10 times in an inner loop; however, we found out that one iteration could be used in practice without much affecting the results.

**Average control**

We first apply the average value of the policy distribution by letting $u_k := K_k x_k$. Figure 3 illustrates the behavior in function of $\sqrt{\varepsilon}$, and compares it to LQR control based on the system linearized at the equilibrium. We see clearly that for the smaller value of $\sqrt{\varepsilon}$, both controllers behave similarly, but when $\sqrt{\varepsilon}$ increases, the gains with the variational control are below the LQR gains, leading to softer controls based on averaging a trigonometric function around its maximum (softer controls may preserve actuators). To underline the effect of $\varepsilon$, we have considered a small cost: $R = \delta t \cdot 0.01 \mathbb{I}_m$, $Q = \delta t \cdot 0.01 \mathbb{I}_d$ and $P_K = \delta t \cdot \mathbb{I}_d$.

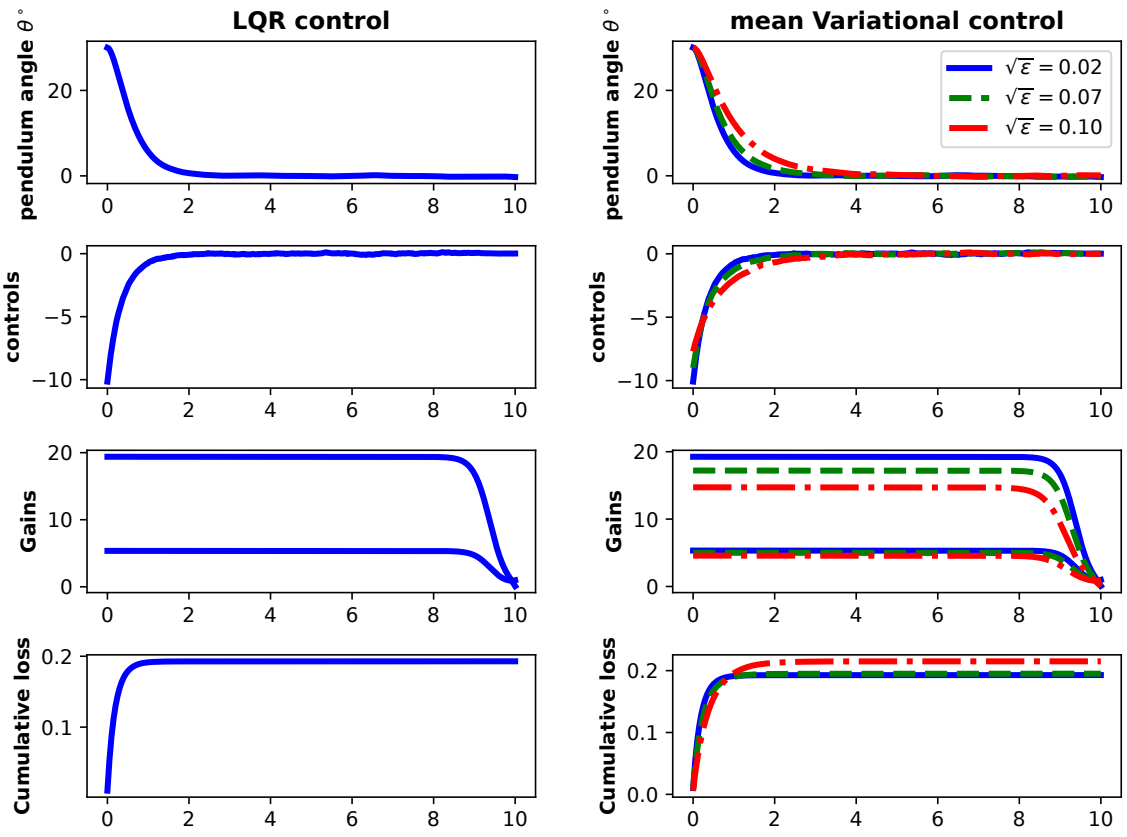


Figure 3: Linearized LQR at equilibrium versus our "variational control" for the inverted pendulum regulation where we apply for the KL the mean policy $\mathbb{E}[u_k|x_k]$ for different values of entropic regulation $\sqrt{\varepsilon} = 0.02, 0.07, 0.10$. From the top to the bottom row, we show the angle $\theta$ converted in degrees, the control, the two gains for angle and angular velocity, and finally, we compare both LQR and variational control with the same LQR quadratic loss.

### Random control

We now sample the control from the actual policy distribution $q(u_k|x_k) = \mathcal{N}(K_k x_k, \varepsilon(R+B^T P_{k+1} B)^{-1})$. We consider a large terminal cost $P_K = \delta t \cdot 1000 \mathbb{I}_d$ but low stage costs $R = \delta t \cdot 0.01 \mathbb{I}_m$, $Q = \delta t \cdot 0.01 \mathbb{I}_d$. In this way, we elicit high entropy along the path (hence exploration of the state space) while enforcing the final equilibrium state. Results are displayed in Figure 4, where we see the empirical distribution of the state when applying random controls. The distribution $p(x_k)$ of the state spreads during the transient phase but shrinks to the equilibrium indeed at the final time $T$.

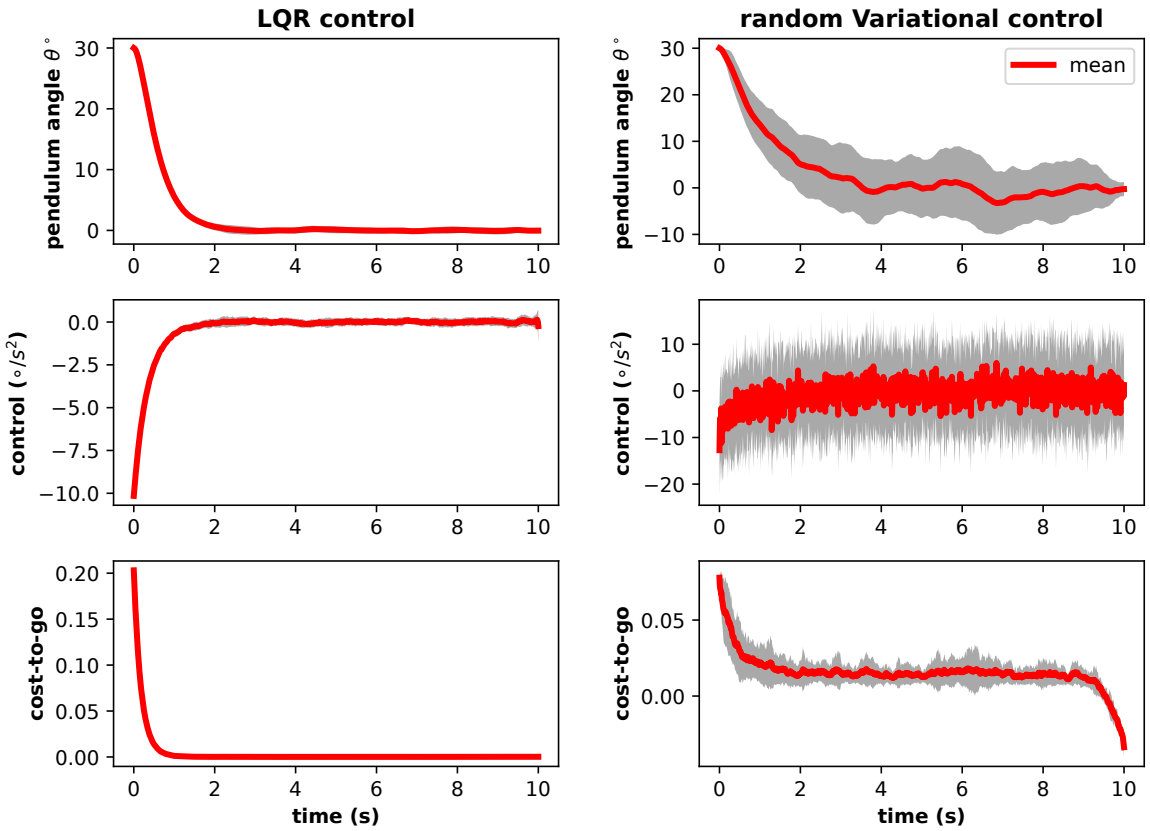


Figure 4: Linearized LQR versus "Variational control" for the inverted pendulum regulation where we sample randomly from the policy $q(u_k|x_k)$. The entropic regulation parameter is fixed to $\sqrt{\varepsilon} = 0.10$. We execute 30 Monte Carlo runs, and for each output, we draw the empirical mean in red and the empirical standard deviation in grey. From the top to the bottom row, we show the angle $\theta$ converted in degrees, the control, and the cost-to-go at the current state. The cost-to-go at $x_k$ is defined by $\frac{1}{2} x_k^T P_k x_k$ for LQR and by $-\varepsilon \log q(x_k) = -\varepsilon \log \mathcal{N}(x_k|0, \varepsilon P_k^{-1})$ for variational control.

## 5.2 Variational Control of a Dubins car

We now consider a Dubins car with a constant velocity $v$ and a heading angle $\theta$, which is controlled by a front wheel with a steering angle $\delta$. This model corresponds to the cinematic bicycle model (two-wheel model). The steering angle can be computed as a function of the desired gyration radius

$\mathcal{R}$ of the trajectory and on the distance between the front and back wheels $L$ as follows:

$$\dot{x} = v\cos(\theta); \quad \dot{y} = v\sin(\theta) \tag{52}$$

$$\dot{\theta} = \frac{1}{L} v \tan\delta + \nu_k \approx \frac{1}{L} v\delta + \nu_k \tag{53}$$

$$\delta = \operatorname{atan}\frac{L}{\mathcal{R}} \approx \frac{L}{\mathcal{R}}. \tag{54}$$

These equations give in discrete time:

$$\begin{pmatrix} x_{k+1} \\ y_{k+1} \\ \theta_{k+1} \end{pmatrix} = \begin{pmatrix} x_k + v\cos(\theta_k)dt \\ y_k + v\sin(\theta_k)dt \\ \theta_k \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \frac{1}{L} v dt \end{pmatrix} \delta_k + \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} w_k. \tag{55}$$

We generate a reference trajectory $(x_k^*, y_k^*, \theta_k^*)$ for a given sequence of piecewise constant gyration radius $\mathcal{R}_t$ such that $\dot{\theta}_t = \frac{v}{\mathcal{R}_t}$. The trajectory is generated up to time $T = 10s$ starting from $(x = 0, y = 0)$ with $v = 2m/s$ and $L = 0.1m$. The gyradion radius are $\mathcal{R}_t = +5$ for $t \in [0, 5]$, $\mathcal{R}_t = -2$ for $t \in [5, 7.5]$ and $\mathcal{R}_t = +3$ for $t \in [7.5, 10]$.

This reference trajectory is used in the transition cost equation (49) where we consider a matrix $Q = 0.01dt\mathbb{I}_3$. We consider an energy cost $R = 0.01dt$ for the control $\delta$. We find experimentally that the controller matches the theoretical $\delta$ steering angle well when no entropic noise is added.

Notice that, contrary to the pendulum case, the transition function si not odd and we will have $\alpha_k \neq 0$ and $\beta_k \neq 0$.

Finally, we add to the acceleration term a stochastic noise $w_k$ of variance $1°$ and, when applying a random control, an entropic noise driven by a temperature parameter $\varepsilon$. The results are shown in Figure 5, where we can see that the car deviates from the reference trajectory as the entropy increases (via the temperature parameter $\varepsilon$), whereas the mean remains close. This results in a desirable (that is, close to the reference) but difficult-to-predict control.

*The sources of the code are available on Github on the following repository:*
*https://github.com/marc-h-lambert/KL-control.*

# 6 Conclusion

We have proposed a new framework for stochastic optimal control based on the entropic regularization of both the dynamics entropy and the policy entropy. This problem was reformulated as a KL divergence between two processes: the first defining the controlled stochastic trajectory and the second defining a reward process. Following a variational dynamic programming principle, we derived new formulas for the exact optimal policy and cost-to-go.

In the linear case, we were able to compute this exact policy, leading to the Max-entropy LQR controller. We also showed that the separation principle can be generalized by designing a Max-entropy LQG controller.

In the nonlinear case, we proposed a method to jointly approximate the control policy and the cost-to-go. We applied this method to a joint Gaussian model in the case of nonlinear dynamics with affine control inputs and quadratic costs. This approximation can be computed in closed form, resulting in a Max-entropy VLQR controller, where the cost-to-go formulas generalize the backward Riccati equation from LQR control.

To illustrate the results, we have performed simulations using our new policy on a second-order system. Using the average policy results in softer control with smaller gains than LQR, whereas

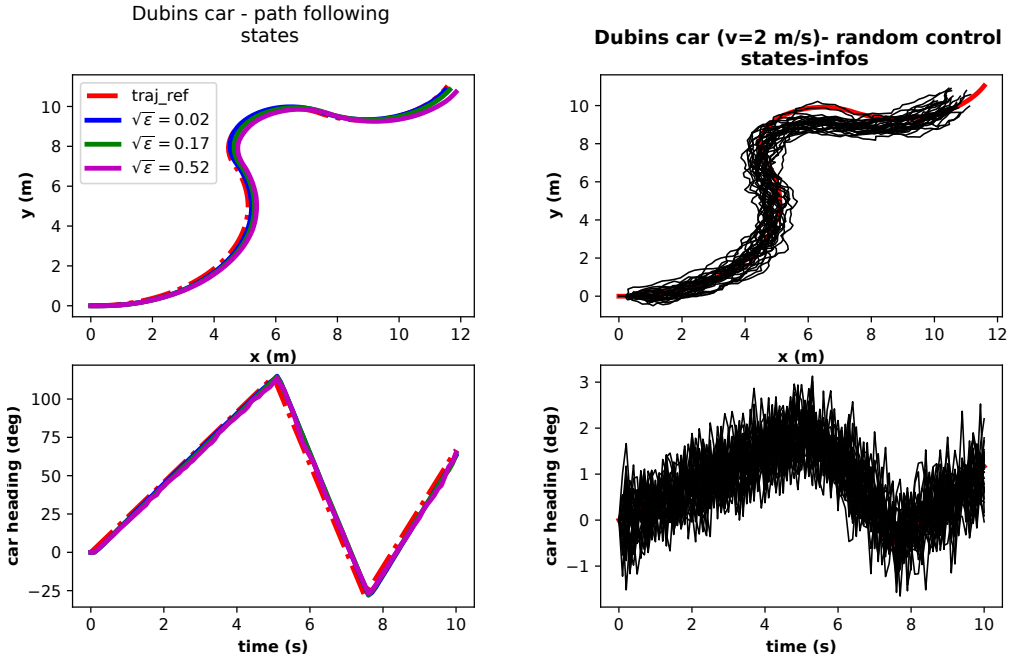


Figure 5: Variational control of a nonlinear Dubins car with mean control (left column) for different temperatures $\varepsilon$ and randomized control (right column) with $\varepsilon = 1°$. The reference is shown in red for the position (top row) and for the car heading (bottom row).

the random policy causes dispersion in the state space during the transient phase. We also applied the method a mobile robotic problem where the entropy generates unpredictability.

The proposed method paves the way for future work: the control affine model can be made richer by considering a state-dependent control matrix $B(x)$ and a state-dependent covariance of Brownian motion $C(x)$. Moreover, we could use richer approximating families, such as mixtures of Gaussians, to more closely capture the value function.

Finally, we may wonder if the separation principle (Theorem 1) can be extended to the nonlinear case. This will be investigated in future work.

## Acknowledgment

This work was funded by the French Defense Procurement Agency (DGA) and by the French National Research Agency, under the France 2030 program with the reference "PR[AI]RIE-PSAI" (ANR-23-IACL-0008).

## References

[1] B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey, "Maximum entropy inverse reinforcement learning," *AAAI Conference on Artificial Intelligence*, pp. 1433–1438, 2008.

[2] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," *International Conference on Machine Learning*, 2018.

[3] B. D. Ziebart, J. A. Bagnell, and A. K. Dey, "Modeling interaction via the principle of maximum causal entropy," *International Conference on Machine Learning*, pp. 1255–1262, 2010.

[4] O. So, Z. Wang, and E. A. Theodorou, "Maximum entropy differential dynamic programming," *International Conference on Robotics and Automation*, pp. 3422–3428, 2022.

[5] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, 2008.

[6] R. A. Howard and J. E. Matheson, "Risk-sensitive markov decision processes," *Management Science*, vol. 18, pp. 356–369, 1972.

[7] W. H. Fleming and R. W. Rishel, *Deterministic and Stochastic Optimal Control*. Springer, 1975.

[8] W. H. Fleming, "Risk sensitive stochastic control and differential games," *Communications In Information And Systems*, 2006.

[9] P. D. Pra, L. Meneghini, and W. J. Runggaldier, "Connections between stochastic control and dynamic games," *Mathematics of Control, Signals and Systems*, vol. 9, pp. 303–326, 1996.

[10] V. S. Borkar, "Learning algorithms for risk-sensitive control," *19th International Symposium on Mathematical Theory of Networks and Systems*, 2010.

[11] V. Anantharam and V. S. Borkar, "A variational formula for risk-sensitive reward," *SIAM Journal on Control and Optimization*, vol. 55, no. 2, pp. 961–988, 2017.

[12] M. Moharrami, Y. Murthy, A. Roy, and R. Srikant, "A policy gradient algorithm for the risk-sensitive exponential cost mdp," *Math. Oper. Res.*, vol. 50, no. 1, p. 431–458, 2025.

[13] P. Whittle, "Risk-sensitive linear quadratic gaussian control," *Advances in Applied Probability*, 1981.

[14] M. C. Campi and M. R. James, "Nonlinear discrete-time risk-sensitive optimal control," *International Journal Of Robust And Nonlinear Control*, vol. 6, pp. 1–19, 1996.

[15] A. Jordana, B. Hammoud, J. Carpentier, and L. Righetti, "Stagewise newton method for dynamic game control with imperfect state observation," *IEEE Control Systems Letters*, vol. 6, pp. 3241–3246, 2022.

[16] A. Jordana, A. Meduri, E. Arlaud, J. Carpentier, and L. Righetti, "Risk-sensitive extended kalman filter," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 10 450–10 456.

[17] H. J. Kappen, "Path integrals and symmetry breaking for optimal control theory," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, no. 11, p. P11011, nov 2005.

[18] E. A. Theodorou and E. Todorov, "Relative entropy and free energy dualities: Connections to path integral and KL control," *Conference on Decision and Control*, pp. 1466–1473, 2012.

[19] G. Williams, A. Aldrich, and E. A. Theodorou, "Model predictive path integral control: From theory to parallel computation," *Journal of Guidance, Control, and Dynamics*, vol. 40, no. 2, pp. 344–357, 2017.

[20] M. Toussaint and A. Storkey, "Probabilistic inference for solving discrete and continuous state markov decision processes," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, p. 945–952.

[21] E. Todorov, "Linearly-solvable markov decision problems," in *Advances in Neural Information Processing Systems*, B. Schölkopf, J. Platt, and T. Hoffman, Eds., vol. 19. MIT Press, 2006.

[22] J. Watson, H. Abdulsamad, and J. Peters, "Stochastic optimal control as approximate input inference," *Proceedings of Machine Learning Research*, vol. 100, pp. 697–716, 30 Oct–01 Nov 2020.

[23] E. Todorov, "Efficient computation of optimal actions," *Proceedings of the National Academy of Sciences*, pp. 11 478–11 483, 2009.

[24] Y. Chen, T. T. Georgiou, and M. Pavon, "Stochastic control liaisons: Richard Sinkhorn meets Gaspard Monge on a Schrödinger bridge," *SIAM Review*, pp. 249–313, 2021.

[25] K. Rawlik, M. Toussaint, and S. Vijayakumar, "On stochastic optimal control and reinforcement learning by approximate inference," *International Joint Conference on Artificial Intelligence*, 2012.

[26] D. Gagliardi and G. Russo, "On a probabilistic approach to synthesize control policies from example datasets," *Automatica*, vol. 137, 01 2022.

[27] M. Lambert, F. Bach, and S. Bonnabel, "Variational dynamic programming for stochastic optimal control," *Conference on Decision and Control (CDC)*, 2024.

[28] E. Todorov, "General duality between optimal control and estimation," *Conference on Decision and Control*, 2008.

[29] M. Toussaint, "Robot trajectory optimization using approximate inference," *International Conference on Machine Learning*, pp. 1049–1056, 2009.

[30] T. M. Cover and J. A. Thomas, *Elements of Information Theory, 2nd Edition.* Wiley-Interscience, 2006.

[31] P. Lancaster and L. Rodman, *Algebraic Riccati Equations.* Clarendon Press, 2002.

[32] R. F. Stengel, *Stochastic optimal control: theory and application.* USA: John Wiley & Sons, Inc., 1986.

[33] M. Lambert, S. Bonnabel, and F. Bach, "The recursive variational gaussian approximation (R-VGA)," *Statistics and Computing*, vol. 32, 2022.

[34] M. W. Spong, "Underactuated mechanical systems," in *Control Problems in Robotics and Automation.* Springer, 2005, pp. 135–150.

[35] L. Montaut, Q. L. Lidec, A. Bambade, V. Petrík, J. Sivic, and J. Carpentier, "Differentiable collision detection: a randomized smoothing approach," *International Conference on Robotics and Automation*, pp. 3240–3246, 2023.

[36] I. Arasaratnam and S. Haykin, "Cubature Kalman filters," *IEEE Trans. Automat. Control*, vol. 54, no. 6, pp. 1254–1269, 2009.

# 7 Appendix

# 8 Proof of Proposition 1

To prove the Proposition 1 we use the following lemma.

**Lemma 2.** *Let's consider the following problem:*

$$\min_{p(z)\in\mathcal{P}(\mathbb{R}^d)} \mathrm{KL}(p(z)p(x|z)\|h(x,z))$$
$$:= \min_{p(z)\in\mathcal{P}(\mathbb{R}^d)} \int\int p(z)p(x|z)\log\frac{p(z)p(x|z)}{h(x,z)}dxdz,$$

*where $p(z)$ and $p(x|z)$ are probability distributions and $h$ is a density function which may be unnormalized. $\mathcal{P}(\mathbb{R}^d)$ is the space of probability distribution smoothed enough to admit a density function. Then, the minimum is attained at*

$$p^*(z) = \frac{1}{Z}\exp\Big(-\int p(x|z)\log\frac{p(x|z)}{h(x,z)}dx\Big),$$

*where $Z$ is the normalization constant of $p^*(z)$. Moreover, the minimum is $-\log Z$.* ■

*Proof.*

$$\begin{aligned}
&\int\int p(z)p(x|z)\log\frac{p(x|z)p(z)}{h(z,x)}dxdz\\
&=\int p(z)\int p(x|z)\log p(z)dxdz\\
&+\int p(z)\int p(x|z)\log\frac{p(x|z)}{h(z,x)}dxdz\\
&=\int p(z)\log p(z)dz-\int p(z)\log f(z)dz\\
&\quad\text{where}\quad f(z)=\exp\Big(-\int p(x|z)\log\frac{p(x|z)}{h(z,x)}dx\Big)\\
&=KL(p(z)||f(z))\quad\text{which is minimal for } p(z)\propto f(z)\\
&=-\log Z\quad\text{where } Z=\int f(z)dz.
\end{aligned}$$

□

Applying Lemma 2 to $z = u_k|x_k$, $x = x_{k+1}$ and $h(z,x) = r(x_k,u_k)\phi^*_{k+1}(x_{k+1})$, we obtain the desired result.

# 9 Proof of corollary 1

The optimal policy $p(u_k|x_k)$ and value function $V^c_k(x_k)$ which satisfy the maximum entropy principle are given by the proposition 1 in their theoretical general form and when the state is fully observable.

We show here that in the case of a linear model and a quadratic cost the optimal policy is necessary a Gaussian distribution and the value is a affine quadratic form. We proceed by recursion.

Starting from $V^{(c)}_K(x_K) := \frac{1}{2}x^T_K P_K x_K$, we show that $p(u_k|x_k)$ is Gaussian and $V_k$ is an affine quadratic form and satisfy the recursion:

$$\frac{1}{2}x^T_k P_k x_k + q_k := \min_{p(u_k|x_k)} \mathbb{E}\Big[\frac{1}{2}(x^T_k Q x_k + u^T_k R u_k) + \varepsilon\log p(u_k|x_k) + \frac{1}{2}x^T_{k+1}P_{k+1}x_{k+1} + q_{k+1}\Big].$$

*Proof.* Applying Proposition 1, the optimal control in the general setting is proportional to the unnormalized distribution:

$$p^*(u_k|x_k) \propto \quad \exp\Big[-\tfrac{1}{\varepsilon}\Big(\tfrac{1}{2}x_k^T Q x_k + \tfrac{1}{2}u_k^T R u_k + \mathbb{E}[\tfrac{1}{2}x_{k+1}^T P_{k+1} x_{k+1}] + q_{k+1}\Big)\Big]. \tag{56}$$

Since the expectation is under the transition $p(x_{k+1}|x_k, u_k) = \mathcal{N}(Ax_k + Bu_k, C)$ we have the relation:

$$\mathbb{E}[\frac{1}{2}x_{k+1}^T P_{k+1} x_{k+1}] = \frac{1}{2}(Ax_k + Bu_k)^T P_{k+1}(Ax_k + Bu_k) + \frac{1}{2}\mathrm{Tr}(P_{k+1}C). \tag{57}$$

Putting this result in (56), we find the optimal control takes the form $p(u_k|x_k) \propto \exp(-\frac{1}{\varepsilon}F(u_k, x_k))$ where $F$ is a affine quadratic form in $u_k$ defined by:

$$F(u_k, x_k) = \frac{1}{2}u_k^T(R + B^T P_{k+1} B)u_k + u_k^T B^T P_{k+1} A x_k + \frac{1}{2}x_k^T(Q + A^T P_{k+1} A)x_k + \frac{1}{2}\mathrm{Tr}(P_{k+1}C) + q_{k+1}.$$

Since $F$ is quadratic in $u_k$ we deduce that the optimal distribution is a Gaussian. Its parameters can be obtained by completing the square on $u_k$. Using the notation $K_k = -(R+B^T P_{k+1} B)^{-1} B^T P_{k+1} A$, we obtain:

$$\begin{aligned} p(u_k|x_k) &= \mathcal{N}(K_k x_k, \varepsilon(R + B^T P_{k+1} B)^{-1}) \\ &:= \frac{1}{Z(x_k)} \exp(-G(x_k, u_k)), \end{aligned} \tag{58}$$

where $Z(x_k) = \int \exp(-G(x_k, u_k))du_k$ is a normalization constant which does not depend on $u_k$. That proves the first part of the Corollary 1.

From proposition 1 again, the optimal value function at step $k$ satisfies the "log-sum-exp" recursion:

$$\begin{aligned} V_k^c(x_k) &= -\varepsilon \log \int \exp\Big[-\frac{1}{\varepsilon}(\frac{1}{2}x_k^T Q x_k + \frac{1}{2}u_k^T R u_k + \mathbb{E}[\frac{1}{2}x_{k+1}^T P_{k+1} x_{k+1}] + q_{k+1}\Big]du_k \\ &:= -\varepsilon \log \int \exp(-F(x_k, u_k))du_k. \end{aligned} \tag{59}$$

If we note $\Phi^c(x_k) = \exp(-V_k^c(x_k)/\varepsilon)$, we observe that $\Phi^c(x_k) = \int \exp(-F(x_k, u_k))du_k$ is also a normalization constant for $p(u|x)$ but $\Phi^c$ is obtained by normalizing $\exp(-F)$ whereas $Z$ was obtained by normalizing $\exp(-G)$. We then have the equality:

$$p(u_k|x_k) = \frac{1}{\Phi^c(x_k)} \exp(-F(x_k, u_k)) \tag{60}$$

$$= \frac{1}{Z(x_k)} \exp(-G(x_k, u_k)), \tag{61}$$

we can then deduce:

$$\Phi^c(x_k) = Z(x_k) \exp(-(F(x_k, u_k) - G(x_k, u_k))).$$

Finally, we obtain:

$$F(x_k,u_k)-G(x_k,u_k)=\frac{1}{2}x_k^T(Q+A^TP_{k+1}A-A^TP_{k+1}B(R+B^TP_{k+1}B)^{-1}B^TP_{k+1}A)x_k \\ +\frac{1}{2}\mathrm{T}r(P_{k+1}C)+q_{k+1} \\ =\frac{1}{2}x_k^TP_kx_k+q_k, \quad (62)$$

where $q_k=q_{k+1}+\frac{1}{2}\mathrm{T}r(P_{k+1}C)$ and $P_k=Q+A^TP_{k+1}A-A^TP_{k+1}B(R+B^TP_{k+1}B)^{-1}B^TP_{k+1}A$ which is the backward Riccati equation. We have found a value function that is also quadratic at index $k$, allowing us to conclude by recursion that all policies are Gaussian and all value functions are quadratic. □

# 10 Proof of the variational dynamic principle for partially observable state

We prove that the total loss for the Max-entropy LQG problem (41) satisfies the recursion (42), using the same method as for LQG [32]. We first define the cost-to-go as:

$$V_k^{(p)}(\mathcal{F}_k):=\min_{\substack{p(u_i|\mathcal{F}_i)\\ k\leq i\leq K-1}}\mathbb{E}\Big[\sum_{i=k}^{K-1}\ell_i(x_i,u_i) \\ +\varepsilon\log p(u_i|\mathcal{F}_i)+L_K(x_K)\mid\mathcal{F}_k\Big], \quad (63)$$

such that $V_0^{(p)}(\mathcal{F}_0)$ correspond to the total loss (41). We show that this cost-to-go can be computed recursively, leading to an extension of the variational dynamic programming principle in the case

of partially observable states. The cost-to-go (63) rewrites:

$$
\begin{aligned}
V_k^{(p)}(\mathcal{F}_k) &= \min_{p(u_k|\mathcal{F}_k)} \Big\{ \mathbb{E}\Big[ \ell_k(x_k,u_k) + \varepsilon \log p(u_k|\mathcal{F}_k) \mid \mathcal{F}_k \Big] \\
&\quad + \min_{\substack{p(u_i|\mathcal{F}_i) \\ k+1 \le i \le K-1}} \mathbb{E}\Big[ \sum_{i=k+1}^{K-1} \ell_i(x_i,u_i) + \varepsilon \log p(u_i|\mathcal{F}_i) + L_K(x_K) \mid \mathcal{F}_k \Big] \Big\} \\
&= \min_{p(u_k|\mathcal{F}_k)} \Big\{ \mathbb{E}\Big[ \ell_k(x_k,u_k) + \varepsilon \log p(u_k|\mathcal{F}_k) \mid \mathcal{F}_k \Big] \\
&\quad + \min_{\substack{p(u_i|\mathcal{F}_i) \\ k+1 \le i \le K-1}} \mathbb{E}\Big[ \mathbb{E}\Big[ \sum_{i=k+1}^{K-1} \ell_i(x_i,u_i) + \varepsilon \log p(u_i|\mathcal{F}_i) + L_K(x_K) \mid \mathcal{F}_{k+1} \Big] \mid \mathcal{F}_k \Big] \Big\}, \\
&= \min_{p(u_k|\mathcal{F}_k)} \Big\{ \mathbb{E}\Big[ \ell_k(x_k,u_k) + \varepsilon \log p(u_k|\mathcal{F}_k) \mid \mathcal{F}_k \Big] \\
&\quad + \mathbb{E}\Big[ \min_{\substack{p(u_i|\mathcal{F}_i) \\ k+1 \le i \le K-1}} \mathbb{E}\Big[ \sum_{i=k+1}^{K-1} \ell_i(x_i,u_i) + \varepsilon \log p(u_i|\mathcal{F}_i) + L_K(x_K) \mid \mathcal{F}_{k+1} \Big] \mid \mathcal{F}_k \Big] \Big\} \\
&= \min_{p(u_k|\mathcal{F}_k)} \Big\{ \mathbb{E}\Big[ \ell_k(x_k,u_k) + \varepsilon \log p(u_k|\mathcal{F}_k) \mid \mathcal{F}_k \Big] \\
&\quad + \mathbb{E}\Big[ V_{k+1}^{(p)}(\mathcal{F}_{k+1}) \mid \mathcal{F}_k \Big] \Big\}.
\end{aligned}
$$

We finally obtain the dynamic programming recursion:

$$
V_k^{(p)}(\mathcal{F}_k) = \min_{p(u_k|\mathcal{F}_k)} \mathbb{E}\Big[ \ell_k(x_k,u_k) + \varepsilon \log p(u_k|\mathcal{F}_k) + V_{k+1}^{(p)}(\mathcal{F}_{k+1}) \mid \mathcal{F}_k \Big],
$$

which is the one stated in (42).

# 11 Proof of the separation principle for max entropy LQG of Theorem 1

We first recall the Kalman filter equations. We need to be careful, because a priori we cannot assume the control inputs to be Gaussian, to prove the theorem which will precisely prove that. Let us thus study what happens to the Kalman filter in the presence of possibly non-Gaussian control inputs.

### The Kalman filter with random non-Gaussian inputs

We assumed control inputs of the form $u_k = \phi_k(\mathcal{F}_k, z_k)$ with $z_k$ is a random variable independent of other noises $\nu_k, w_k$ for any $k$ and where $\mathcal{F}_k = \{y_{0:k}, u_{0:k-1}\}$ is the past available information.

The linear Kalman filter (KF) with non-Gaussian random control inputs is defined analogously to the standard linear Kalman filter. Assume at step $k$ that we have

$$x_k \mid \mathcal{F}_k \sim \mathcal{N}(\hat{x}_k, \Sigma_k). \tag{64}$$

We will show by induction an analog at time $k+1$ is obtained using the standard Kalman equations. In other words, $x_k$ may not be Gaussian, but it remains so conditionally on the past. The proof mimics that of the KF, but as the pair $(x_k, y_k)$ can no longer be assumed a Gaussian vector, owing to $u_k$ intervening in their definition, one needs to adapt the proof and check it still holds. The main spring that supports our results is that the state $x_k - \hat{x}_k$ error evolves independently of the control inputs $u_k$ in the linear case, so that one may still may use Gaussian conditioning theory, provided appropriate conditioning is performed.

### 11.0.1 propagation step

Assume at step $k$ that we have

$$x_k \mid y_{0:k}, u_{0:k-1} \sim \mathcal{N}(\hat{x}_k, \Sigma_k). \tag{65}$$

Consider the following propagation step:

$$\hat{x}_{k+1|k} = A\hat{x}_k + Bu_k.$$

Let us define the state estimation error $\tilde{x}_{k+1|k} := x_{k+1} - \hat{x}_{k+1|k}$ and analogously $\tilde{x}_k := x_k - \hat{x}_k$. We have

$$\tilde{x}_{k+1|k} = A\tilde{x}_k + \nu_k. \tag{66}$$

Of course $\tilde{x}_k \mid y_{0:k}, u_{0:k-1} \sim \mathcal{N}(0, \Sigma_k)$, so that

$$\tilde{x}_{k+1|k} \mid y_{0:k}, u_{0:k-1} \sim \mathcal{N}(0, A\Sigma_k A^T + C), \tag{67}$$

where we recall that $\nu_k \sim \mathcal{N}(0, C)$.

Now, note that as $z_k$ is independent of $\tilde{x}_{k+1|k}$, owing to the error propagation (66) not depending on $u_k$, we have the equalities in law $\tilde{x}_{k+1|k} \mid y_{0:k}, u_{0:k-1} = \tilde{x}_{k+1|k} \mid \{y_{0:k}, u_{0:k-1}, \phi(y_{0:k}, z_k)\} := \tilde{x}_{k+1|k} \mid y_{0:k}, u_{0:k}$. This yields

$$\tilde{x}_{k+1|k} \mid y_{0:k}, u_{0:k} \sim \mathcal{N}(0, A\Sigma_k A^T + C), \tag{68}$$

hence

$$x_{k+1} \mid y_{0:k}, u_{0:k} \sim \mathcal{N}(\hat{x}_{k+1|k}, A\Sigma_k A^T + C). \tag{69}$$

We define $\Sigma_{k+1|k} := A\Sigma_k A^T + C$.

We need now to perform the update step, so as to obtain the analog of our initalization formula (64) at next $k+1$. We let $e_{k+1} = y_{k+1} - H\hat{x}_{k+1|k} = H\tilde{x}_{k+1|k} + v_{k+1}$ be the innovation. As $(\tilde{x}_{k+1|k}, e_{k+1}) \mid y_{0:k}, u_{0:k} = (\tilde{x}_{k+1|k}, H\tilde{x}_{k+1|k} + v_{k+1}) \mid y_{0:k}, u_{0:k}$ is a Gaussian vector, we may apply the Gaussian conditioning theorems. Recalling our prior (68) we find that $\tilde{x}_{k+1|k} \mid y_{0:k}, u_{0:k}, e_{k+1} \sim \mathcal{N}(L_{k+1}e_{k+1}, (I - L_{k+1})\Sigma_{k+1|k})$ where $L_{k+1} := \Sigma_{k+1|k}H^T(H\Sigma_{k+1|k}H^T + W)^{-1}$ is the Kalman gain.

What we need, to complete our recursion, though, is to compute $x_{k+1} \mid y_{0:k+1}, u_{0:k}$. Hence we need to transfer the conditioning on $y_{0:k}, u_{0:k}, e_{k+1}$ to a conditioning on $y_{0:k}, u_{0:k}, y_{k+1}$. As $e_{k+1} = y_{k+1} - H\hat{x}_{k+1|k}$ and $\hat{x}_{k+1|k}$ is a function of $y_{0:k}, u_{0:k}$, both conditioning are equivalent so

$$\tilde{x}_{k+1|k} \mid y_{0:k+1}, u_{0:k} \sim \mathcal{N}(L_{k+1}e_{k+1}, (I - L_{k+1})\Sigma_{k+1|k}).$$

This finally yields

$$x_{k+1} \mid y_{0:k+1}, u_{0:k} \sim \mathcal{N}(\hat{x}_{k+1|k} + L_{k+1}e_{k+1}, (I - L_{k+1})\Sigma_{k+1|k}),$$

which may rewrite

$$x_{k+1} \mid y_{0:k+1}, u_{0:k} \sim \mathcal{N}(\hat{x}_{k+1}, (I - L_{k+1})\Sigma_{k+1|k}),$$

recalling the KF update

$$\hat{x}_{k+1} := \hat{x}_{k+1|k} + L_{k+1}e_{k+1}, \tag{70}$$

and we recover (64) at next step, as by definition the KF equations ensure $(I - L_{k+1})\Sigma_{k+1|k} := \Sigma_{k+1}$.

From the above derivation, we obtain the following properties, which will be useful for computing the conditional expectations involved in the dynamic programming recursion:

- The covariance of the innovation error given the past information $\mathcal{F}_k$ is:

$$\mathbb{E}\Big[e_{k+1}e_{k+1}^\top \mid \mathcal{F}_k\Big] = H\Sigma_{k+1|k}H^T + N. \tag{71}$$

- As $\hat{x}_k$ is $\sigma(\mathcal{F}_k)$-measurable, and the conditional expectation is linear, we have

$$\mathbb{E}\Big[\hat{x}_k e_{k+1}^\top \mid \mathcal{F}_k\Big] = \hat{x}_k \mathbb{E}\Big[e_{k+1}^\top \mid \mathcal{F}_k\Big] = 0. \tag{72}$$

- Since we have supposed $u_k|\mathcal{F}_k$ independent of all other noises, and $\tilde{x}_{k+1|k}$ is defined independently of $u_k$, see (66), then $u_k$ is independent of $e_{k+1}$ given $\mathcal{F}_k$. Thus $\mathbb{E}\Big[u_k e_{k+1}^\top \mid \mathcal{F}_k\Big] = \mathbb{E}\Big[u_k \mid \mathcal{F}_k\Big]\mathbb{E}\Big[e_{k+1}^\top \mid \mathcal{F}_k\Big]$ such that:

$$\mathbb{E}\Big[u_k e_{k+1}^\top \mid \mathcal{F}_k\Big] = 0. \tag{73}$$

## Proof of theorem 1

We now have the tools to prove Theorem 1. To show that the value function remains quadratic throughout the backward dynamic programming process, we proceed by recursion. Let's first check the final value has the desired form for a quadratic final cost $L_K(x_K) = \frac{1}{2}x_K^T P_K x_K$. The final value function is defined as:

$$V_K^{(p)}(\mathcal{F}_K) := \mathbb{E}[L_K(x_K)|\mathcal{F}_K].$$

We use the following lemma, which is well-known and easily proved:

**Lemma 3.** *If* $x \sim \mathcal{N}(\mu, \Sigma)$ *then* $\mathbb{E}[x^T Q x] = \mu^T Q \mu + \mathrm{T}r(Q\Sigma)$.

As we have proved the Kalman filter ensures $x_K|\mathcal{F}_K \sim \mathcal{N}(\hat{x}_K, \Sigma_K)$ even if control inputs are non Gaussian, we have:

$$V_K^{(p)}(\mathcal{F}_K) = \mathbb{E}[L_K(x_K)|\mathcal{F}_K] = \frac{1}{2}\hat{x}_K^T P_K \hat{x}_K + \frac{1}{2}\mathrm{T}r(P_K\Sigma_K).$$

To show the value function as the desired form for $k < K$, we continue the recursion backwards using the dynamic programming defnition (42) that we recall presently

$$\begin{aligned} V_k^{(p)}(\mathcal{F}_k) = \min_{p(u_k|\mathcal{F}_k)} \mathbb{E}\Big[\frac{1}{2}(x_k^T Q x_k + u_k^T R u_k) + \varepsilon \log p(u_k|\mathcal{F}_k) \\ + \frac{1}{2}\hat{x}_{k+1}^T P_{k+1}\hat{x}_{k+1} + q_{k+1} \mid \mathcal{F}_k\Big]. \end{aligned} \tag{74}$$

Since from (64) we have $x_k \mid \mathcal{F}_k \sim \mathcal{N}(\hat{x}_k, \Sigma_k)$, Lemma 3 yields:

$$\mathbb{E}\Big[\frac{1}{2}x_k^T Q x_k \mid \mathcal{F}_k\Big] = \frac{1}{2}\hat{x}_k^T Q \hat{x}_k + \frac{1}{2}\mathrm{T}r(Q\Sigma_k).$$

The rightmost term expands as follows:

$$\begin{aligned} &\mathbb{E}\Big[\frac{1}{2}\hat{x}_{k+1}^\top P_{k+1}\hat{x}_{k+1} + q_{k+1} \mid \mathcal{F}_k\Big] \\ \overset{(70)}{=}&\mathbb{E}\Big[\frac{1}{2}(A\hat{x}_k + Bu_k + L_{k+1}e_{k+1})^\top P_{k+1}(A\hat{x}_k + Bu_k \\ &+ L_{k+1}e_{k+1}) + q_{k+1} \mid \mathcal{F}_k\Big] \\ =&\mathbb{E}\Big[\frac{1}{2}(A\hat{x}_k + Bu_k)^\top P_{k+1}(A\hat{x}_k + Bu_k) \\ &+ \frac{1}{2}(L_{k+1}e_{k+1})^\top P_{k+1}(L_{k+1}e_{k+1}) \\ &+ (A\hat{x}_k + Bu_k)^\top P_{k+1}(L_{k+1}e_{k+1}) + q_{k+1} \mid \mathcal{F}_k\Big] \end{aligned}$$

$$
\begin{aligned}
=&\mathbb{E}\Big[\frac{1}{2}(A\hat{x}_k + Bu_k)^\top P_{k+1}(A\hat{x}_k + Bu_k) \mid \mathcal{F}_k\Big] \\
&+ \mathrm{Tr}(\frac{1}{2}L_{k+1}^\top P_{k+1}L_{k+1}\mathbb{E}\Big[e_{k+1}e_{k+1}^\top \mid \mathcal{F}_k\Big]) \\
&+ \mathrm{Tr}(A^\top P_{k+1}L_{k+1}\mathbb{E}\Big[e_{k+1}\hat{x}_k^\top \mid \mathcal{F}_k\Big]) \\
&+ \mathrm{Tr}(B^\top P_{k+1}L_{k+1}\mathbb{E}\Big[e_{k+1}u_k^\top \mid \mathcal{F}_k\Big]) + q_{k+1} \\
=&\mathbb{E}\Big[\frac{1}{2}(A\hat{x}_k + Bu_k)^\top P_{k+1}(A\hat{x}_k + Bu_k) \mid \mathcal{F}_k\Big] \\
&+ \mathrm{Tr}(\frac{1}{2}L_{k+1}^\top P_{k+1}L_{k+1}(H\Sigma_{k+1|k}H^T + N)) + q_{k+1},
\end{aligned}
$$

where we have used the equations (71), (72) and (73) to simplify the latter expression. Since $\hat{x}_k$ is a function of $\mathcal{F}_k$ and $u_k = \phi_k(\mathcal{F}_k, z_k)$, with $z_k$ independent from $\mathcal{F}_k$ the conditional expectation above boils down to:

$$
\begin{aligned}
&\mathbb{E}\Big[\frac{1}{2}(A\hat{x}_k + Bu_k)^\top P_{k+1}(A\hat{x}_k + Bu_k) \mid \mathcal{F}_k\Big] \\
&= \mathbb{E}_{p(u_k|\mathcal{F}_k)}\Big[\frac{1}{2}(A\hat{x}_k + Bu_k)^\top P_{k+1}(A\hat{x}_k + Bu_k)\Big],
\end{aligned}
$$

which should be understood in the sense of a simple expectation:

$$
\int \frac{1}{2}(A\hat{x}_k + B\phi_k(\mathcal{F}_k, z_k))^\top P_{k+1}(A\hat{x}_k + B\phi_k(\mathcal{F}_k, z_k))p(z_k)dz_k,
$$

and similarly $\mathbb{E}\Big[\frac{1}{2}u_k^T Ru_k|\mathcal{F}_k\Big] = \mathbb{E}_{p(u_k|\mathcal{F}_k)}\Big[\frac{1}{2}u_k^T Ru_k\Big]$ in the same sense. The remaining term may be simplified as in the LQG framework using the definition of the Kalman gain $L_{k+1}$ and the Riccati equation on $\Sigma_k$, which yields

$$
\begin{aligned}
&\frac{1}{2}\mathrm{Tr}(L_{k+1}^\top P_{k+1}L_{k+1}(H\Sigma_{k+1|k}H^T + N)) &(75)\\
&= \frac{1}{2}\mathrm{Tr}(P_{k+1}(\Sigma_{k+1|k} - \Sigma_k)). &(76)
\end{aligned}
$$

Combining all the results, we have largely got rid of the conditional expectations, and we end up with

$$
\begin{aligned}
V_k^{(p)}(\mathcal{F}_k) = \min_{p(u_k|\mathcal{F}_k)} &\mathbb{E}_{p(u_k|\mathcal{F}_k)}\Big[\frac{1}{2}\hat{x}_k^T Q\hat{x}_k + \frac{1}{2}u_k^T Ru_k \\
&+ \varepsilon \log p(u_k|\mathcal{F}_k) \\
&+ \frac{1}{2}(A\hat{x}_k + Bu_k)^\top P_{k+1}(A\hat{x}_k + Bu_k)\Big] \\
&+ \frac{1}{2}\mathrm{Tr}(Q\Sigma_k) + \frac{1}{2}\mathrm{Tr}(P_{k+1}(\Sigma_{k+1|k} - \Sigma_k)) + q_{k+1}.
\end{aligned} \tag{77}
$$

This problem is equivalent to the following problem:

$$V_k^{(p)}(\mathcal{F}_k) = \min_{p(u_k|\mathcal{F}_k)} \mathrm{KL}(p(u_k \mid \mathcal{F}_k) \| r(u_k, \hat{x}_k)), \tag{78}$$

where:

$$\begin{aligned} r(u_k, \hat{x}_k) \propto \exp -\frac{1}{\varepsilon}\Big(&\frac{1}{2}\hat{x}_k^T Q \hat{x}_k + \frac{1}{2}u_k^T R u_k \\ &+ \frac{1}{2}(A\hat{x}_k + Bu_k)^\top P_{k+1}(A\hat{x}_k + Bu_k) + q_k\Big), \end{aligned} \tag{79}$$

where we let $q_k = \frac{1}{2}\mathrm{T}r(Q\Sigma_k) + \frac{1}{2}\mathrm{T}r(P_{k+1}(\Sigma_{k+1|k} - \Sigma_k)) + q_{k+1}$. The optimal policy is $p^*(u_k|\mathcal{F}_k) \propto r(u_k, \hat{x}_k)$ which is clearly a Gaussian. We now proceed as in Corollary 1, Equation (56), completing the square in $u_k$ to compute the Gaussian mean and covariance. Following again Corollary 1 we can then deduce the backward Riccati equation for the value function. This leads to the desired separation principle stated in Theorem 1.

# 12 Proof of Proposition 2 and Lemma 1

To show Proposition 2, we first reformulate the problem (22) for our particular control-affine setting.

**Reformulation of the problem**

$$\begin{aligned} &\mathrm{KL}(q(x_k, u_k)p(x_{k+1}|x_k, u_k) \| \exp(-\ell(x_k, u_k)/\varepsilon)q(x_{k+1})) \\ &= -H(q(x_k, u_k)) - H(p(x_{k+1}|x_k, u_k)) \\ &+ \int q(x_k, u_k)\frac{1}{\varepsilon}\ell(x_k, u_k)dx_k du_k \\ &- \int q(x_k, u_k) \int p(x_{k+1}|x_k, u_k) \log q(x_{k+1}) dx_{k+1} dx_k du_k, \end{aligned}$$

where $H$ is the entropy operator which writes $H(q(x_k, u_k)) = \frac{1}{2}\log|\Sigma_k| + c$ and $H(p(x_{k+1}|x_k, u_k)) = c'$ where $c$ and $c'$ are constants independent of the variational parameters. The last integral on $p(x_{k+1}|x_k, u_k)$ simplifies as follows, denoting $p(x_{k+1}|x_k, u_k)$ by $p$:

$$\begin{aligned} &\int p(x_{k+1}|x_k, u_k) \log q(x_{k+1}) dx_{k+1} := \mathbb{E}_p[\log q(x_{k+1})] \\ &= \mathbb{E}_p[\frac{1}{2\varepsilon}(x_{k+1} - \alpha_{k+1})^\top P_{k+1}(x_{k+1} - \alpha_{k+1})] \\ &= \frac{1}{2\varepsilon}(f(x_k) + Bu_k - \alpha_{k+1})^\top P_{k+1}(f(x_k) + Bu_k - \alpha_{k+1}) \\ &+ \mathbb{E}_p[\nu_k^\top P_{k+1}\nu_k], \end{aligned}$$

where $\mathbb{E}_p[\nu_k^\top P_{k+1}\nu_k] = \operatorname{tr} C P_{k+1}$ interestingly does not depend on variational parameters. Finally, (22) reduces to:

$$\min_{\mu_k,\Sigma_k} \quad \mathbb{E}_q[g(x_k,u_k)] - \frac{1}{2}\log|\Sigma_k|, \tag{80}$$

where $\mathbb{E}_q$ denotes the expectation under $q(x_k,u_k) = \mathcal{N}(\mu_k,\Sigma_k)$ and where $g$ is defined as follows:

$$\begin{aligned} g(x_k,u_k) &= \frac{1}{2\varepsilon}((x_k - x_k^*)^T Q(x_k - x_k^*) + u_k^T R u_k) \\ &+ \frac{1}{2\varepsilon}(f(x_k) + Bu_k - \alpha_{k+1})^\top P_{k+1}(f(x_k) + Bu_k - \alpha_{k+1}). \end{aligned}$$

**Closed form solution**

To solve (80), we use the property of integration under Gaussian distribution described in the following result known as Stein's Lemma.

**Lemma 4.** *For a function $f : \mathbb{R}^d \to \mathbb{R}$,*

$$\begin{aligned} \nabla_\mu \int \mathcal{N}(x|\mu,\Sigma) f(x) dx &= \int \mathcal{N}(x|\mu,\Sigma)\nabla f(x) dx \\ \nabla_\Sigma \int \mathcal{N}(x|\mu,\Sigma) f(x) dx &= \frac{1}{2}\int \mathcal{N}(x|\mu,\Sigma)\nabla^2 f(x) dx. \end{aligned}$$

*Proof.* The proof comes from integration by part and using the symmetric properties of Gaussians $\nabla_\mu \mathcal{N}(x|\mu,\Sigma) = -\nabla_x \mathcal{N}(x|\mu,\Sigma)$ and $\nabla_\Sigma \mathcal{N}(x|\mu,\Sigma) = \frac{1}{2}\nabla_x^2 \mathcal{N}(x|\mu,\Sigma)$. □

Using this lemma and the relation $\nabla_\Sigma \log|\Sigma| = \Sigma^{-1}$, the derivative with respect to $\Sigma_k$ of the quantity (80) writes:

$$\frac{1}{2}\mathbb{E}_q\left[\begin{pmatrix} \nabla_{xx} g_k(x_k,u_k) & \nabla_{xu} g_k(x_k,u_k) \\ \nabla_{ux} g_k(x_k,u_k) & \nabla_{uu} g_k(x_k,u_k) \end{pmatrix}\right] - \frac{1}{2}\Sigma_k^{-1}.$$

Writing $\nabla_\Sigma(\cdot) = 0$ yields for the problem at hand

$$\Sigma_k^{-1} = \frac{1}{\varepsilon}\left[\begin{pmatrix} \varepsilon\mathbb{E}_q\Big[\nabla_{xx} g(x_k,u_k)\Big] & \mathbb{E}_q\Big[\frac{\partial f}{\partial x}(x_k)^T\Big] P_{k+1} B \\ B^T P_{k+1}\mathbb{E}_q\Big[\frac{\partial f}{\partial x}(x_k)\Big] & R + B^T P_{k+1} B \end{pmatrix}\right].$$

Recalling our model for the joint covariance as a $2\times 2$ block matrix $\Sigma_k$ (24), we can compare the above matrix with the inverse $\Sigma_k^{-1}$ given by :

$$\Sigma_k^{-1} = \varepsilon^{-1}\begin{pmatrix} P_k + K_k^\top S_k K_k & -K_k^\top S_k \\ -S_k K_k & S_k \end{pmatrix}. \tag{81}$$

By identification, this readily yields

$$\begin{aligned} S_k &= R + B^T P_{k+1} B \\ -S_k K_k &= B^T P_{k+1} \mathbb{E}_q\Big[\frac{\partial f}{\partial x}(x_k)\Big] \\ P_k + K_k^\top S_k K_k &= Q + \mathbb{E}_q\Big[\frac{\partial f}{\partial x}(x_k)^T P_{k+1} \frac{\partial f}{\partial x}(x_k) + H_k\Big], \end{aligned} \tag{82}$$

where the last equation comes from a computation of the upper left term $\mathbb{E}_q\Big[\nabla_{xx} g_k(x_k, u_k)\Big]$. We then deduce the expression for $K_k$ and $P_k$.

From Stein's lemma, the derivative w.r.t. $\mu_k$ of (80) is $\Big(\mathbb{E}_q\Big[\frac{\partial g}{\partial x}(x_k)\Big], \mathbb{E}_q\Big[\frac{\partial g}{\partial u}(u_k)\Big]\Big)$. Setting it to zero gives :

$$\begin{aligned} 0 &= Q(\alpha_k - x_k^*) + \mathbb{E}_q\Big[\frac{\partial f}{\partial x}(x_k)^T P_{k+1}(f(x_k) + Bu_k - \alpha_{k+1})\Big] \\ 0 &= \frac{1}{\varepsilon}(R\beta_k + B^T P_{k+1} B\beta_k + B^T P_{k+1}(\mathbb{E}_q\big[f(x_k)\big] - \alpha_{k+1})), \end{aligned}$$

from which we deduce the expression for $\alpha_k$ and $\beta_k$.

**Proof of Lemma 1**

We now show the general equations (50)-(51) may be simplified under oddness conditions. Assume $\alpha_{k+1} = 0$. We let $\alpha_k = 0$ and $\beta_k = 0$, and we want to show the equations on $\alpha_k, \beta_k$ are satisfied. By doing so, we are dealing with centered expectancies. We have $\mathbb{E}[f(x_k)] = 0$, proving $\beta_k = 0$ is consistent with $\alpha_k = 0$. Besides, we have $\frac{\partial f}{\partial x}(-x) = \frac{\partial f}{\partial x}(x)$, which entails $\alpha_k = 0 \Rightarrow \mathbb{E}[\frac{\partial f}{\partial x}(x_k)^T P_{k+1} f(x_k)] = 0$. Finally, we write using the law of total expectation

$$\begin{aligned} &\mathbb{E}_{q(x_k,u_k)}\Big[\frac{\partial f}{\partial x}(x_k)^T P_{k+1} B u_k\Big] \\ &= \mathbb{E}_{q(x_k)}\Big[\frac{\partial f}{\partial x}(x_k)^T P_{k+1} B \mathbb{E}[u_k \mid x_k]\Big] \\ &= \mathbb{E}_{q(x_k)}\Big[\frac{\partial f}{\partial x}(x_k)^T P_{k+1} B K_k x_k\Big], \end{aligned}$$

and we use the fact we integrate an odd function w.r.t. a centered Gaussian.