On Entropy Regularized Path Integral Control for Trajectory Optimization

Tom Lefebvre [1,2,*] and **Guillaume Crevecoeur [1,2]**

[1] Department of Electromechanical, Systems and Metal Engineering, Ghent University, 9000 Ghent, Belgium; guillaume.crevecoeur@ugent.be
[2] EEDT Decision & Control, Flanders Make, 3001 Heverlee, Belgium
[*] Correspondence: tom.lefebvre@ugent.be

Abstract: In this article, we present a generalized view on Path Integral Control (PIC) methods. PIC refers to a particular class of policy search methods that are closely tied to the setting of Linearly Solvable Optimal Control (LSOC), a restricted subclass of nonlinear Stochastic Optimal Control (SOC) problems. This class is unique in the sense that it can be solved explicitly yielding a formal optimal state trajectory distribution. In this contribution, we first review the PIC theory and discuss related algorithms tailored to policy search in general. We are able to identify a generic design strategy that relies on the existence of an optimal state trajectory distribution and finds a parametric policy by minimizing the cross-entropy between the optimal and a state trajectory distribution parametrized by a parametric stochastic policy. Inspired by this observation, we then aim to formulate a SOC problem that shares traits with the LSOC setting yet that covers a less restrictive class of problem formulations. We refer to this SOC problem as Entropy Regularized Trajectory Optimization. The problem is closely related to the Entropy Regularized Stochastic Optimal Control setting which is often addressed lately by the Reinforcement Learning (RL) community. We analyze the theoretical convergence behavior of the theoretical state trajectory distribution sequence and draw connections with stochastic search methods tailored to classic optimization problems. Finally we derive explicit updates and compare the implied Entropy Regularized PIC with earlier work in the context of both PIC and RL for derivative-free trajectory optimization.

Keywords: entropic inference; entropy regularization; stochastic search methods; path integral control

1. Introduction

Finding controllers for systems that are high dimensional and continuous in space is still one of the most challenging problems faced by the robotic and control community. The goal is to find a feedback policy that stabilizes the system and that possibly also encodes rich and complex dynamic behavior such as locomotion [1]. A feedback policy is a state- and time-dependent function $u(t, x)$ with u representing the input applied to the system that is in a state x at time t. A well-known paradigm to design such policies is Stochastic Optimal Control (SOC) [2]. The policy is determined such that when applied to the system, it is expected to accumulate a minimized cost over a specified finite time horizon. Finding such an optimal feedback policy is an elegant and appealing theoretical idea but meets significant difficulties in practice. Explicit expressions for the optimal policy $u^*(t, x)$ rarely exist and one often has to resort to local solutions instead. Such local solutions can be found by so called trajectory optimization algorithms (In this work we focus on discrete-time system and model-based strategies, in the sense that a simulator is available that closely approximates the actual dynamics) [3,4]. These yield open-loop solutions of the form $u^*(t) = u^*(t, x^*(t))$ where the signals $u^*(t)$ and $x^*(t)$ are optimal in the sense that they minimize the accumulated cost starting from a given fixed initial state.

In hierarchical control approaches such trajectory optimizers are used to discover rich and complex dynamic behaviors "offline" [5]. The system is then stabilized "online" using a linearized feedback-controller around the optimal trajectory. The locally linear feedback-controller is usually designed such that $u^*_{\text{fb}}(t,x) = k(t) + K(t)x(t) = u^*(t) + K(t)(x(t) - x^*(t))$ where $K(t) \approx \nabla_x u(t, x^*(t))$ and $k(t) = u^*(t) + K(t)x^*(t)$. The gradient $K(t)$ is approximated by defining a Linear Quadratic Regulator (LQR) problem around the optimal trajectory using a second order Taylor expansion. In fact, this LQR problem then also provides a correction to $\{x^*(t), u^*(t)\}$ were this linearization point may not be optimal. As such, a new trajectory $\{u^\star(t), x^\star(t)\}$ is obtained around which a new LQR can be defined. Iterating this approach will incrementally improve the trajectory. This is exactly how gradient based trajectory optimization algorithms work such as Differential Dynamic Programming (DDP) [3,6] and iterative LQR (iLQR) [7]. It is implied that gradient based trajectory optimizers require a differentiable model and cost function. Furthermore, they are notoriously ill-suited to handle state constraints. A second application is that of Model Predictive Control (MPC). The idea is to compute an optimal trajectory initialized with the current state measurement and apply the optimal open-loop controller every sample instant. This approach provides a "just-in-time" service and circumvents computation of the explicit policy altogether. MPC is feasible provided that the trajectory optimization problem can be solved in a single sample period [8]. This poses hard real-time requirements on the trajectory optimizer that are rarely met in practice, specifically for nonlinear dynamics with state and input constraints.

1.1. Stochastic Search Methods

In this article, we are interested in solving the trajectory optimization problem above using a sample based optimization method, or a so-called stochastic search method. Stochastic search methods rely on randomness to probe the optimization space and maintain mechanisms that eventually guide that randomness towards prosperity. Broadly speaking, a stochastic search method maintains a prior population and generates a posterior population based on the prosperity of the individuals.

Evolutionary Strategies (ESs) refer to a particular subclass of stochastic search algorithms tailored to static optimization, that, as opposed to population based algorithms [9], engage a parametric (θ) *search distribution* model, $\pi(x|\theta) : \mathcal{X} \mapsto \mathbb{R}_{\geq 0}$. Every main generation, g, a sample population, $\mathcal{X}_g = \{x_g^k\}$, is generated from a search distribution $\pi(x|\theta_g)$ and the search distribution parameters, $\theta_{g+1} \leftarrow \theta_g$, are updated based on the relative success of the individual samples. The objective function value $f(x^k)$ is used as a discriminator between prosperous and poor behavior of each individual, x^k [10]. When iterated this concept spawns a sequence of distributions, $\{\pi_g\}$. The update procedure is devised so that the distribution sequence migrates gradually through the optimization space and concentrates around the optimal solution eventually. Although that limiting the sequence to a parametric distribution family may compromise the inherent expressiveness or elaborateness of the associated search, it also elevates the determination of update rules, from what are basically heuristics, to a more rigorous and theoretical body [11,12]. Well-known members are the Covariance Matrix Adaptation Evolutionary Strategy (CMA-ES) [11,13] and the Natural Evolutionary Strategy (NES) [14]. Most ESs consider a multivariate Gaussian parametric distribution and provide appropriate update procedures for the mean, μ_g, and covariance, Σ_g. Examples of application on medium to high scale problems for non-differential design optimization are [15–21].

$$\pi_g(x) = \mathcal{N}(\theta_g) = \mathcal{N}(x|\mu_g, \Sigma_g)$$

In the present work, we aim to address trajectory optimization problems, exploiting stochasticity as a natural means of exploration, in a similar fashion as how ESs address static optimization problem. The trajectory optimization problem is however fundamentally different in the sense that we can no longer probe the optimization space directly. We can only do so by applying stochastic policies to the system, then observe how the system evolves and infer an updated search distribution from these system *rollouts*.

Specifically, we are interested in locally linear Gaussian feedback policies of the following form, that we can apply to the system to artificiality inject the required stochasticity. Clearly this policy resembles the locally linear feedback trajectory as was describe in the previous paragraph and the idea can in that sense be understood as a sample based implementation of the iLQR or DDP algorithm.

$$\pi_{g,t}(u|x) = \mathcal{N}(u|x; \theta_{g,t}) = \mathcal{N}(u|k_{g,t} + \mathrm{K}_{g,t}x, \Sigma_{g,t})$$

The covariance determines the explorativeness of the policy and therefore the diverseness of the samples. Our goal is to actively shape it to stimulate exploration.

1.2. Path Integral Control

In the previous decade, a novel class of stochastic search algorithms was discovered that partially answer our question. This class is known as Path Integral Control (PIC). PIC is closely related to the Linearly Solvable Optimal Control (LSOC) prolem. LSOC is a restrictive subset of the Stochastic Optimal Control (SOC) framework (see Section 2) and is characterized by a number of remarkable properties (see Section 3.1) [22–24]. It was already pointed out that PIC and ES exhibit structural similarities [25,26]. This lead to the development of a PIC-CMA algorithm. This method deviates from the theory of LSOC and adapts the policy covariance in analogy with CMA-ES. This modification improved the convergence properties significantly yet ignores the underlying theory of LSOC. Other attempted generalizations include [27,28].

There has been keen interest in such algorithms as stochastic search algorithms may exhibit several advantages over gradient based algorithms [25,26,28]. So far it has been used in guided policy search to generate a set of prior optimal trajectories that were then used to fit a global policy [29] and is one of the two algorithms promoted by the *Lyceum* robot learning environment [30]. The use of PIC algorithms for real-time control applications was only recently considered as their execution is similar to Monte Carlo (MC) algorithms and were therefore thought to be too time critical to perform in real-time. However, with the rise of affordable GPUs and the ease of parallelization of MC based methods, it may become feasible in the near future to iterate dynamic stochastic search algorithms in real-time [31,32]. Nevertheless, practitioners of such methods have raised issues concerning the update of the covariance matrix [25,29,31,32]. It seems a mechanism is inherent to the existing framework that makes the covariance matrix vanish, compromising exploration. In other words, the search distribution collapses prematurely. Many authors suggested that the issue is limitedly understood and that it seems unlikely that it can be resolved with the theory at hand.

Furthermore, it is clear that ESs and PIC methods are closely related and that therefore PIC method may benefit from the rich body of work concerning ESs. However, as PIC are derived solely from the theory of LSOC, we argue that they are only limitedly understood and their similarity has been circumstantial. Recently, Williams et al. provide a novel derivation of the PI^2 method from an information-theoretic background [32]. Other authors have explored the relation between LSOC and information-theory [32–34], however these studies aimed for a physical connection and understanding.

In this paper, we venture on a different strategy and aim to identify a generalized set of PIC methods. We aim to describe an overarching optimization principle that leads us to derive to ESs in the context of static optimization and PIC methods in the context of dynamic optimization. Therefore, we approach the problem from an algorithmic point of view, rather than searching for a deeper physical interpretation or understanding. To establish the overarching framework, we identify the principle of entropic inference as a suitable setting to synthesize stochastic search algorithms and derive an entropic optimization framework from it. This will also allow us to derive a generalized set of PIC methods which are no longer limited to the LSOC setting and therefore do not inherit any of its inherent limitations. Furthermore, the mutual theoretical background paves way for a knowledge transfer from ES to PIC. Finally, this viewpoint provides us with a unique opportunity to relate PIC to existing Entropy Regularization paradigms in Reinforcement Learning [35–38].

1.3. Contributions

As our efforts span a rather extensive body of earlier work and aim to provide a holistic view of related subjects, we feel inclined to list our original contributions.

- We provide a comprehensive overview of existing Path Integral Control tailored to policy search. Here we point out a mutual underlying design principle which allows for a formal comparison and classification.
- We propose an original and intuitive argument for the introduction of entropy regularization terms in the context of optimization that is based on the principle of entropic inference.
- We untie the derivation of Path Integral Control methods from its historical roots in Linearly Solvable Optimal Control and illustrate that a similar set of algorithms can be derived directly from the more timely framework of Entropy Regularized Optimal Control. Therefore, we introduce the framework of Entropy Regularized Trajectory Optimization and derive the Entropy Regularized Path Integral Control (ERPIC) method. We consider this to be our primary contribution. Furthermore, this work elevates the structural similarity between Evolutionary Strategies, such as the CMA-ES [13] and NES [14], and PIC methods originally pointed out and exploited in [25], to a formal equivalence.
- We give a formal comparison of preceding PIC methods and ERPIC tailored to derivative-free trajectory optimization with control affine dynamics and locally linear Gaussian policies.

2. Preliminaries and Notation

Let us first introduce a number of relevant concepts and establish notation.

2.1. General Notation

We denote the set of continuous probability density functions over any closed set $\mathcal{X} \subseteq \mathbb{R}^n$ as $\mathcal{P}(\mathcal{X}) = \{p : \mathcal{X} \mapsto \mathbb{R}_{\geq 0} | \int_{\mathcal{X}} \pi(x) \mathrm{d}x = 1\}$. We will not always specify the argument as it is implied by the distribution definition. The statement $\pi \propto p$ with $\pi \in \mathcal{P}$ denotes that π is proportional to p up to a normalization constant. The expectation of f over a probability density function π is denoted as $\mathbb{E}_\pi[f] = \int_{\mathcal{X}} \pi(x) f(x) \mathrm{d}x$. We will also need a number of information theoretic measures.

The differential entropy is defined as

$$\mathbb{H}[\pi] = \mathbb{E}_\pi[-\log \pi] = -\int \log \pi(x) \pi(x) \mathrm{d}x$$

This is widely interpreted as a metric of uncertainty. However, more precisely it is to say that it is a measure for the amount of information left to be specified about some epistemological uncertain variable X for which we have represented our belief with a probability density function π [39].

The Kullback–Leibler divergence or relative entropy between two probability density functions π and ρ is defined as

$$\mathbb{D}[\pi \parallel \rho] = \mathbb{E}_\pi\left[\log \tfrac{\pi}{\rho}\right] = \int_{\mathcal{X}} \pi(x) \log \frac{\pi(x)}{\rho(x)} \mathrm{d}x \geq 0$$

The relative entropy is always positive but not symmetric $\mathbb{D}[\pi \parallel \rho] \neq \mathbb{D}[\rho \parallel \pi]$. An interpretation is given in Section 4.

2.2. Dynamic System Models

Further, we will consider controlled stationary deterministic or stochastic discrete-time systems. We use $x \in \mathcal{X}$ to represent the system state and $u \in \mathcal{U}$ to denote the input or control effort. Deterministic systems are modeled with a state-space function $x_{t+1} = f(x_t, u_t)$, $f : \mathcal{X} \times \mathcal{U} \mapsto \mathcal{X}$. Stochastic systems are modeled by a state transition distribution function $x_{t+1} \sim p(x_{t+1}|x_t, u_t)$, $p : \mathcal{X} \times \mathcal{X} \times \mathcal{U} \mapsto \mathbb{R}_{\geq 0}$. We also define two classes of non-stationary feedback policies. Deterministic policies are denoted as $u_t = u(t, x_t)$, $u : \mathbb{Z} \times \mathcal{X} \mapsto \mathcal{U}$, and stochastic policies are denoted as

$u_t \sim \pi(u_t|t, x_t)$, $\pi : \mathbb{Z} \times \mathcal{U} \times \mathcal{X} \mapsto \mathbb{R}_{\geq 0}$. Note that any deterministic dynamic system can also be represented as a stochastic system using the Dirac delta, $p(x'|x, u) = \delta(x' - f(x, u))$. Analogously, any deterministic policy can also be represented as a stochastic policy, $\pi(u_t|t, x_t) = \delta(u - u(t, x_t))$.

A realization of a system with policy u or π over a horizon T results into a trajectory $\tau = \{x_0, u_0, x_1, u_1, \ldots, x_{T-1}, u_{T-1}, x_T\} \in \mathcal{T} \subset \mathcal{S}_{1:T} \times \mathcal{A}_{0:T-1}$ or a state trajectory $\tau_x = \{x_0, x_1, \ldots, x_{T-1}, x_T\} \in \mathcal{T}_{\mathcal{X}} \subset \mathcal{S}_{1:T}$. In any case, we will assume that trajectories $\tau \in \mathcal{T}$ agree with the underlying system dynamics and is induced by a stochastic (or deterministic) policy.

Considering that we work within a stochastic setting, we can associate a probability to each trajectory. The trajectory distribution or path distribution induced by a policy π is denoted as $p(\tau|\pi)$. The state trajectory distribution induced by a policy π is denoted as $p_\pi(\tau_x)$. These distributions can be factorized in a product of transition probabilities.

$$p(\tau|\pi) = \prod_{t=0}^{T-1} p(x_{t+1}|x_t, u_t)\pi(u_t|t, x_t)$$
$$p_\pi(\tau_x) = \prod_{t=0}^{T-1} p_\pi(x_{t+1}|t, x_t)$$

Here, we defined the controlled state transition distribution. Note that the control effort is now entirely explicit.

$$p_\pi(x_{t+1}|t, x_t) = \int p(x_{t+1}|x_t, u_t)\pi(u_t|t, x_t)\mathrm{d}u_t$$

This definition also implies two state trajectory distributions of particular interest. The distribution p_u induced by a deterministic policy u and the uncontrolled or free trajectory distribution which we denote as p_0.

We can associate a cost to a trajectory $\tau \in \mathcal{T}$. In a general setting, we consider the cost $R : \mathcal{T} \mapsto \mathbb{R}$ that accumulates at a rate $r : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$.

$$R(\tau) = r_T(x_T) + \sum_{t=0}^{T-1} r_t(x_t, u_t)$$

Analogously, we can associate a cost $C : \mathcal{T}_{\mathcal{X}} \mapsto \mathbb{R}$ to state trajectories $\tau_x \in \mathcal{T}_{\mathcal{X}}$. Here, the control efforts is accounted for implicitly through the accumulation rate $c : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ that penalizes transitions $x \to x'$ (Note that we will often change between the use of time subscripts and the prime notation to indicate an increment in the time dimension)

$$C(\tau_x) = c_T(x_T) + \sum_{t=0}^{T-1} c_t(x_t, x_{t+1})$$

If there exists an inverse dynamic function f^{-1} so that $u = f^{-1}(x, x')$ when $x' = f(x, u)$, C and R are equivalent in the sense that $r_t(x, f^{-1}(x, x')) = c_t(x, x')$ and $r_T(x) = c_T(x)$. We emphasize here that the existence of such an inverse model is a mere formal assumption. In the practical setting to be presented, we will never actually have to invert the system dynamics given that we will have access to u.

Finally, we are also interested in the exotic cost function $L : \mathcal{T}_{\mathcal{X}} \mapsto \mathbb{R}$. The total cost can be decomposed as the superposition of a cost $\tilde{L}$ that accumulates with a solely state dependent rate $l : \mathcal{X} \mapsto \mathbb{R}$ and a term that accounts for the control effort. Specifically we are interested in the situation where the control efforts are penalized indirectly by introducing the logarithm of the ratio between the controlled and free state transition probabilities. This specific choice will have a remarkable technical consequence later on. Further, note that L is therefore of the same type as C and operates on $\mathcal{T}_{\mathcal{X}}$. It follows that if the transition $x \to x'$ is likelier when the system is controlled, i.e., $p_u(x'|x) > p(x'|x)$ the ratio is larger than 1 and so its logarithm is positive, effectively inducing a cost. Alternatively, when the transition becomes less probable, this induces a negative penalty, again inducing a cost. As a result it will depend on $\tilde{L}$ whether the use of p_u over p_0 is beneficial or not.

$$L(\tau_x) = \underbrace{l_T(x_T) + \sum_{t=0}^{T-1} l_t(x_t)}_{\check{L}(\tau_x)} + \sum_{t=0}^{T-1} \log \frac{p_u(x_{t+1}|t, x_t)}{p_0(x_{t+1}|t, x_t)}$$

2.3. Stochastic Optimal Control

We are interested in finding policies that minimize the expected induced cost. Hence we address the Stochastic Optimal Control (SOC) problem. Such policies are known to be stabilizing and are capable of encoding very rich and complex dynamic behavior even for considerably simple cost formulations.

$$u^* = \arg\min_u \mathbb{E}_{p(\tau|u)}[R(\tau)] \tag{1}$$

Solving (1) poses a so-called dynamic optimization problem. The prefix dynamic emphasizes that the optimization variables are constrained by a causal structure which allows to break the problem apart into several subproblems that can be solved recursively. This problem property is also known as *optimal substructure* which can be exploited by dedicated solution algorithms.

Correspondingly, the problem above can also be represented with the recursive Bellman equation. Here, $V_t : \mathcal{X}_t \mapsto \mathbb{R}$ represents the value function or optimal cost-to-go, i.e., the cost that is accumulated if we initialize the system in state x at time t and control it optimally until the horizon T.

$$V(t, x) = \min_\pi \int \pi(u|t, x) \left(r_t(x, u) + \int V(t+1, x')p(x'|t, x, u)\mathrm{d}x' \right) \mathrm{d}u$$

We emphasize that we presume the policy to be stochastic. However, unless there is the incentive to maintain a stochastic policy for the purpose of exploration, the expression minimizes for a deterministic policy so that the expectation over the policy can be omitted [40]. This redundant form is however appealing for later comparison. Further, note that if also $p(x'|x, u) = \delta(x' - f(x, u))$, this problem reduces to deterministic optimal control.

$$V(t, x) = \min_u r_t(x, u) + \int V(t+1, x')p(x'|x, u)\mathrm{d}x' \tag{2}$$

Throughout this manuscript we will further assume that the initial state is fixed and known exactly, i.e., $x(0) = x_0$. We could stress this formally with our notation and exemplify that any function associated to problem (2), be it $u(t, x)$ $\pi(u|t, x)$ or $V(t, x)$ or any of the implied functions $p_\pi(x'|t, x)$ and $p_\pi(\tau_x)$ are conditional on x_0. In order not to overload the notational burden we simply assume this to be true throughout the paper. This assumption will allow us to concentrate on local solutions or so called trajectory optimizers.

2.4. Local Parametric Policies

The algorithms we present here do not aim to retrieve an exact solution to the problem above, but instead operate on a restricted class of parametrized policies $\pi(u|t, x, \theta)$ where θ refers to the policy parameter. Considering that we are interested in trajectory optimizers, we restrict our focus to locally linear Gaussian policies of the form $\pi(u|t, x, \theta) = \mathcal{N}(u|u_t + K_t x, \Sigma_t)$, so that $\theta = \{\theta_t\}_{t=0}^T$ with $\theta_t = \{u_t, K_t, \Sigma_t\}$. Here K_t represented a feedback gain matrix and Σ_t the covariance matrix which controls the stochasticity and therewith the explorative incentive of the policy. This parametrization allows to approximate the exact solutions in the neighborhood of an optimal open-loop trajectory initialized at for example $x(0) = x_0$.

3. Path Integral Control

In this section we give a brief introduction to the theory of LSOC and give an overview of existing PIC methods rooting directly from it. In almost all preceding contributions (apart from those following [22]), LSOC is addressed in a continuous time setting and a discretization is only performed afterwards in order to derive practical methods. Here we will directly address the LSOC problem in a discrete time setting as it better suits our requirements and as such also avoid the tedious technicalities involved with the discretization.

Second, we give a comprehensive overview of the class of PIC methods. As far as we are aware of, this is the first time such a formal comparison is made, revealing an overarching design principle shared by all algorithmic formulations. The identification of this design principle makes it easier to pinpoint the core prerequisite of any PIC method and will allow us to introduce a generalized set of PIC methods in Section 4.3.

3.1. Discrete Time Linearly Solvable Optimal Control

Formally, the discrete time LSOC framework refers to the following SOC problem,

$$V(t,x) = \min_u l_t(x) + \int p(x'|t,x,u)\left(\log\frac{p_u(x'|t,x)}{p_0(x'|t,x)} + V(t+1,x')\right)dx' \tag{3}$$

$$= \min_u l_t(x) + \mathbb{D}\left[p_u \parallel p_0\right] + \mathbb{E}_{p_u}[V(t+1,x')]$$

We emphasize that the control effort is penalized implicitly through the Kullback–Leibler divergence between the controlled and free state transition probabilities. For a formal motivation for this particular penalty formulation, we refer the reader to the work in [22,33] and recall the intuitive justification that was already given in Section 2. Here, we are mostly interested in the technical implications that are associated with this problem formulation.

Problem (3) implies the existence of an optimal policy $u^*_{\text{LSOC}}(t,x)$

$$u^*_{\text{LSOC}}(t,x) = \arg\min_u l_t(x) + \mathbb{D}\left[p_u \parallel p_0\right] + \mathbb{E}_{p_u}[V(t+1,x')]$$

In general the problem can not be solved exactly for $u^*_{\text{LSOC}}(t,x)$. However, the following theorem establishes a relation between the optimal policy $u^*_{\text{LSOC}}(t,x)$, the optimal state transition distribution $p_{u^*_{\text{LSOC}}}(x'|t,x)$ and the optimal state trajectory distribution $p_{u^*_{\text{LSOC}}}(\tau_x)$, and summarizes the most profound results in the context of LSOC.

Theorem 1. *With $V(t,x)$ defined as in (3), the following problems are equivalent,*

$$u^*_{LSOC}(t,x) = \arg\min_u l_t(x) + \mathbb{D}\left[p_u \parallel p_0\right] + \mathbb{E}_{p_u}[V(t+1,x')] \tag{4}$$

$$p_{u^*_{LSOC}}(x'|t,x) = \arg\min_{p_u \in \mathcal{P}} l_t(x) + \mathbb{D}\left[p_u \parallel p_0\right] + \mathbb{E}_{p_u}[V(t+1,x')] \tag{5}$$

$$p_{u^*_{LSOC}}(\tau_x) = \arg\min_{p_u \in \mathcal{P}} \mathbb{E}_{p_u}[L] = \mathbb{E}_{p_u}[\tilde{L}] + \mathbb{D}\left[p_u \parallel p_0\right] \tag{6}$$

The latter can be solved explicitly

$$p_{u^*_{LSOC}}(x'|t,x) \propto p_0(x'|t,x)e^{-V(t+1,x')}$$

$$p_{u^*_{LSOC}}(\tau_x) \propto p_0(\tau_x)e^{-\tilde{L}(\tau_x)}$$

where $V(t,x)$ is governed by the recursion

$$V(t,x) = l_t(x) + \log\int p_0(x'|t,x)e^{-V(t+1,x')}dx'$$

and where $u^*_{LSOC}(t, x)$ *and* $p_{u^*_{LSOC}}(x'|t, x)$ *are related as*

$$p_{u^*_{LSOC}}(x'|t, x) = p(x'|x, u^*_{LSOC}(t, x))$$

The proof relies on the calculus of variations. These peculiar results follow directly from the fact that the problem does no longer depend on u explicitly. For details we refer the reader to earlier references.

Most remarkable is that the equivalence of problems (4) and (5) implies that we can lift the optimization problem from the control space to the state transition distribution space. Second, although we can not solve to $u^*_{LSOC}(t, x)$ directly, we can solve for the induced optimal state transition distribution, $p_{u^*_{LSOC}}(x'|t, x)$. Note that therefore we should still identify $V(t, x)$. The equivalence of (5) and (6) follows from the recursive definition of $V(t, x)$ that can be evaluated exactly, and thus implies the existence of an explicit optimal state trajectory distribution, $p_{u^*_{LSOC}}(\tau_x)$ also. This is a unique trait of the LSOC setting and lies at the very root of every PIC method. The reason why there exists an explicit solution is a direct consequence from the control effort penalization which makes it possible to lift the optimization problem. As an unfortunate side effect, the system uncertainty and the control penalization are now somehow entangled. This we identify as a fundamental limitation of the LSOC setting and related methods.

A final remark can be made with respect to control affine system dynamics, $x' = a(x) + B(x)u$, disturbed by unbiased Gaussian noise $\xi \sim \mathcal{N}(\xi|0, \Sigma)$ in the control subspace. In this case, it is possible to derive an explicit expression for the optimal control policy $u^*_{LSOC}(t, x)$. The controlled transition distribution is then given by $p(x'|x, u) = \mathcal{N}(x'|a(x) + B(x)u, B(x)\Sigma B(x)^\top)$ and the control policy $u^*_{LSOC}(t, x)$ can be extracted by comparing the expected value of x' based on either distribution $p(x'|x, u^*_{LSOC}(t, x))$ or $p_{u^*_{LSOC}}(x'|t, x)$.

It is easily verified that

$$u^*_{LSOC}(t, x) = \frac{\mathbb{E}_{p_0(x'|t,x)}\left[\xi \cdot e^{-V(t+1, x')}\right]}{\mathbb{E}_{p_0(x'|t,x)}\left[e^{-V(t+1, x')}\right]} = \frac{\mathbb{E}_{p_0(\tau_x|t,x)}\left[\xi \cdot e^{-\tilde{L}(\tau_x|t+1, x')}\right]}{\mathbb{E}_{p_0(x'|t,x)}\left[e^{-\tilde{L}(\tau_x|t+1, x')}\right]} \tag{7}$$

This is the discrete time version of the property given in for example [41] Theorem 1, and clearly illustrates the *path integral* terminology. Moreover, it follows that the optimal policy can be estimated with Monte Carlo sampling from the free system dynamics making it appealing to use directly in an MPC setting [31,32,42]. However, as noted in the introduction we are mostly interested in its applications as a trajectory optimization method.

In the following section, we detail a number of policy search methods that effectively exploit this peculiar setting. All of them would rely on the stochastic system dynamics to explore the solution space.

3.2. Path Integral Control (PIC) Methods

The unique traits of the LSOC framework have been exploited to derive a class of so called PIC methods. The interested reader is referred to earlier references [26–28,30,32,41,43–47]. An overview of applications was already given in the introduction. Originally, the optimal policy was estimated directly from (7) using Monte Carlo samples generated with the free system, i.e., p_0. Other methods evolved beyond that idea but the basic principle remained the same. In any case the goal is to find the optimal policy $u^*_{LSOC}(t, x)$ relying on the inherent stochasticity of the system. To be able to give a concise overview, we distilled a generic design principle that allows to derive different PIC methods. Note that the authors have identified this principle and that the derivations included in the references may not explicitly state this concept.

The principle is based on two distributions. We assume the existence of a formal but explicit goal state trajectory distribution, say $p_{u^\star}$, and, use a parametrized path distribution, $p_{u(\theta)}$, induced by some parametric policy, $u(\theta)$.

It is then possible to infer the optimal policy by projecting the parametrized distribution $p_{u(\theta)}$ onto the optimal distribution $p_{u^\star}$, and, determine the corresponding optimal policy parameter by minimizing the projection distance. Based on the cross-entropy argument originally made by [41], the Kullback–Leibler divergence is used as a projection operator.

The following optimization problem is addressed.

$$\min_{\theta} \mathbb{D} \left[p_{u^\star} \parallel p_{u(\theta)} \right] = -\mathbb{E}_{p_u^\star}[\log p_{u(\theta)}] + \text{constant}$$

$$= -J(\theta|p_{u^\star}) + \text{constant}$$

Provided that we can sample from the distribution $p_{u^\star}$, this expectation can be approximated using a Monte Carlo estimator. Assuming that we can evaluate but not sample from $p_{u^\star}$, but dispose of a distribution p_{u_g} that we can both evaluate and sample from, another estimate can be obtained by the idea of importance sampling. Such a distribution is easily constructed by applying some policy u_g to the system. Particularly interesting is the setting where u_g is somehow informed about the desired distribution $p_{u^\star}$. We denote this objective with subscript g to emphasize the use of the guiding policy u_g.

$$J_g(\theta|p_{u^\star}) = \mathbb{E}_{p_{u_g}} \left[\frac{p_{u^\star}}{p_{u_g}} \cdot \log p_{u(\theta)} \right]$$

We introduce the Monte Carlo estimator $\hat{J}_g(\theta|p_{u^\star})$ for the formal objective defined above. As a guiding policy we substitute $u(\theta_g)$ for u_g

$$J_g(\theta|p_{u^\star}) \approx \hat{J}_g(\theta|p_{u^\star}) = \frac{1}{M} \sum_{k=1}^{M} \frac{p_{u^\star}(\tau_x^k)}{p_{u(\theta_g)}(\tau_x^k)} \cdot \log p_{u(\theta)}(\tau_x^k) \tag{8}$$

This framework admits to define an iterative strategy where θ_{g+1} is found by optimizing the objective $\hat{J}_g(\theta|p_{u^\star})$. As the existence of an explicit path distribution was so far a unique trait of the LSOC setting, the derivation of PIC methods has always been tied to the framework.

Depending on the optimal distribution substituted for $p_{u^\star}$, the parametrization of the policy $u(\theta)$ and the strategy used to solve the optimization problem, different PIC methods are obtained that can be useful to tackle different problems. As already stated in the introduction we are interested in its application in the context of trajectory optimization, i.e., finding local optimal control solutions for a fixed given initial state. Therefore, we will be interested in local policy parametrizations.

Regarding the optimization strategy we can make a distinction between two approaches, denoted as exact methods and gradient descent methods.

3.2.1. Exact Methods

For a particular subset of policy representations in combination with a control affine system model, the objective in (8) can be maximized exactly. If these conditions are met, the exact solution θ^* can be substituted for the next guiding policy parameter θ_{g+1}. We refer to these approaches as exact methods. Such are particularly interesting to find local policies. We refer to Section 5 for details.

$$\theta_{g+1} = \arg\max_{\theta} \hat{J}_g(\theta|p_{u^\star})$$

Regarding the choice for the goal distribution $p_{u^\star}$ we can make an additional distinction between the Sample Efficient Path Integral Control (SEPIC) and Path Integral Relative Entropy Search (PIREPS) method.

Sample Efficient Path Integral Control Method

Most convenient is to substitute the desired optimal path distribution $p_{u^*_{\mathrm{LSOC}}}$ for $p_{u^\star}$

$$p_{u^\star} = p_{u^*_{\mathrm{LSOC}}}$$

This strategy boils down to the one proposed by Williams et al. [30,32,44] and partially with the one proposed in [48] (In [48], the particular estimation of the system dynamics allows for a more general solution in terms of Equation (7). However this is out of scope here as we are interested in local policy parametrizations). The terminology stems from the fact that the sampling is done using the informed distribution p_{u_g} as a guiding post. Furthermore, if the parametric policy is sufficiently expressive to cover the optimal policy u^*_{LSOC}, this methods converges to the actual solution of the LSOC framework. The approximate objective is evaluated as

$$\hat{J}_g(\theta|p_{u^*_{\mathrm{LSOC}}}) = \frac{1}{M} \sum_{k=1}^{M} p_{u(\theta_g)}(\tau_x^k)^{-1} \cdot p_0(\tau_x^k) \cdot e^{-\tilde{L}(\tau_x^k)} \cdot \log p_{u(\theta)}(\tau_x^k) \tag{9}$$

SEPIC aims to solve for $p_{u^*_{\mathrm{LSOC}}}$ directly. The problem is that if the original guiding policy is too far removed from the optimal policy, the method will not converge. Because the interesting regions in the solution space are simply not sampled and the optimal policy is never discovered.

Path Integral Relative Entropy Policy Search

To remedy this issue [47] proposed PIREPS.

Retrospectively, this is an information-theoretic trust-region strategy which introduces a regularization term penalizing the Kullback–Leibler divergence between the new and old path distribution to promote more conservative policy updates. In fact, this is a form of entropy regularization which we will come back to in Section 4. This idea was introduced independently from [47] in [29] to generate local policies in the context of guided policy search.

$$p_{u_{g+1}} = \arg\min_{p_u \in \mathcal{P}} \mathbb{D}\left[p_u \parallel p_{u^*_{\mathrm{LSOC}}}\right] + \lambda \mathbb{D}\left[p_u \parallel p_{u_g}\right]$$

The approach generates a sequence of intermediate pseudo-optimal path probabilities that aim to improve the convergence properties. This problem can be solved explicitly and the intermediate probabilities are substituted for $p_{u^\star}$ consequently. The idea is that the distribution $p_{u_{g+1}}$ will be closer to the solution space sampled by $p_{u(\theta_g)}$ and that as a direct result the policy updates will be more robust. Note that for $g \to \infty$ the sequence of pseudo-optimal path probabilities collapses on $p_{u^*_{\mathrm{LSOC}}}$.

$$p_{u^\star} = p_{u_{g+1}} \propto p_{u_g}^{\frac{\lambda}{1+\lambda}} \cdot p_0^{\frac{1}{1+\lambda}} \cdot e^{-\frac{1}{1+\lambda}\tilde{L}} = p_{u_0}^{\left(\frac{\lambda}{1+\lambda}\right)^{g+1}} \cdot p_{u^*_{\mathrm{LSOC}}}^{1-\left(\frac{\lambda}{1+\lambda}\right)^{g+1}}$$

The approximate objective is evaluated as

$$\hat{J}_g(\theta|p_{u_{g+1}}) = \frac{1}{M} \sum_{k=1}^{M} p_{u(\theta_g)}^{\frac{-1}{1+\lambda}}(\tau_x^k) \cdot p_0^{\frac{1}{1+\lambda}}(\tau_x^k) \cdot e^{-\frac{1}{1+\lambda}\tilde{L}(\tau_x^k)} \cdot \log p_{u(\theta)}(\tau_x^k) \tag{10}$$

3.2.2. Gradient Ascent Methods

Alternatively one can try maximize the objective estimator $\hat{J}(\theta|p_{u^\star})$ using a gradient ascent method. Such a strategy is suited for general policy parametrizations that aim to find a global approximation of $u^*_{\mathrm{LSOC}}(t,x)$.

$$\theta_{g+1} = \theta_g + \eta \cdot \nabla_\theta \hat{J}_g(\theta|p_{u^\star})$$

Again regarding the choice of the goal distribution $p_{u^\star}$ either the Path Integral Cross Entropy (PICE) or the Adaptive Smoothing Path Integral Control (ASPIC) method are obtained.

Path Integral Cross Entropy Method

PICE can be considered the gradient ascent version of SEPIC and was proposed in [41].

$$p_{u^\star} = p_{u^*_{\text{LSOC}}}$$

Adaptive Smoothing Path Integral Control (ASPIC) Method

ASPIC can be considered the gradient ascent version of PIREPS and was proposed more recently in [49].

$$p_{u_{g+1}} = \arg \min_{p_u \in \mathcal{P}} \mathbb{E}_{p_u}[L] + \lambda \mathbb{D}\left[p_u \parallel p_{u_g}\right] p_{u^\star} \qquad = p_{u_{g+1}} \propto p_{u_g}^{\frac{\lambda}{1+\lambda}} \cdot p_0^{\frac{1}{1+\lambda}} \cdot e^{-\frac{1}{1+\lambda}\tilde{L}}$$

3.3. Other Noteworthy PIC Methods

An important subset of PIC methods, or methods that are associated to the framework of LSOC at least, are Path Integral Policy Improvement algorithms or PI^2. These methods are hard to classify as they are somewhere in between Evolutionary Search methods and policy optimization methods and rely on a heuristic temporal averaging strategy to resolve the conflicting policy parameter update schemes. The most important members of this class are PI^2 [45], PI^2 with Covariance Matrix Adaptation [25] and PI^2 with Population Adaptation [50]. The authors of PI^2-CMA, were the first to pursue the structural equivalence between Evolutionary Strategies and PIC methods. Based on this equivalence they proposed to adapt the policy covariance in correspondence to the CMA-ES algorithm, improving the convergence but destroying the underlying assumption of LSOC. This idea was successfully repeated in the context of trajectory optimization by [29].

3.4. Other Remarks

An issue of widespread concern in RL is how to shape a deliberately stochastic policy to obtain an explorative incentive without compromising safety measures or becoming risk seeking by unfortunate coincidence. PIC-based methods were one of the first strategies that somewhat address this issue with the implied connection between the uncertainty and cost. Especially, in case of deterministic systems where a deliberate stochastic policy could be introduced. Unfortunately the underlying theory of LSOC enforces an inverse proportionality between the control noise magnitude and the control penalty and moreover requires the control to be penalized quadratically (see later). Although justifiable from a control engineering perspective, this also poses severe practical limitations. Nowadays, entropy regularization seems to be a fruitful resolution to address the issue of stochastic policy shaping in a less restrictive or predetermined fashion. As for now it is not really clear how PIC methods relate to the framework of entropy regularized RL and whether they can benefit from recent advances made by this community. We shall address this question in the following section.

4. Entropy Regularized Path Integral Control

The main purpose of this section is to identify a novel SOC problem that can be solved explicitly yielding a formal optimal state trajectory distribution. The problem formulation is in that sense similar to the LSOC setting, yet it addresses a less restricted set of optimal control problems. Based on the general design principle for PIC methods discussed in Section 3.2, this optimal distribution can be substituted for the goal distribution $p_{u^\star}$ hinting at a generalized set of PIC methods which will be investigate structurally for the purpose of model-free trajectory optimization in Section 5.

Our problem formulation relies heavily on the concept of entropy regularization. Entropy regularization is a setting of widespread use in the context of RL nowadays, yet it is less studied in the context of (static) stochastic optimization or stochastic search, at least in the way that we will treat it.

As we will show our SOC framework shares properties with static stochastic optimization and therefore the results that we derive in the latter setting will also hold in the context that enjoys our interest. Furthermore, we note that our results related to static stochastic optimization may be of interest to the Evolutionary Search community. Vice versa, the methods in Section 5 may benefit from prevailing ideas in the stochastic optimization community in order to deal with the problem of sample efficiency which is still considered to be a major challenge by the RL community. An example of such a strategy that is potentially interesting is importance mixing [51,52].

4.1. Entropy Regularized Optimization

There exists a large body of work that addresses the relation between inference and control. A lesser amount of work investigates the relation between inference and optimization. In this brief section, we provide an original and convincing argument to introduce an entropic regularization term into formal stochastic optimization problems that does not rely on an information-theoretic but on a strictly inference related argument. Although the resulting framework and associated distributions are known, our justification is original and in our opinion more intuitive. The concept results into a distribution sequence which exhibits a number of interesting properties and allows to make formal statements about convergence rates of derived practical search methods. As it will turn out these properties also directly apply to the entropy regularized optimal control problem that we will introduce in Section 4.3. To appreciate the argument, we must give an introduction to entropic inference.

4.1.1. Entropic Inference

Inductive inference refers to the problem of how a rational agent should update its state of knowledge, or so-called belief, about some quantity when new information about that quantity becomes available. Beliefs about the quantity $x \in \mathcal{X} \subset \mathbb{R}^n$ are modeled as distributions. An inference procedure refers to the computational framework that establishes how to integrate new information with information held by any prior belief, say ρ, to determine an informed posterior, say π, consistently.

A well-known framework is that of Bayesian inference which allows to process new information in the form of data. A lesser known inference framework deals with the setting where information is available in the form of a constraint on the family of acceptable posteriors [53]. Specifically constraints in the form of the expected value of some function $f : \mathcal{X} \to \mathbb{R}^n$, i.e., $\mathbb{E}[f] = \mu$. Consequently we can focus our attention to the subset of distributions that agree with it $\mathcal{C}_f(\mu) := \{\pi \in \mathcal{P} : \mathbb{E}_\pi[f] = \mu\}$. Any $\pi \in \mathcal{C}_f(\mu)$ that satisfies the information constraint qualifies as a potential posterior belief. The challenge thus reduces to identifying a unique posterior from among all those that could give rise to the constraint. The solution is to establish a ranking on the set $\mathcal{C}_f(\mu)$ by determining a functional $\mathcal{F}$ that associates a value to any posterior π relative to a given prior ρ. The inference procedure is then effectively cast into an optimization problem.

$$\pi^* = \arg \min_{\pi \in \mathcal{C}_f(\mu)} \mathcal{F}[\pi, \rho]$$

The problem remains in finding a suitable and meaningful functional $\mathcal{F}$. The measure should promote a distribution that is maximally unbiased (i.e., maximally ignorant) about all the information we do not possess. This principle is referred to as *minimal updating*. This problem setting roots back at least to the maximum entropy principle first introduced by Jaynes [54,55]. As many authors provided compelling theoretical arguments for the relative entropy as the only consistent measure [39,56–60], here it is important to emphasize that no interpretation of the measure is given. It simply is the only measure that agrees with the axioms of minimal updating.

The following variational problem determines the framework of entropic inference.

$$\pi^* = \arg\min_{\pi \in \mathcal{P}} \mathbb{D}\left[\pi \;\|\; \rho\right] \text{ s.t. } \mathbb{E}_\pi[f] = \mu \tag{11}$$

We now possess of a rational argument as to why an entropy regularization term is added to any problem and what its effect will be on the solution.

4.1.2. Optimization as an Inference Problem

Let us now argue how the principle of entropic inference can be practiced to serve the purpose of static or classic optimization. For convenience we assume that the objective $f : \mathcal{X} \subset \mathbb{R}^n \mapsto \mathbb{R}$ has a unique global minimum.

$$x^* = \arg\min_{x \in \mathcal{X}} f(x)$$

Assume we can model any beliefs we have about the solution x^* by some prior distribution, say $\rho \in \mathcal{P}$. Second, instead of supposing information in the form of the expected value of the objective f, here we only require that the expected value with respect to the posterior, π, is, some amount $\Delta > 0$, smaller than the expected value is with respect to the prior. In this fashion, we change our prior belief about the optimal solution but only to the minimal extent required to decrease the expectation taken over f with some arbitrary value Δ. Put differently, we obtain a posterior that makes least claim to being informed about the optimal solution beyond the stated lower limit on the expectation. This idea can be formalized accordingly

$$\min_{\pi \in \mathcal{P}} \mathbb{D}\left[\pi \;\|\; \rho\right] \text{ s.t. } \mathbb{E}_\pi[f] + \Delta \le \mathbb{E}_\rho[f] \tag{12}$$

As the relative entropy minimizes for $\pi = \rho$, it follows that the inequality tightens into an equality. Nonetheless, we should be careful when we pick a value for $\Delta > 0$ that respects the bound $\Delta \le \mathbb{E}_\rho[f] - f^*$. This is a practical concern that does not interfere with what we wish to accomplish, which is to construct a minimal update procedure that we can practice to serve the purpose of optimization. It suffices to solve the problem above for π using variational calculus. This yields the following entropic update rule where $\lambda > 0$ denotes the Lagrangian multiplier associated with the inequality constraint. This update principle is well known and can be subjected to an interesting interpretation (The posterior distribution, π, is equal to the prior distribution, ρ, multiplied with a cost driven probability shift, $e^{-\lambda f}$, that makes rewarding regions more probable, resembling the concept of Bayesian inference. The transformation $p(f) \propto e^{-\lambda f}$ maps costs to probabilities. Indeed one may recognize the inverse log-likelihood transformation from probability to cost as it is often used in the context of Bayesian inference) However, as far we are aware of, it has never been derived from the theory of entropic inference or thus minimal updating, which makes it possible to appreciate it in a more general light.

$$\pi \propto \rho \cdot e^{-\lambda f}$$

The exact value of λ can be determined by solving the dual optimization problem. Alternatively, we could also pick any $\lambda > 0$ without the risk of overshooting the constraint $\Delta \le \mathbb{E}_\rho[f] - f^*$. We will show that when $\lambda \to \infty$, the expectation $\mathbb{E}_\pi[f]$ collapses on the exact solution f^*. Therefore, $\frac{1}{\lambda}$ reduces to a *temperature* like quantity that determines the amount of information that is added to ρ, where in the limit exactly so much information is admitted to precisely determine f^* and thus x^*.

The same strategy is obtained by considering the entropy regularized stochastic optimization problem below. Here an information-theoretic trust-region is introduced to promote conservative distribution updates but apart from the fact that the problem can be solved exactly, there is no proper motivation, at least not in the sense of minimal updating. This strategy has been adopted by previous authors from the optimization community [11,61].

$$\min_{\pi \in \mathcal{P}} \mathbb{E}_\pi[f] \text{ s.t. } \mathbb{D}\left[\pi \parallel \rho\right] \leq \Delta$$

Again it is easier to choose a suitable value for the implied Lagrangian multiplier $\lambda > 0$ than it is for $\Delta > 0$. The problem above is therefore often relaxed using a penalty function.

$$\pi^* = \arg\min_{\pi \in \mathcal{P}} \mathbb{E}_\pi[f] + \lambda\mathbb{D}\left[\pi \parallel \rho\right] \propto \rho \cdot e^{-\frac{1}{\lambda}f} \tag{13}$$

4.1.3. Theoretical Search Distribution Sequences

The entropic updating procedure suggested in (13) can be solved exactly and implies a theoretical distribution sequence, substituting the previous posterior for the next prior.

$$\pi_{g+1} \propto \pi_g \cdot e^{-\frac{1}{\lambda}f} = \min_{\pi \in \mathcal{P}} \mathbb{E}_\pi[f] + \lambda\mathbb{D}\left[\pi \parallel \pi_g\right] \tag{14}$$

$$\pi_g \propto \pi_0 \cdot e^{-\frac{g}{\lambda}f} \tag{15}$$

The update mechanism in (14) can be used as a theoretical model to shape practical search distribution sequences that can be used to solve the underlying optimization problem. In practice one seeks algorithms that estimate the posterior distribution π_{g+1} from samples taken with the prior π_g. The theory now implies that the sequence will get gradually more informed about the optimum.

As far as we are aware of, the properties of the theoretical distribution sequence in (15) have not been studied before. In the following Theorem we summarize some of its properties. For the proof we refer to the work in Appendix A.

Theorem 2. *Assume that, without loss of generality, objective f attains a unique global minimum at the origin. Further define the sequence of search distributions $\{\pi_g\}$ so that $\pi_g \propto \pi_0 \cdot e^{-gf}$. Then, it holds that*

- *the distribution sequence $\{\pi_g\}$ collapses in the limit on the Dirac delta distribution in the sense that*

$$\lim_{g \to \infty} \pi_g = \begin{cases} \infty, & x = x^* \\ 0, & x \neq x^* \end{cases} \tag{16}$$

- *the function $\mathbb{E}_{\pi_g}[f]$ of g is monotonically decreasing regardless of π_0*

$$\mathbb{E}_{\pi_{g+1}}[f] < \mathbb{E}_{\pi_g}[f] \tag{17}$$

- *the function $\mathbb{H}\left[\pi_g\right]$ of g is monotonically decreasing if π_0 is chosen uniform on $\mathcal{X}$*

$$\mathbb{H}\left[\pi_{g+1}\right] < \mathbb{H}\left[\pi_g\right] \tag{18}$$

The property (17) is suggested by the problem definition in (12). It follows that the the sequence $\{\mathbb{E}_{\pi_g}[f]\}$ converges monotonically to f^*, implying that $\{\pi_g\}$ should converge to x^* (16). If we thus construct a stochastic optimization algorithm that maintains a sequence of search distributions governed by (14), the algorithm should converge monotonically to the optimal solution.

This sequence of search distributions is increasingly more informed about the solution and as a result its entropy content deteriorates as the sequence converges to the minimum (see Appendix A).

Properties (16) and (18) thus imply that the entropy slowly evaporates. This is an important observation indicating that the explorative incentive of the search distribution deteriorates for increasing g. Therefore, when we would use this sequence to shape a practical search algorithm, this property suggest that the entropy of the sample population will slowly diminish. As a result of the finiteness of the population it will become increasingly more likely that the correct distribution parameters cannot be inferred from the population so that the sequence collapses prematurely.

In order to stimulate the sequence to preserve an explorative incentive, i.e., manage the entropic content, problem (13) can be relaxed using an entropy term ($\gamma > 0$), which will force the distribution to maintain a finite entropy.

$$\pi_{g+1} = \arg\min_{\pi \in \mathcal{P}} \mathbb{E}_\pi[f] + \lambda \mathbb{D}\left[\pi \parallel \pi_g\right] - \gamma \mathbb{H}\left[\pi\right] \tag{19}$$

The entropy regularized problem above implies the following distribution sequence. This is easily verified relying on the calculus of variations.

$$\pi_{g+1} \propto \pi_g^{\frac{\lambda}{\lambda+\gamma}} \cdot e^{-\frac{1}{\lambda+\gamma}f} \tag{20}$$

$$\pi_g \propto \pi_0^{\left(\frac{\lambda}{\lambda+\gamma}\right)^g} \cdot e^{-\frac{1}{\gamma}\left(1-\left(\frac{\lambda}{\lambda+\gamma}\right)^g\right)f} \propto \pi_0^{\left(\frac{\lambda}{\lambda+\gamma}\right)^g} \cdot \pi_\infty^{1-\left(\frac{\lambda}{\lambda+\gamma}\right)^g} \tag{21}$$

As opposed to (15), sequence (21) does not collapse on the minimum, but instead converges to $\pi_\infty \propto e^{-\frac{1}{\gamma}f}$. The latter is equal to the exact solution of the entropy relaxed optimization problem defined below.

$$\pi_\infty = \arg\min_{\pi \in \mathcal{P}} \mathbb{E}_\pi[f] - \gamma \mathbb{H}\left[\pi\right]$$

The derivative of the implied functions $\mathbb{E}_{\pi_g}[f]$ and $\mathbb{H}\left[\pi_g\right]$ are given by

$$\frac{\mathrm{d}}{\mathrm{d}g} \mathbb{E}_{\pi_g}[f] = -\gamma\alpha \log\left(\frac{\lambda+\gamma}{\lambda}\right) \operatorname{Cov}_{\pi_g}\left[\log \pi_\infty, \log \frac{\pi_\infty}{\pi_0}\right] \tag{22}$$

$$\frac{\mathrm{d}}{\mathrm{d}g} \mathbb{H}\left[\pi_g\right] = -\alpha \log\left(\frac{\lambda+\gamma}{\lambda}\right) \left(\operatorname{Cov}_{\pi_g}\left[\log \pi_\infty, \log \frac{\pi_\infty}{\pi_0}\right] - \alpha\operatorname{Var}_{\pi_g}\left[\log \frac{\pi_\infty}{\pi_0}\right]\right) \tag{23}$$

where $\alpha(g) = (\frac{\lambda}{\lambda+\gamma})^g > 0$ so that $\lim_{g \to \infty} \alpha(g) = 0$.

The convergence rate of the implied function $\mathbb{E}_{\pi_g}[f]$ now also depends on the interaction between f and π_0. If the initial distribution π_0 is taken to be uniform on $\mathcal{X}$, so that $\operatorname{cov}_{\pi_g}\left[\pi_0, f\right] = 0$, the expected value will decrease monotonically. As opposed to (18) the rate will stall for large g. Derivations are provided in Appendix B.

In Section 4.3, we introduce a stochastic optimal control problem that is characterized by a similar distribution sequence. Therefore it inherits the properties described above. The usefulness of these theoretical distributions, is illustrated in Appendix C where we derive a stochastic search method.

Next, let us return to our original problem and briefly review entropy regularized SOC and its usefulness to RL as a stepping stone to Section 4.3.

4.2. Entropy Regularized Optimal Control

As we already explained in the introduction, there is no reason to maintain a stochastic policy in the generic SOC setting (2). On the other hand, a stochastic policy can be of interest to derive policy updating algorithms that depend on stochasticity for improved exploration rather than to rely on the stochastic system dynamics only.

As was justified rigorously in the previous section this can be achieved by introducing an entropy relaxation term in the optimization objective. Such a term actively stimulates a stochastic policy. Nowadays, this is a setting of widespread use in the RL community and was explored extensively by Levine et al. particularly in the context of Actor-Critic RL algorithms [35,62]. In recent literature, this problem has been regularized successfully using a relative entropy relaxation term in order to limit policy oscillations between updates [36–38,63,64]. These references address the following entropy regularized SOC problem.

$$V_{g+1}(t,x) = \min_{\pi \in \mathcal{P}} \mathbb{E}_\pi \left[r_t(x,u) + \mathbb{E}_p \left[V_{g+1}(t+1,x') \right] \right] - \gamma \mathbb{H}[\pi] + \lambda \mathbb{D}[\pi \parallel \pi_g] \tag{24}$$

$$\pi_{g+1}(u|t,x) = \arg \min_{\pi \in \mathcal{P}} \mathbb{E}_\pi \left[r_t(x,u) + \mathbb{E}_p \left[V_{g+1}(t+1,x') \right] \right] - \gamma \mathbb{H}[\pi] + \lambda \mathbb{D}[\pi \parallel \pi_g] \tag{25}$$

Relying on the calculus of variations, one can verify that the solution is given by [38,65]

$$V_{g+1} = (\lambda + \gamma) \log \int \pi_g^{\frac{\lambda}{\lambda+\gamma}} e^{-\frac{1}{\lambda+\gamma}\left(r + \mathbb{E}_p\left[V'_{g+1}\right]\right)} \mathrm{d}u$$

$$\pi_{g+1} \propto \pi_g^{\frac{\lambda}{\lambda+\gamma}} e^{-\frac{1}{\lambda+\gamma}\left(r + \mathbb{E}_p\left[V'_{g+1}\right]\right)}$$

The authors in the respective references do use this problem formulation as a starting point to derive policy search methods. Because of the expectation in the exponent, it is impossible to practice the recursion and find an explicit expression for the corresponding optimal path distribution which would circumvent the dependency on the value function. Therefore all of them require to estimate the value, $V_g(t,x)$, or state-action value function, $Q_g(t,x,u)$, either using a general function approximator or a local approximation. As a consequence, the entropy regularized optimal control framework can also not be used to derive a PIC method, at least not in the sense discussed in Section 3.2.

In this article, we specifically limit our focus to the class of PIC methods described in Section 3.2, which as stated assume the existence of an explicit goal state trajectory distribution. In the following section, we introduce an adjusted entropy regularized stochastic optimal problem for which there does exists a formal yet explicit optimal path distribution, and that consequently can be treated with the PIC machinery put in place.

4.3. Entropy Regularized Trajectory Optimization

As discussed in the previous section, there exists no explicit expression for the optimal path distribution sequence in the setting of entropy regularized SOC. Consequently, the framework can not be leveraged by the PIC design principle described in Section 3.2. Here, we aim to identify an entropy regularized SOC problem that can be solved for an explicit trajectory distribution. As was identified to be an essential condition in the setting of LSOC, therefore we must get rid of the stochastic policy, π. Therefore, we suggest to absorb it into the implied state transition distribution, p_π. Analogously, this will elevate the problem from the policy to the state transition distribution optimization space, provided that we also can get rid of the dependency of the cost rate $r_t(x,u)$ on the control effort u.

Let us therefore assume there exists an inverse dynamic function f^{-1} so that $u = f^{-1}(t,x,x')$ when $x' = f(t,x,u)$. This implies a trajectory cost rate function c_t that accumulates into a trajectory cost $C = \sum_{t=0}^{T} c_t$. Specifically, we have that $c_t(x,x') = r_t(x, f^{-1}(x,x'))$ and $c_T(x) = r_T(x)$ (We emphasize here that the existence of such an inverse model is a mere formal assumption. In the practical setting to be presented in Section 5, we will never actually have to invert the system dynamics given that we have direct access to the control values that have been applied to the system).

These definitions allow to define a stochastic state trajectory optimization problem. We can rewrite the stochastic Bellman Equation (2) as follows,

$$V(t,x) = \min_\pi \int \int \left(r_t(x,u) + V(t+1,x') \right) p(x'|x,u) \pi(u|t,x) \mathrm{d}u \mathrm{d}x'$$

and then substitute the expressions above. This yields a similar problem

$$V(t,x) = \min_\pi \int \left(c_t(x,x') + V(t+1,x') \right) p_\pi(x'|t,x) \mathrm{d}x'$$

As there is a one-to-one correspondence between π and p_π, one can replace the minimization with respect to the policy by a minimization with respect to the implied state transition distribution p_π.

Finally, we can regularize and relax the problem according to the general entropic principles justified in Section 4.1 and as are thus also common in the RL community.

We end up with what we refer to as the Entropy Regularized Trajectory Optimization (ERTO) problem.

$$V_{g+1}(t,x) = \min_{p_\pi \in \mathcal{P}} \mathbb{E}_{p_\pi}[c_t(x,x') + V_{g+1}(t+1,x')] - \gamma \mathbb{H}[p_\pi] + \lambda \mathbb{D}\left[p_\pi \parallel p_{\pi_g}\right] \tag{26}$$

$$p_{\pi_{g+1}}(x'|t,x) = \arg\min_{p_\pi \in \mathcal{P}} \mathbb{E}_{p_\pi}[c_t(x,x') + V_{g+1}(t+1,x')] - \gamma \mathbb{H}[p_\pi] + \lambda \mathbb{D}\left[p_\pi \parallel p_{\pi_g}\right] \tag{27}$$

Remark that this problem is fundamentally different from that given in (24) considering that the penalties, $\mathbb{D}\left[\pi \parallel \pi_g\right]$ and $\mathbb{D}\left[p_\pi \parallel p_{\pi_g}\right]$, and, $\mathbb{H}[\pi]$ and $\mathbb{H}[p_\pi]$, do not express the same restrictions.

Analogously to the LSOC, we summarize the most profound properties of problem (26) in the following theorem. The theorem also establishes a relation between the optimal stochastic policy $\pi_g(t,x)$, the optimal state transition distribution $p_{\pi_g}(x'|t,x)$ and the optimal state trajectory distribution $p_{\pi_g}(\tau_x)$. Again, the proof relies on the calculus of variations mostly. We direct the reader to Appendix D for details.

Theorem 3. *With $V_{g+1}(t,x)$ defined as in (26), the following problems are equivalent,*

$$\pi_{g+1}(u|t,x) = \arg\min_{\pi \in \mathcal{P}} \mathbb{E}_{p_\pi}[c_t(x,x') + V_{g+1}(t+1,x')] - \gamma \mathbb{H}[p_\pi] + \lambda \mathbb{D}\left[p_\pi \parallel p_{\pi_g}\right] \tag{28}$$

$$p_{\pi_{g+1}}(x'|t,x) = \arg\min_{p_\pi \in \mathcal{P}} \mathbb{E}_{p_\pi}[c_t(x,x') + V_{g+1}(t+1,x')] - \gamma \mathbb{H}[p_\pi] + \lambda \mathbb{D}\left[p_\pi \parallel p_{\pi_g}\right] \tag{29}$$

$$p_{\pi_{g+1}}(\tau_x) = \arg\min_{p_\pi \in \mathcal{P}} \mathbb{E}_{p_\pi}[C] - \gamma \mathbb{H}[p_\pi] + \lambda \mathbb{D}\left[p_\pi \parallel p_{\pi_g}\right] \tag{30}$$

The latter can be solved explicitly,

$$p_{\pi_{g+1}}(x'|t,x) = p_{\pi_g}(x'|t,x)^{\frac{\lambda}{\lambda+\gamma}} e^{-\frac{1}{\lambda+\gamma}(c_t(x,x')+V(t+1,x'))}$$

$$p_{\pi_{g+1}}(\tau_x) = p_{\pi_g}(\tau_x)^{\frac{\lambda}{\lambda+\gamma}} e^{-\frac{1}{\lambda+\gamma}C(\tau_x)}$$

where $V_{g+1}(t,x)$ is governed by the recursion

$$V_{g+1}(t,x) = (\lambda + \gamma) \log \int p_{\pi_g}(x'|t,x)^{\frac{\lambda}{\lambda+\gamma}} e^{-\frac{1}{\lambda+\gamma}(c_t(x,x')+V(t+1,x'))} dx'$$

and where $\pi_g(u|t,x)$ and $p_{\pi_g}(x'|t,x)$ are related as

$$p_{\pi_g}(x'|t,x) = \int p(x'|x,u)\pi_g(u|t,x)du$$

We list the most important implications of Theorem 3:

- As a result of the entropy regularization and the state trajectory lifted optimization space, it is now possible to obtain another formal yet explicit optimal state trajectory distribution. Note that this is not the same optimal distribution as we derived in the LSOC setting given that here the control is not penalized through a Kullback–Leibler divergence term but is penalized implicitly through the cost C. When we evaluate C and have access to the state-action trajectory τ, we can simply replace C by R. Here, we emphasize that p_{π_g} still represents the state trajectory distribution which is now also a function of the actions.

$$p_{\pi_{g+1}} \propto p_{\pi_g}^{\frac{\lambda}{\lambda+\gamma}} \cdot e^{-\frac{1}{\lambda+\gamma}C} \equiv p_{\pi_g}^{\frac{\lambda}{\lambda+\gamma}} \cdot e^{-\frac{1}{\lambda+\gamma}R} \tag{31}$$

- Second, the theorem implies that we can readily apply the PIC design strategy described in Section 3.2, substituting the optimal path distribution sequence (31) for $p_{\pi}^{\star}$, and, the parametrized path distribution $p_{\pi(\theta)}$ induced by some stochastic parametric policy $\pi(\theta)$ for $p_{u(\theta)}$, to derive a generalized class of PIC methods. We refer to this class as Entropy Regularized Path Integral Control or ERPIC.

$$\min_{\theta} \mathbb{D}\left[p_{\pi_{g+1}} \parallel p_{\pi(\theta)}\right] \propto \mathbb{E}_{p_{\pi(\theta_g)}}\left[\frac{p_{\pi_{g+1}}}{p_{\pi(\theta_g)}} \log p_{\pi(\theta)}\right] + \text{constant} \tag{32}$$

$$\approx \frac{1}{M}\sum_{k=1}^{M} p_{\pi(\theta_g)}(\tau_x^k)^{-\frac{\gamma}{\lambda+\gamma}} \cdot e^{-\frac{1}{\lambda+\gamma}C(\tau_x^k)} \cdot \log p_{\pi(\theta)}(\tau_x^k) + \text{constant}$$

- As their is no longer a formal difference between problem (19) and (30), it follows that the properties derived for the sequence (21) also hold for the optimal state trajectory distribution sequence $\{p_{\pi_g}\}$ and imply monotonic convergence to $p_{\pi_\infty}(\tau_x) \propto \exp\left(-\frac{1}{\gamma}C(\tau_x)\right)$.

We conclude this section with a final remark related to control affine deterministic system dynamics, $x' = a(x) + B(x)u$, in this setting controlled by a stochastic policy of the form $\pi_g(u|t,x) = \mathcal{N}(u|u_g(t,x), \Sigma_g(t,x))$. Again, it is possible to derive an explicit expression for the optimal policy. The controlled transition distribution is given by $p(x'|t,x,u) = \mathcal{N}(x'|a(x) + B(x)u_g(t,x), B(x)\Sigma_g(t,x)B(x)^\top)$. By matching the moments of the two distributions $\mathcal{N}(x'|a(x) + B(x)u_{g+1}(t,x), B(x)\Sigma_{g+1}(t,x)B(x)^\top)$ and $p_{\pi_{g+1}}(x'|t,x)$ we find expressions for $u_{g+1}(t,x)$ and $\Sigma_{g+1}(t,x)$.

It is easily verified that

$$u_{g+1}(t,x) = u_g(t,x) + \frac{\mathbb{E}_{p_{\pi_g}(\tau_x|t,x)}\left[\xi \cdot e^{-\frac{1}{\lambda+\gamma}C(\tau_x|t,x)}\right]}{\mathbb{E}_{p_{\pi_g}(\tau_x|t,x)}\left[e^{-\frac{1}{\lambda+\gamma}C(\tau_x|t,x)}\right]} \tag{33}$$

and

$$\Sigma_{g+1}(t,x) = \frac{\mathbb{E}_{p_{\pi_g}(\tau_x|t,x)}\left[\xi\xi^\top \cdot e^{-\frac{1}{\lambda+\gamma}C(\tau_x|t,x)}\right]}{\mathbb{E}_{p_{\pi_g}(\tau_x|t,x)}\left[e^{-\frac{1}{\lambda+\gamma}C(\tau_x|t,x)}\right]}$$
$$-\frac{\mathbb{E}_{p_{\pi_g}(\tau_x|t,x)}\left[\xi \cdot e^{-\frac{1}{\lambda+\gamma}C(\tau_x|t,x)}\right]}{\mathbb{E}_{p_{\pi_g}(\tau_x|t,x)}\left[e^{-\frac{1}{\lambda+\gamma}C(\tau_x|t,x)}\right]}\frac{\mathbb{E}_{p_{\pi_g}(\tau_x|t,x)}\left[\xi^\top \cdot e^{-\frac{1}{\lambda+\gamma}C(\tau_x|t,x)}\right]}{\mathbb{E}_{p_{\pi_g}(\tau_x|t,x)}\left[e^{-\frac{1}{\lambda+\gamma}C(\tau_x|t,x)}\right]} \tag{34}$$

where $\xi \sim \mathcal{N}(0, \Sigma_g(t,x))$.

We point out the similarities with (7). Further note that for $\lambda = 0$, the information-theoretic trust-region is lifted in which case (33) can be applied in the same MPC sense as (7).

5. Formal Comparison of Path Integral Control Methods

In this final section, we give a formal comparison of three different PIC methods based on the general design principle identified in Section 3.2 that are tailored to trajectory optimization. The goal distribution $p_{u^\star}$ can now either root from the LSOC or the ETRO setting, respectively, introduced in Sections 3.1 and 4.3. The methods we discuss are tailored to find local solutions. That is policies that will control the system when the initial state is equal to a given state x_0. Sometimes on refers to such strategies as trajectory-based policy optimization methods. All of the methods are model-free and sample-based. Currently, such methods are used in the literature either to deal with complex simulation environments where traditional gradient based trajectory optimizers fall short [28,42] or to derive reinforcement learning algorithms [29,41,49].

The methods discussed here are in that sense closest related to [38]. However, this algorithm derives from the generic entropy regularized SOC problem (24) which differs significantly from the problem formulation proposed in Section 4.3. We come back to this later.

Remark that depending on the theoretical setting, either LSOC or ETRO, we have to use a different parametrized state trajectory distribution which has an effect on the PIC objectives. In case of LSOC based PIC we address the following objective,

$$\hat{J}_{g,\text{LSOC}}(\theta) = \sum_{k=1}^{K} \frac{w_g^k}{\sum_{k=1}^{K} w_g^k} \log p_{u(\theta)}(\tau_x^k) \tag{35a}$$

while in case of ERTO based PIC we address

$$\hat{J}_{g,\text{ERTO}}(\theta) = \sum_{k=1}^{K} \frac{w_g^k}{\sum_{k=1}^{K} w_g^k} \log p_{\pi(\theta)}(\tau_x^k) \tag{35b}$$

The values w_g^k differ for each methods and will be discussed in the following section.

5.1. Control Affine Systems

We concentrate on systems governed by deterministic control affine dynamics. In either case this is the sole condition for which an explicit solution exists for the optimal policies $u_{\text{LSOC}}(t, x)$ or $\pi_g(t, x)$, given in (7) or (33)), respectively . Furthermore, it is also the sole condition for which w_g^k can be evaluated exactly. This assumption introduces little practical limitations as many systems comply to this system model.

$$x' = a(x) + B(x)u$$

Further let us assume Gaussian policies of the form

$$\pi(u|t, x) = \mathcal{N}\left(u|u_t(x), \Sigma_t(x)\right)$$

In the case that we discuss methods derived from the LSOC setting, this means that we deliberately introduce a stochastic policy to mimic stochastic system dynamics. We emphasize that in this case $\Sigma_t(x)$ can not be chosen freely as it is directly related to the control effort penalization. From here on forth, we can thus consider stochastic policies in the LSOC setting as long as we silently acknowledge that the covariance can not be chosen freely.

Then the state transition distribution is given by

$$p_\pi(x'|t, x) = \mathcal{N}\left(x'|a(x) + B(w)u_t(x), B(x)\Sigma_t(x)B(x)^\top\right)$$

We can now substitute these expression into the different PIC objectives defined throughout the rest of the paper to find expressions for w_g^k.

Therefore, we will require the following intermediate results,

$$\log p_0(\tau_x^k) = \sum_{t=0}^{T-1} \log p_0\left(x_{t+1}^k \,\middle|\, t, x_t^k\right) \propto -\sum_{t=0}^{T-1} \frac{1}{2} \left\| u_{g,t} + \delta u_t^k \right\|_{R_t^k}^2$$

$$\log p_{\pi(\theta_g)}(\tau_x^k) = \sum_{t=0}^{T-1} \log p_{\pi(\theta_g)}\left(x_{t+1}^k \,\middle|\, t, x_t^k\right) \propto -\sum_{t=0}^{T-1} \frac{1}{2} \left\| \delta u_t^k \right\|_{R_t^k}^2$$

where $R_t = B^\top \left(B\Sigma_t B^\top\right)^{-1} B \approx \Sigma_t^{-1}$ and $\delta u_t^k \sim \mathcal{N}(0, \Sigma_t(x_t^k))$. These results follow from the fact that the samples satisfy the following condition

$$x_{t+1}^k = a_t^k + B_t^k u_{g,t}^k + B_t^k \delta u_t^k$$

Finally we can write the weights w_g^k as $e^{-P(\tau_x^k)}$ where the function $P(\cdot)$ depends on the specific objective. We will address two LSOC based objectives associated with SEPIC (or PICE) (9) and PIREPS (or ASPIC) (10) and the ETRO based objective associated to ERPIC (32). When we substitute the appropriate expressions in the corresponding objectives, we obtain, respectively,

- SEPIC (or PICE)

$$P_{\text{SEPIC}} = l_T(x_T) + \sum_{t=0}^{T-1} l_t(x_t) + \sum_{t=0}^{T-1} \tfrac{1}{2} \left\| u_{g,t} + \delta u_t \right\|_{R_t}^2 - \sum_{t=0}^{T-1} \tfrac{1}{2} \left\| \delta u_t \right\|_{R_t}^2$$

The function P reads as the cost accumulated over the trajectory τ^k where the states and control efforts are penalized separately. The trailing term is included to compensate for the full noise penalization [41]. To penalize the states any nonlinear function can be used, the control is penalized using a quadratic penalty term which depends on the noise added to the system. A crucial limitations is clearly that the exploration noise and the control cost are therefore coupled.

- PIREPS (or ASPIC)

$$P_{\text{PIREPS}} = \tfrac{1}{1+\lambda} P_{\text{SEPIC}}$$

It turns out that in this setting the weights w_g^k are simply a smoothed version of those associated to SEPIC. For $\lambda \gg 1$ (i.e., strong regularization) the weights will all have approximately the same value and therefore $u_{g+1}(t, x) \approx u_g(t, x)$. For $1 \gg \lambda > 0$ (i.e., weak regularization), the method reduces to SEPIC.

- ERPIC

$$P_{\text{ERPIC}} = \tfrac{1}{\lambda+\gamma} r_T(x_T) + \tfrac{1}{\lambda+\gamma} \sum_{t=0}^{T-1} r_t(x_t, u_t) - \tfrac{\gamma}{\lambda+\gamma} \sum_{t=0}^{T-1} \tfrac{1}{2} \left\| \delta u_t \right\|_{R_t}^2$$

One can easily verify that in the ERPIC setting, function P represents the cost accumulated over the trajectory τ^k where the states and control efforts are no longer penalized separately. Here, a discount terms is included that promotes uncertain trajectories making sure that the entropy of the search distribution does not evaporate eventually. Second, we wish to point out the obvious similarities with the stochastic search method in Appendix C.

5.2. Locally Linear Gaussian Policies

In conclusion, we solve the optimization problem defined in (32) exactly. Therefore, we will approximate the Gaussian policy $\mathcal{N}(u|u_t(x), \Sigma_t(x))$ with a locally linear Gaussian feedback policy of the form $\mathcal{N}(u|u_{g,t} + K_{g,t}x, \Sigma_{g,t})$.

In the ERPIC setting, the proportionality between the control penalization and the injected noise is lifted and therefore we gain access to the full parametrization of the policies, specifically $\theta_{g,t} = \{a_{g,t}, K_{g,t}, \Sigma_{g,t}\}$. Starting from (35b) it follows that

$$\begin{aligned}
\{a_{g,t}, K_{g,t}, \Sigma_{g,t}\} &= \max_{\theta_{g,t}} \sum_{k=1}^{K} \frac{w_g^k}{\sum_{k=1}^{K} w_g^k} \log p_{\pi(\theta)}(\tau_x^k) \\
&= \max_{\theta_{g,t}} \sum_{k=1}^{K} \frac{w_g^k}{\sum_{k=1}^{K} w_g^k} \sum_{t=0}^{T-1} \log \mathcal{N}(u|u_{g,t} + K_{g,t}x, \Sigma_{g,t}) \\
&= \max_{\theta_{g,t}} \sum_{k=1}^{K} \frac{w_g^k}{\sum_{k=1}^{K} w_g^k} \log \mathcal{N}(u|u_{g,t} + K_{g,t}x, \Sigma_{g,t})
\end{aligned} \qquad (36)$$

This procedure then yields the following elaborate updates (we refer to Appendix E for a proper derivation). Notation $\langle\langle f \rangle\rangle$ is shorthand for the likelihood weighted average $\sum_{k=1}^{K} \frac{w_g^k}{\sum_{k=1}^{K} w_g^k} f^k$ while $\langle f \rangle$ is shorthand for the empirical mean $\frac{1}{K} \sum_{k=1}^{K} f^k$.

$$u_{g+1,t} = u_{g,t} + \Delta\hat{\mu}_{u,g,t} - \hat{\Sigma}_{ux,g,t}\hat{\Sigma}_{xx,g,t}^{-1} \left(\hat{x}_{g,t} + \Delta\hat{\mu}_{x,g,t}\right) \tag{37}$$

$$\mathrm{K}_{g+1,t} = \hat{\Sigma}_{ux,g,t}\hat{\Sigma}_{xx,g,t}^{-1} \tag{38}$$

$$\Sigma_{g+1,t} = \hat{\Sigma}_{uu,g,t} - \hat{\Sigma}_{ux,g,t}\hat{\Sigma}_{xx,g,t}^{-1}\hat{\Sigma}_{xu,g,t} \tag{39}$$

with

$$\Delta\hat{\mu}_{u,g,t} = \langle\langle \Delta u_t^k \rangle\rangle$$

$$\Delta\hat{\mu}_{x,g,t} = \langle\langle \Delta x_t^k \rangle\rangle$$

$$\hat{\Sigma}_{uu,g,t} = \langle\langle \Delta u_t^k \Delta u_t^{k,\top} \rangle\rangle$$

$$\hat{\Sigma}_{ux,g,t} = \langle\langle \Delta u_t^k \Delta x_t^{k,\top} \rangle\rangle$$

$$\hat{\Sigma}_{xx,g,t} = \langle\langle \Delta x_t^k \Delta x_t^{k,\top} \rangle\rangle$$

where $\Delta x_t^k = x_t^k - \hat{x}_{g,t}$ with $\hat{x}_{g,t} = \langle x_t^k \rangle$ and $\Delta u_t^k = u_t^k - \hat{u}_{g,t}$ with $\hat{u}_{g,t} = \langle u_t^k \rangle = u_{g,t} + \mathrm{K}_{g,t}\hat{x}_{g,t}$ so that $\Delta u_t^k = \delta u_t^k + \mathrm{K}_{g,t}\Delta x_t^k$.

5.3. Discussion

As was made clear in the introduction of the paper, it is not our intention to provide a numerical analysis or study. We are merely concerned with the relation between all the topics that we touched upon. In conclusion, we will discuss therefore a number of observations that are of interest to fully grasp the relation between Path Integral Control methods, Stochastic Search methods and Reinforcement Learning and the implied limitations.

5.3.1. Remarks Related to Stochastic Search Methods and Variance

In this section, we comment on Stochastic Search Methods from two perspectives: The first perspective is related to the entropy regularized optimization framework and the theoretical search distribution sequences introduced in Section 4.1. The second perspective is related to the apparent connections between Entropy Regularized Optimization tailored to standard optimization and Entropy Regularized Trajectory Optimization which is tailored to optimal control problems.

As far as we are aware of, we provide an original argument to address the standard optimization problem as a probabilistic inference problem. The idea of treating the belief about the solution as a distribution function and aiming to reduce the expected cost value with each belief update is to our opinion very intuitive. It seem to be ideas worth pursuing whether existing stochastic search methods fit this abstract framework and whether other practical methods can be modeled after the theoretical distribution sequences. However, the mathematical framework can also be rewritten as a variational optimization problem with respect to a search distribution regularized by information-theoretic or information-geometric trust-regions. From this perspective our entropic inference argument sheds a new light on earlier work in the context of stochastic search methods [11,12,14,36,61,64].

By conducting an Entropy Regularization in the context of Stochastic Optimal Control that draws direct inspiration from the LSOC setting, we were able to formulate ETRO. This problem shares the crucial trait with the LSOC problem that made it a topic of interest in learning-based control. As a result the problems looses the so called optimal substructure property which is characteristic to Optimal Control and can therefore be treated as a standard entropy regularized optimization problem. This condition has two consequences: The theoretical analysis become easier to handle. On the other hand, the framework infers the temporal policy parameters as if they influence the entire trajectory instead of only from time instant t.

Following the latter observation, it is interesting to compare the updates in (37) with those in Appendix C. In retrospect, these are almost identical apart from one crucial difference. The same weights w_g^k are used to update the temporal parameters $\{u_{g,t}, K_{g,t}, \Sigma_{g,t}\}$ regardless of t. This relates directly to the comment above and has a dreadful practical implication. The effect of a random policy variations δu_t^k at time t is accounted for by the complete associated trajectory. The trajectory itself is however influenced by random policy variations at different time instants $t' \in \{0, \dots, t-1, t+1, \dots, T-1\}$. We remark that the approach is therefore expected to suffer from high variance, especially for $T \gg 1$, which will ultimately destabilize the convergence. This observation is a fundamental flaw of the framework described in Section 3.2 and therefore holds for any PIC methods including any existing methods. The high variance can partially be alleviated by acknowledging that the parameter update at time instant t is independent from the stochastic variables $\{x_{t'}^k, u_{t'}^k\}_{t'=0}^{t-1}$ and therefore the optimization problem in (36) can be revised using temporal weights

$$\max_{\theta_{g,t}} \sum_{k=1}^{K} \frac{w_{g,t}^k}{\sum_{k=1}^{K} w_{g,t}^k} \log \mathcal{N}(u|u_{g,t} + K_{g,t}x, \Sigma_{g,t})$$

where $w_{g,t}^k = w_{g,t}(\tau^k) = \exp(-P_{t,\text{ERPIC}}(\tau^k))$ with

$$P_{t,\text{ERPIC}}(\tau) = \tfrac{1}{\lambda+\gamma} r_T(x_T) + \tfrac{1}{\lambda+\gamma} \sum_{t'=t}^{T-1} r_t(x_{t'}, u_{t'}) - \tfrac{\gamma}{\lambda+\gamma} \sum_{t'=t}^{T} \tfrac{1}{2} \|\delta a_{t'}\|_{R_{t'}}^2$$

The update is then more in line with the explicit expression for the optimal stochastic policy that is given in (33).

5.3.2. Positioning of PIC Methods within the Field of RL

Formerly it was unclear how the framework of Linearly Solvable Optimal Control was related exactly to other RL methods. By identifying the framework of Entropy Regularized Trajectory Optimization and comparing it with the framework of Entropy Regularized Stochastic Optimal Control this ambiguity is lifted. Second, the model-free (local) trajectory optimization algorithms presented in this section are closest related to the method presented in [38]. However, this algorithm is based on problem formulation (24) and therefore requires to estimate the Q_g-function which is not required in this strictly PIC-based setting. Related to the final comment above, this circumvents the problem of temporal variance induced by random policy variations at different time instants and is therefore expected to perform superiorly in practice.

6. Conclusions

In this paper, we addressed PIC methods, a class of policy search methods which have been closely tied to the theoretical setting of LSOC. Our work was motivated primarily by an unsatisfactory understanding and generality of Path Integral Control methods in relation with the setting of LSOC. Nevertheless, referred class of policy search methods enjoyed interest in the RL community with applications ranging from robust trajectory optimization and guided policy search to robust model based predictive control.

The LSOC setting was considered to be unique in the sense that an explicit expression exists for the optimal state trajectory distribution. The former we identified to be fundamental to the class of PIC methods. We illustrate that the existence of such a solution is not a unique trait to the setting of LSOC and argue that a similar solution can be derived from within the setting of Entropy Regularized Stochastic Optimal Control, a framework of widespread use nowadays in the RL community. The setting of Entropy Regularized Stochastic Optimal Control allows to treat a more general class of optimal control problems than the setting of LSOC, rendering LSOC obsolete.

In either case, the result follows from lifting the stochastic optimal control problem from the policy to the state transition distribution space making the control implicit. As a result of this characteristic, the properties of the implied state trajectory distribution sequence can be analyzed. We show that the sequence converges monotonically to the solution of the underlying deterministic optimal control problem.

To treat this sequence more formally, we give an original and compelling argument for the use of information-geometric measures in the context of stochastic search algorithms based on the principle of entropic inference. The main idea is to maintain a belief function over the solution space that is least committed to any assumption about the solution, apart from the requirement that the expectation over the objective should decrease monotonically between updates. The resulting Entropy Regularized Optimization framework may serve as an overarching paradigm to analyze and derive stochastic search algorithms.

In conclusion, the value of this article is in that it identifies PIC methods for what they really are, a class of model-free and approximation free policy search algorithms that are theoretically founded on a specific subclass of Entropy Regularized Stochastic Optimal Control. Therewith, the underlying machinery associated to preceding derivations is untied from the peculiar setting of LSOC and the relation between PIC methods and state-of-the-art Reinforcement Learning is finally demystified. Nevertheless, our investigation suggests that PIC methods in fact are structurally closer related to Stochastic Search Methods tailored to classic optimization problem than they are to state-of-the-art Reinforcement Learning methods tailored to optimal control and decision-making problems and that the associated policy parameter updates will therefore be prone to higher variances.

Appendix A. Proof of Theorem 2

Proof. Proofs are given for (16)–(18), respectively.

- Consider any $x^\bullet \in \mathcal{X} : x^\bullet \neq x^*$ and define $f^\bullet = f(x^\bullet) > f^*$, then there exists a set $\mathcal{X}_{f^\bullet > f} \subset \mathcal{X}$ so that $\forall x \in \mathcal{X}_{f^\bullet > f} : f^\bullet > f(x)$ and a set $\mathcal{X}_{f^\bullet \leq f} = \mathcal{X}/\mathcal{X}_{f^\bullet > f}$ so that $\forall x \in \mathcal{X}_{f^\bullet \leq f} : f^\bullet \leq f(x)$. These definitions allow to derive an upper bound for the value of $\pi_g(x^\bullet)$, specifically

$$
\begin{aligned}
\pi_g(x^\bullet) = \pi_g^\bullet &= \frac{e^{-gf^\bullet}}{\int_{\mathcal{X}_{f^\bullet > f}} e^{-gf} \mathrm{d}x + \int_{\mathcal{X}_{f^\bullet \leq f}} e^{-gf} \mathrm{d}x} \\
&= \frac{1}{\int_{\mathcal{X}_{f^\bullet > f}} e^{g(f^\bullet - f)} \mathrm{d}x + \int_{\mathcal{X}_{f^\bullet \leq f}} e^{g(f^\bullet - f)} \mathrm{d}x} \\
&\leq \frac{1}{\int_{\mathcal{X}_{f^\bullet > f}} e^{g(f^\bullet - f)} \mathrm{d}x}
\end{aligned}
$$

  Now as $\exp(f^\bullet - f) > 1, \forall x \in \mathcal{X}_{f^\bullet > f}$ it follows that $\pi_g^\bullet$ tends to 0 for $g \to \infty$. On the other hand, if we choose $x^\bullet = x^*$, one can easily verify that the denominator tends to 0 and thus $\pi_g^\bullet$ tends to ∞ as $g \to \infty$. This limit behavior agrees with that of the Dirac delta and the statement follows.

- To proof that $\mathbb{E}_{\pi_g}[f]$ is a monotonically decreasing function of g, we simply have to verify whether the derivative to g is strictly negative. Therefore, let us first express the expectation explicitly, introducing the normalizer

$$\mathbb{E}_{\pi_g}[f] = \frac{\int f a_g \mathrm{d}x}{\int a_g \mathrm{d}x}$$

where $a_g = e^{-g\frac{1}{\lambda}f}\pi_0$ with $\frac{\mathrm{d}}{\mathrm{d}g}a_g = a_g' = -\frac{1}{\lambda}f a_g$

Taking the derivative to g yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}g}\mathbb{E}_{\pi_g}[f] &= \frac{\int f a_g' \mathrm{d}x}{\int a_g \mathrm{d}x} - \frac{\int f a_g \mathrm{d}x \int a_g' \mathrm{d}x}{\int a_g \mathrm{d}x \int a_g \mathrm{d}x} \\ &= -\frac{1}{\lambda}\mathbb{E}_{\pi_g}[f^2] + \frac{1}{\lambda}\mathbb{E}_{\pi_g}[f]^2 \\ &= -\frac{1}{\lambda}\mathrm{Var}_{\pi_g}[f]\end{aligned}$$

As the variance is a strictly positive operator, expect for π_∞, the statement follows.

- The entropy of the distribution π_g is equal to

$$\mathbb{H}\left[\pi_g\right] = \log\int a_g \mathrm{d}x - \frac{\int \log(a_g)a_g \mathrm{d}x}{\int a_g \mathrm{d}x}$$

We will also need the logarithm of a_g which is $\log a_g = \log\pi_0 - g\frac{1}{\lambda}f$.

Taking the derivative of $\mathbb{H}\left[\pi_g\right]$ to g yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}g}\mathbb{H}\left[\pi_g\right] &= \frac{\int a_g' \mathrm{d}x}{\int a_g \mathrm{d}x} - \frac{\int \frac{1}{a_g}a_g a_g' \mathrm{d}x}{\int a_g \mathrm{d}x} - \frac{\int \log(a_g)a_g' \mathrm{d}x}{\int a_g \mathrm{d}x} + \frac{\int \log(a_g)a_g \mathrm{d}x \int a_g' \mathrm{d}x}{\int a_g \mathrm{d}x \int a_g \mathrm{d}x} \\ &= -\frac{\int \log(a_g)a_g' \mathrm{d}x}{\int a_g \mathrm{d}x} + \frac{\int \log(a_g)a_g \mathrm{d}x \int a_g' \mathrm{d}x}{\int a_g \mathrm{d}x \int a_g \mathrm{d}x} \\ &= \frac{1}{\lambda}\frac{\int \log(a_g)f a_g \mathrm{d}x}{\int a_g \mathrm{d}x} - \frac{1}{\lambda}\frac{\int \log(a_g)a_g \mathrm{d}x \int f a_g \mathrm{d}x}{\int a_g \mathrm{d}x \int a_g \mathrm{d}x} \\ &= \frac{1}{\lambda}\frac{\int \log\pi_0 f a_g \mathrm{d}x}{\int a_g \mathrm{d}x} - \frac{g}{\lambda^2}\frac{\int f^2 a_g \mathrm{d}x}{\int a_g \mathrm{d}x} - \frac{1}{\lambda}\frac{\int \log\pi_0 a_g \mathrm{d}x \int f a_g \mathrm{d}x}{\int a_g \mathrm{d}x \int a_g \mathrm{d}x} + \frac{g}{\lambda^2}\frac{\int f a_g \mathrm{d}x \int f a_g \mathrm{d}x}{\int a_g \mathrm{d}x \int a_g \mathrm{d}x} \\ &= -\frac{g}{\lambda^2}\mathrm{Var}_{\pi_g}[f] + \frac{1}{\lambda}\mathrm{Cov}_{\pi_g}[\log\pi_0, f]\end{aligned}$$

In case we choose π_0 uniform on $\mathcal{X}$ the entropy decreases monotonically with g. Otherwise, the entropy might temporarily increase especially when π_0 and π_∞ are far apart.

Further note that the rate of convergence increases with g.

□

Appendix B. Derivation of Derivatives in (22) and (23)

Consider the implied function of g by taking the expectation of f

$$\mathbb{E}_{\pi_g}[f] = \frac{\int f a_g \mathrm{d}x}{\int a_g \mathrm{d}x}$$

Here, function a_g is defined as $a_g = \pi_0^\alpha \cdot \pi_\infty^{1-\alpha}$, where $\alpha = (\frac{\lambda}{\lambda+\gamma})^g$ and $\pi_\infty \propto \exp(-\frac{1}{\gamma}f)$. We will also needs the derivatives

$$\alpha' = -\alpha\log\left(\frac{\lambda+\gamma}{\lambda}\right)$$

$$a_g' = -\alpha'\log\frac{\pi_\infty}{\pi_0}a_g = \alpha\log\left(\frac{\lambda+\gamma}{\lambda}\right)\log\frac{\pi_\infty}{\pi_0}a_g$$

Taking the derivative of $\mathbb{E}_{\pi_g}[f]$ to g yields

$$\frac{\mathrm{d}}{\mathrm{d}g}\mathbb{E}_{\pi_g}[f] = -\gamma\frac{\mathrm{d}}{\mathrm{d}g}\mathbb{E}_{\pi_g}[\log\pi_\infty] = -\gamma\left(\frac{\int\log\pi_\infty a'_g\mathrm{d}x}{\int a_g\mathrm{d}x} + \frac{\int\log\pi_\infty a_g\mathrm{d}x\int a'_g\mathrm{d}x}{\int a_g\mathrm{d}x\int a_g\mathrm{d}x}\right)$$

$$= -\gamma\alpha\log\left(\frac{\lambda+\gamma}{\lambda}\right)\left(\frac{\int\log\pi_\infty\log\frac{\pi_\infty}{\pi_0}a_g\mathrm{d}x}{\int a_g\mathrm{d}x} + \frac{\int\log\pi_\infty a_g\mathrm{d}x\int\log\frac{\pi_\infty}{\pi_0}a_g\mathrm{d}x}{\int a_g\mathrm{d}x\int a_g\mathrm{d}x}\right)$$

$$= -\gamma\alpha\log\left(\frac{\lambda+\gamma}{\lambda}\right)\mathrm{Cov}_{\pi_g}\left[\log\pi_\infty, \log\frac{\pi_\infty}{\pi_0}\right]$$

when π_0 is chosen uniform on $\mathcal{X}$ this is always smaller than 0 due to the positive definiteness of the covariance operator.

In a similar fashion we can address the entropy

$$\mathbb{H}\left[\pi_g\right] = \log\int a_g\mathrm{d}x - \frac{\int\log(a_g)a_g\mathrm{d}x}{\int a_g\mathrm{d}x}$$

Here we will also need the logarithm of a_g which is given by

$$\log a_g = \log\pi_\infty - \alpha\left(\log\pi_\infty - \log\pi_0\right)$$

Taking the derivative of $\mathbb{H}\left[\pi_g\right]$ to g yields

$$\frac{\mathrm{d}}{\mathrm{d}g}\mathbb{H}\left[\pi_g\right] = \frac{\int a'_g\mathrm{d}x}{\int a_g\mathrm{d}x} - \frac{\int\frac{1}{a_g}a_g a'_g\mathrm{d}x}{\int a_g\mathrm{d}x} - \frac{\int\log(a_g)a'_g\mathrm{d}x}{\int a_g\mathrm{d}x} + \frac{\int\log(a_g)a_g\mathrm{d}x\int a'_g\mathrm{d}x}{\int a_g\mathrm{d}x\int a_g\mathrm{d}x}$$

$$= -\frac{\int\log(a_g)a'_g\mathrm{d}x}{\int a_g\mathrm{d}x} + \frac{\int\log(a_g)a_g\mathrm{d}x\int a'_g\mathrm{d}x}{\int a_g\mathrm{d}x\int a_g\mathrm{d}x}$$

$$= -\alpha\log\left(\frac{\lambda+\gamma}{\lambda}\right)\left(\frac{\int\log(a_g)\log\frac{\pi_\infty}{\pi_0}a_g\mathrm{d}x}{\int a_g\mathrm{d}x} - \frac{\int\log(a_g)a_g\mathrm{d}x\int\log\frac{\pi_\infty}{\pi_0}a_g\mathrm{d}x}{\int a_g\mathrm{d}x\int a_g\mathrm{d}x}\right)$$

$$= -\alpha\log\left(\frac{\lambda+\gamma}{\lambda}\right)\mathrm{Cov}_{\pi_g}\left[\log a_g, \log\pi_\infty - \log\pi_0\right]$$

$$= -\alpha\log\left(\frac{\lambda+\gamma}{\lambda}\right)\left((1-\alpha)\mathrm{Cov}_{\pi_g}\left[\log\pi_\infty, \log\frac{\pi_\infty}{\pi_0}\right] + \alpha\mathrm{Cov}_{\pi_g}\left[\log\pi_0, \log\frac{\pi_\infty}{\pi_0}\right]\right)$$

Again, when π_0 is uniform on $\mathcal{X}$, the sequence decreases monotonically.

Appendix C. Entropy Regularized Evolutionary Strategy

In order to illustrate the practical use of the search distribution sequence in (20)–(21), let us demonstrate how to cast the theoretical distribution into a practical search algorithm. To that end, we project π_g onto a parametric distribution family $\pi(\theta)$ and manipulate the resulting expression into an expectation over the prior belief π_g.

Note that this strategy is analogous to the one described in Section 3.2. As such we establish a calculable update procedure that infers parameters from an estimated expectation using samples taken from the prior. In particular we are interested in the Gaussian family, i.e., $\pi(x|\theta) = \mathcal{N}(x|\mu, \Sigma)$, which is commonly used in the context of evolutionary strategies [11,13,14]. As a projection operator again the relative entropy is used relying on the same argument as in [41].

$$\theta_{g+1} = \arg\min_\theta\mathbb{D}\left[\pi_{g+1}\parallel\pi(\theta)\right] = -\underbrace{\mathbb{E}_{\pi(\theta_g)}\left[\frac{\pi_{g+1}}{\pi(\theta_g)}\log\pi(\theta)\right]}_{J_g(\theta)} + \text{constant}$$

This problem can be solved explicitly. We propose solving it for each parameter independently using a coordinate descent strategy, substituting the previous value of the respective other in the objective. This strategy renders each independent problem concave [11].

The derivative of J_g to μ and Σ equal

$$\nabla_\mu J_g|_{\mu,\Sigma_g} = -\Sigma_g^{-1} \mathbb{E}_{\pi_g} \left[w_g(x) \left(\delta x_g - \mu + \mu_g \right) \right]$$

$$\nabla_\Sigma J_g|_{\mu_g,\Sigma} = \mathbb{E}_{\pi_g} \left[w_g(x) \Sigma^{-1} \delta x_g \delta x_g^\top \Sigma^{-1} \right] - \Sigma^{-1}$$

where $w_g(x) = \pi(x|\theta_g)^{\frac{-\gamma}{\lambda+\gamma}} e^{-\frac{1}{\lambda+\gamma} f(x)}$ and $\delta x_g = x - \mu_g$.

We can solve these equations to provide independent update procedures for the distribution parameters.

$$\mu_{g+1} = \mu_g + \mathbb{E}_{\pi(\theta_g)} \left[w_g \delta x_g \right]$$

$$\Sigma_{g+1} = \mathbb{E}_{\pi(\theta_g)} \left[w_g \delta x_g \delta x_g^\top \right]$$

Finally we substitute Monte Carlo estimates estimates for the expectations. We obtain the following update which readers, familiar with the class of Evolutionary Strategies, will recognize to be similar to those they are accustomed with (Note that if it was not for the diffusion factor $\frac{\gamma}{\lambda+\gamma}$, the weights would simply equal the exponential transformed objective. We emphasize here that f does not need to represent the physical objective but could be any rank preserving mapping implying these updates correspond with those of any other ES provided that the mapping is known)

$$\mu_{g+1} = \mu_g + \sum_k \frac{w_g^k}{\sum_k w_g^k} \delta x_g^k$$

$$\Sigma_{g+1} = \sum_k \frac{w_g^k}{\sum_k w_g^k} \delta x_g^k \delta x_g^{k,\top}$$

where $w_g^k = w_g(x^k)$ and $\delta x_g^k = x^k - \mu_g$.

Finally, we note that the weights can also be expressed as illustrated below.

$$w_g^k \propto \exp \left(-\frac{1}{\lambda+\gamma} \left(f(x^k) - \frac{\gamma}{2} \left\| \delta x_g^k \right\|_{\Sigma_g^{-1}}^2 \right) \right)$$

It is noted that this strategy is closely related to the work in [61].

Appendix D. Proof of Theorem 3

Proof. The equivalence of (28) and (29) follows directly from the definition of p_π.

The equation in (29) determines a variational problem in p_π which we can solve correspondingly

$$p_{\pi_g}(x'|t,x) \propto p_{\pi_g}^{\frac{\lambda}{\lambda+\gamma}} e^{-\frac{1}{\lambda+\gamma} c_t(x,x') - \frac{1}{\lambda+\gamma} V_{g+1}(t+1,x')}$$

and that we can substitute back into (26) to yield the recursive equation

$$\frac{1}{\lambda+\gamma} V_{g+1}(t,x) = \log \int p_{\pi_g}(x'|t,x)^{\frac{\lambda}{\lambda+\gamma}} \exp \left(-\frac{1}{\lambda+\gamma} c_t(x,x') - \frac{1}{\lambda+\gamma} V_{g+1}(t+1,x') \right) dx'$$

Either by iterating the recursion or by solving problem (30) directly, it is then also easily verified that

$$p_{\pi_{g+1}}(\tau_x) \propto p_{\pi_g}(\tau_x)^{\frac{\lambda}{\lambda+\gamma}} e^{-\frac{1}{\lambda+\gamma} C(\tau_x)}$$

proving the second statement. $\square$

Appendix E. Derivation of Equation (37)

Here, we address the optimization problems defined in (32). We address the specific context where the system dynamics are control affine and where we use a locally linear Gaussian feedback policy.

In general, we then obtain an expression of the following form and where w_g^k, x, μ, and Λ are problem-specific.

$$\min_{u,K,\Sigma} \hat{J}_g(a, K, \Sigma) = \sum_k \frac{w_g^k}{\sum_k w_g^k} \log \mathcal{N}\left(x^k | \mu^k, \Lambda^k\right) = \left\langle\!\!\left\langle \log \mathcal{N}\left(x^k | \mu^k, \Lambda^k\right) \right\rangle\!\!\right\rangle$$

Recall we introduced the notation $\langle\!\langle \cdot \rangle\!\rangle$ to denote the likelihood weighted average in addition to notation $\langle \cdot \rangle$ for the empirical mean.

The logarithm can be expressed as

$$\log \mathcal{N}(x|\mu, \Lambda) \propto -\log|\Lambda| - \mathrm{tr}\left(\Lambda^{-1}(x-\mu)(x-\mu)^\top\right) + c$$

where c is some constant.

For problem (32), we have that

$$x^k = x_{t+1}^k = a_t^k + B_t^k u_{g,t} + B_t^k K_{g,t} x_t^k + B_t^k \delta u_t^k$$
$$\mu^k = a_t^k + B_t^k u + B_t^k K x_t^k$$
$$\Lambda^k = B_t^k \Sigma B_t^{k,\top}$$

so that

$$x - \mu = B_t^k \left(u_{g,t} - u + (K_{g,t} - K)x_t^k + \delta u_t^k\right)$$

Regardless of the procedure that is used, we can express the first order optimality conditions, where the proportionality includes matrix multiplications with positive definite matrices.

$$\nabla_u \hat{J}_g \propto \left\langle\!\!\left\langle u_{g,t} - u + (K_{g,t} - K)x_t^k + \delta u_t^k \right\rangle\!\!\right\rangle = 0$$
$$\nabla_K \hat{J}_g \propto \left\langle\!\!\left\langle \left(u_{g,t} - u + (K_{g,t} - K)x_t^k + \delta u_t^k\right) x_t^{k,\top} \right\rangle\!\!\right\rangle = 0$$
$$\nabla_\Sigma \hat{J}_g \propto \Sigma^{-1} - \left\langle\!\!\left\langle \Sigma^{-1}(x^k - \mu^k)(x^k - \mu^k)^\top \Sigma^{-1} \right\rangle\!\!\right\rangle = 0$$

These equations can be solved to yield expressions for $u_{g+1,t}$, $K_{g+1,t}$ and $\Sigma_{g+1,t}$

$$u_{g+1,t} = u_{g,t} + \Delta \hat{\mu}_{u,g,t} - \hat{\Sigma}_{ux,g,t} \hat{\Sigma}_{xx,g,t}^{-1} \left(\hat{x}_{g,t} + \Delta \hat{\mu}_{x,g,t}\right)$$
$$K_{g+1,t} = \hat{\Sigma}_{ux,g,t} \hat{\Sigma}_{xx,g,t}^{-1}$$
$$\Sigma_{g+1,t} = \hat{\Sigma}_{uu,g,t} - \hat{\Sigma}_{ux,g,t} \hat{\Sigma}_{xx,g,t}^{-1} \hat{\Sigma}_{xu,g,t}$$
$$\Delta \hat{\mu}_{u,g,t} = \left\langle\!\!\left\langle \Delta u_t^k \right\rangle\!\!\right\rangle$$
$$\Delta \hat{\mu}_{x,g,t} = \left\langle\!\!\left\langle \Delta x_t^k \right\rangle\!\!\right\rangle$$
$$\hat{\Sigma}_{uu,g,t} = \left\langle\!\!\left\langle \Delta u_t^k \Delta u_t^{k,\top} \right\rangle\!\!\right\rangle$$
$$\hat{\Sigma}_{ux,g,t} = \left\langle\!\!\left\langle \Delta u_t^k \Delta x_t^{k,\top} \right\rangle\!\!\right\rangle$$
$$\hat{\Sigma}_{xx,g,t} = \left\langle\!\!\left\langle \Delta x_t^k \Delta x_t^{k,\top} \right\rangle\!\!\right\rangle$$

where $\Delta x_t^k = x_t^k - \hat{x}_{g,t}$ with $\hat{x}_{g,t} = \langle x_t^k \rangle$ and $\Delta u_t^k = u_t^k - \hat{u}_{g,t}$ with $\hat{u}_{g,t} = \langle u_t^k \rangle = u_{g,t} + K_{g,t} \hat{x}_{g,t}$ so that $\Delta u_t^k = \delta u_t^k + K_{g,t} \Delta x_t^k$.

This concludes the derivation.

References

1. Heess, N.; Dhruva. T.B.; Sriram, S.; Lemmon, J.; Merel, J.; Wayne, G.; Tassa, Y.; Erez, T.; Wang, Z.; Eslami, S.; et al. Emergence of locomotion behaviours in rich environments. *arXiv* **2017**, arXiv:1707.02286.

2. Todorov, E. Optimal control theory. In *Bayesian Brain: Probabilistic Approaches to Neural Coding*; MIT Press: Cambridge, UK, 2006; pp. 269–298.

3. Mayne, D. A Second-order Gradient Method for Determining Optimal Trajectories of Non-linear Discrete-time Systems. *Int. J. Control* **1966**, *3*, 85–95. [CrossRef]

4. Tassa, Y.; Mansard, N.; Todorov, E. Control-limited differential dynamic programming. In Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 31 May–7 June 2014; pp. 1168–1175. [CrossRef]

5. Erez, T.; Todorov, E. Trajectory optimization for domains with contacts using inverse dynamics. In Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, Vilamoura, Portugal, 7–12 October 2012; pp. 4914–4919.

6. Tassa, Y.; Erez, T.; Todorov, E. Synthesis and stabilization of complex behaviors through online trajectory optimization. In Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, Vilamoura, Portugal, 7–12 October 2012; pp. 4906–4913.

7. Todorov, E.; Li, W. A generalized iterative LQG method for locally-optimal feedback control of constrained nonlinear stochastic systems. In Proceedings of the 2005 American Control Conference, Portland, OR, USA, 8–10 June 2005; Volume 1, pp. 300–306

8. Diehl, M.; Bock, H.; Diedam, H.; Wieber, P. Fast direct multiple shooting algorithms for optimal robot control. In *Fast Motions in Biomechanics and Robotics*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 65–93.

9. Kennedy, J.; Eberhart, R. Particle swarm optimization. In Proceedings of the ICNN'95-International Conference on Neural Networks, Perth, WA, Australia, 27 November–1 December 1995; Volume 4, pp. 1942–1948. [CrossRef]

10. Schwefel, H. *Numerische Optimierung von Computer-Modellen Mittels der Evolutionsstrategie: Mit Einer Vergleichenden Einführung in die Hill-Climbing-und Zufallsstrategie*; Springer: Berlin/Heidelberg, Germany, 1977; Volume 1.

11. Abdolmaleki, A.; Price, B.; Lau, N.; Reis, L.; Neumann, G. Deriving and improving CMA-ES with information geometric trust regions. In Proceedings of the Genetic and Evolutionary Computation Conference, ACM, Berlin, Germany, 15–19 July 2017; pp. 657–664.

12. Ollivier, Y.; Arnold, L.; Auger, A.; Hansen, N. Information-geometric optimization algorithms: A unifying picture via invariance principles. *J. Mach. Learn. Res.* **2017**, *18*, 564–628.

13. Hansen, N. The CMA evolution strategy: A tutorial. *arXiv* **2016**, arXiv:1604.00772.

14. Wierstra, D.; Schaul, T.; Glasmachers, T.; Sun, Y.; Peters, J.; Schmidhuber, J. Natural evolution strategies. *J. Mach. Learn. Res.* **2014**, *15*, 949–980.

15. Winter, S.; Brendel, B.; Pechlivanis, I.; Schmieder, K.; Igel, C. Registration of CT and intraoperative 3-D ultrasound images of the spine using evolutionary and gradient-based methods. *IEEE Trans. Evol. Comput.* **2008**, *12*, 284–296. [CrossRef]

16. Hansen, N.; Niederberger, A.; Guzzella, L.; Koumoutsakos, P. A method for handling uncertainty in evolutionary optimization with an application to feedback control of combustion. *IEEE Trans. Evol. Comput.* **2008**, *13*, 180–197. [CrossRef]

17. Villasana, M.; Ochoa, G. Heuristic design of cancer chemotherapies. *IEEE Trans. Evol. Comput.* **2004**, *8*, 513–521. [CrossRef]

18. Gholamipoor, M.; Ghadimi, P.; Alavidoost, M.; Feizi Chekab, M. Application of evolution strategy algorithm for optimization of a single-layer sound absorber. *Cogent Eng.* **2014**, *1*, 945820. [CrossRef]

19. Hansen, N.; Kern, S. Evaluating the CMA evolution strategy on multimodal test functions. In *Lecture Notes in Computer Science, Proceedings of the International Conference on Parallel Problem Solving from Nature, Birmingham, UK, 18–22 September 2004*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 282–291.

20. Gregory, G.; Bayraktar, Z.; Werner, D. Fast Optimization of Electromagnetic Design Problems Using the Covariance Matrix Adaptation Evolutionary Strategy. *IEEE Trans. Antennas Propag.* **2011**, *59*, 1275–1285. [CrossRef]

21. Kothari, D. Power system optimization. In Proceedings of the 2012 2nd National Conference on Computational Intelligence and Signal Processing (CISP), Guwahati, India, 2–3 March 2012; pp. 18–21. [CrossRef]

22. Todorov, E. Linearly-solvable Markov decision problems. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, UK, 2007; pp. 1369–1376.

23. Yong, J.; Zhou, X. *Stochastic Controls: Hamiltonian Systems and HJB Equations*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 1999; Volume 43.

24. Kappen, H. Linear theory for control of nonlinear stochastic systems. *Phys. Rev. Lett.* **2005**, *95*, 200201. [CrossRef] [PubMed]

25. Stulp, F.; Sigaud, O. Path integral policy improvement with covariance matrix adaptation. *arXiv* **2012**, arXiv:1206.4621.

26. Stulp, F.; Sigaud, O. Policy Improvement Methods: Between Black-Box Optimization and Episodic Reinforcement Learning. 2012. Available online: https://hal.archives-ouvertes.fr/hal-00738463 (accessed on 1 October 2020)

27. Lefebvre, T.; Crevecoeur, G. Path Integral Policy Improvement with Differential Dynamic Programming. In Proceedings of the 2019 IEEE International Conference on Advanced Intelligent Mechatronics (AIM), Hong Kong, China, 8–12 July 2019.

28. Rajamäki, J.; Naderi, K.; Kyrki, V.; Hämäläinen, P. Sampled differential dynamic programming. In Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Korea, 9–14 October 2016; pp. 1402–1409.

29. Chebotar, Y.; Kalakrishnan, M.; Yahya, A.; Li, A.; Schaal, S.; Levine, S. Path integral guided policy search. In Proceedings of the 2017 IEEE international conference on robotics and automation (ICRA), Singapore, 29 May–3 June 2017; pp. 3381–3388.

30. Summers, C.; Lowrey, K.; Rajeswaran, A.; Srinivasa, S.; Todorov, E. Lyceum: An efficient and scalable ecosystem for robot learning. *arXiv* **2020**, arXiv:2001.07343.

31. Williams, G.; Drews, P.; Goldfain, B.; Rehg, J.; Theodorou, E. Aggressive driving with model predictive path integral control. In Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 16–21 May 2016; pp. 1433–1440. [CrossRef]

32. Williams, G.; Drews, P.; Goldfain, B.; Rehg, J.; Theodorou, E. Information-Theoretic Model Predictive Control: Theory and Applications to Autonomous Driving. *IEEE Trans. Robot.* **2018**, *34*, 1603–1622. [CrossRef]

33. Theodorou, E.; Krishnamurthy, D.; Todorov, E. From information theoretic dualities to path integral and kullback-leibler control: Continuous and discrete time formulations. In Proceedings of the Sixteenth Yale Workshop on Adaptive and Learning Systems, Yale, CT, USA, 5–7 June 2013.

34. Theodorou, E. Nonlinear stochastic control and information theoretic dualities: Connections, interdependencies and thermodynamic interpretations. *Entropy* **2015**, *17*, 3352–3375. [CrossRef]

35. Haarnoja, T.; Tang, H.; Abbeel, P.; Levine, S. Reinforcement learning with deep energy-based policies. In Proceedings of the 34th International Conference on Machine Learning, Sydney, 7–9 August 2017; pp. 1352–1361.

36. Peters, J.; Mulling, K.; Altun, Y. Relative entropy policy search. In Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, Atlanta, GA, USA, 11–15 July 2010. .

37. Neu, G.; Jonsson, A.; Gómez, V. A unified view of entropy-regularized markov decision processes. *arXiv* **2017**, arXiv:1705.07798.

38. Akrour, R.; Abdolmaleki, A.; Abdulsamad, H.; Peters, J.; Neumann, G. Model-free trajectory-based policy optimization with monotonic improvement. *J. Mach. Learn. Res.* **2018**, *19*, 565–589.

39. Caticha, A. Entropic inference: Some pitfalls and paradoxes we can avoid. In Proceedings of the MaxEnt 2012, The 32nd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Garching, Germany, 15–20 July 2012; Volume 1553, pp. 200–211.

40. Szepesvári, C. *Reinforcement Learning Algorithms for MDPs*; Wiley Encyclopedia of Operations Research and Management Science; John Wiley & Sons: Hoboken, NJ, USA, 2010.

41. Kappen, H.; Ruiz, H.C. Adaptive importance sampling for control and inference. *J. Stat. Phys.* **2016**, *162*, 1244–1266. [CrossRef]

42. Williams, G.; Aldrich, A.; Theodorou, E. Model predictive path integral control: From theory to parallel computation. *J. Guid. Control. Dyn.* **2017**, *40*, 344–357. [CrossRef]

43. Kappen, H.; Wiegerinck, W.; van den Broek, B. A path integral approach to agent planning. In *Autonomous Agents and Multi-Agent Systems*; Springer Science & Business Media: Berlin/Heidelberg, Germany; 2007.

44. Drews, P.; Williams, G.; Goldfain, B.; Theodorou, E.; Rehg, J. Vision-Based High-Speed Driving With a Deep Dynamic Observer. *IEEE Robot. Autom. Lett.* **2019**, *4*, 1564–1571. [CrossRef]

45. Theodorou, E.; Buchli, J.; Schaal, S. A generalized path integral control approach to reinforcement learning. *J. Mach. Learn. Res.* **2010**, *11*, 3137–3181.

46. Theodorou, E.; Buchli, J.; Schaal, S. Reinforcement learning of motor skills in high dimensions: A path integral approach. In Proceedings of the 2010 IEEE International Conference on Robotics and Automation, Anchorage, AK, USA, 3–7 May 2010; pp. 2397–2403.

47. Gómez, V.; Kappen, H.; Peters, J.; Neumann, G. Policy search for path integral control. In *Lecture Notes in Computer Science, Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Nancy, France, 15–19 September 2014*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 482–497.

48. Pan, Y.; Theodorou, E.; Kontitsis, M. Sample efficient path integral control under uncertainty. In Proceedings of the Advances in Neural Information Processing Systems, Montréal, QC, Candada, 7–12 December 2015; pp. 2314–2322.

49. Thalmeier, D.; Kappen, H.; Totaro, S.; Gómez, V. Adaptive Smoothing Path Integral Control. *arXiv* **2020**, arXiv:2005.06364.

50. Yamamoto, K.; Ariizumi, R.; Hayakawa, T.; Matsuno, F. Path Integral Policy Improvement With Population Adaptation. *IEEE Trans. Cybern.* **2020**. [CrossRef] [PubMed]

51. Sun, Y.; Wierstra, D.; Schaul, T.; Schmidhuber, J. Efficient natural evolution strategies. In Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation, ACM, Montreal, QC, Canada, 8–12 July 2009; pp. 539–546.

52. Pourchot, A.; Perrin, N.; Sigaud, O. Importance mixing: Improving sample reuse in evolutionary policy search methods. *arXiv* **2018**, arXiv:1808.05832.

53. Giffin, A.; Caticha, A. Updating probabilities with data and moments. In Proceedings of the 27th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Saratoga Springs, NY, USA, 8–13 July 2007; Volume 954, pp. 74–84.

54. Jaynes, E. Information theory and statistical mechanics. *Phys. Rev.* **1957**, *106*, 620. [CrossRef]

55. Jaynes, E. Information theory and statistical mechanics. II. *Phys. Rev.* **1957**, *108*, 171. [CrossRef]

56. Kullback, S. *Information Theory and Statistics*; John Wiley & Sons: Hoboken, NJ, USA, 1959.

57. Johnson, R.; Shore, J. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Inf. Theory* **1980**, *26*, 26–37.

58. Jaynes, E. On the rationale of maximum-entropy methods. *Proc. IEEE* **1982**, *70*, 939–952. [CrossRef]

59. Tikochinsky, Y.; Tishby, N.; Levine, R. Consistent inference of probabilities for reproducible experiments. *Phys. Rev. Lett.* **1984**, *52*, 1357. [CrossRef]

60. Caticha, A. Entropic inference. In Proceedings of the MaxEnt 2010, the 30th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Chamonix, France, 4–9 July 2010; Volume 1305, pp. 20–29.

61. Abdolmaleki, A.; Lioutikov, R.; Peters, J.; Lau, N.; Reis, L.; Neumann, G. Model-based relative entropy stochastic search. In Proceedings of the Advances in Neural Information Processing Systems, Montréal, QC, Canada, 7–12 December 2015; pp. 3537–3545.

62. Haarnoja, T.; Zhou, A.; Hartikainen, K.; Tucker, G.; Ha, S.; Tan, J.; Kumar, V.; Zhu, H.; Gupta, A.; Abbeel, P.; et al. Soft actor-critic algorithms and applications. *arXiv* **2018**, arXiv:1812.05905.

63. Rubin, J.; Shamir, O.; Tishby, N. Trading value and information in MDPs. In *Decision Making with Imperfect Decision Makers*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 57–74.

64. Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; Moritz, P. Trust region policy optimization. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 1889–1897.

65. Rawlik, K.; Toussaint, M.; Vijayakumar, S. On stochastic optimal control and reinforcement learning by approximate inference. In Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, Beijing, China, 3–9 August 2013; AAAI Press: Menlo Park, CA, USA, 2013.