Unifying Entropy Regularization in Optimal Control:
From and Back to Classical Objectives via Iterated Soft Policies and Path Integral Solutions [*]

Ajinkya Bhole [*,**] **Mohammad Mahmoudi Filabadi** [*,**]
Guillaume Crevecoeur [*,**] **Tom Lefebvre** [*,**]

** Department of Electromechanical, Systems and Metal Engineering,
Ghent University, Ghent, Belgium (e-mail: ajinkya.bhole@ugent.be).
** Core lab MIRO, Flanders Make, Belgium.*

Abstract: This paper develops a unified perspective on several stochastic optimal control formulations through the lens of Kullback-Leibler regularization. We propose a central problem that separates the KL penalties on policies and transitions, assigning them independent weights, thereby generalizing the standard trajectory-level KL-regularization commonly used in probabilistic and KL-regularized control. This generalized formulation acts as a generative structure allowing to recover various control problems. These include the classical Stochastic Optimal Control (SOC), Risk-Sensitive Optimal Control (RSOC), and their policy-based KL-regularized counterparts. The latter we refer to as soft-policy SOC and RSOC, facilitating alternative problems with tractable solutions. Beyond serving as regularized variants, we show that these soft-policy formulations majorize the original SOC and RSOC problem. This means that the regularized solution can be iterated to retrieve the original solution. Furthermore, we identify a structurally synchronized case of the risk-seeking soft-policy RSOC formulation, wherein the policy and transition KL-regularization weights coincide. Remarkably, this specific setting gives rise to several powerful properties such as a linear Bellman equation, path integral solution, and, compositionality, thereby extending these computationally favourable properties to a broad class of control problems.

Keywords: KL-Regularized Control, Risk-Sensitive Stochastic Control, Path Integral Solutions.

1. INTRODUCTION

Optimal control problems arise in a wide range of application areas, from robotics and autonomous systems (Toussaint (2009); Williams et al. (2018)) to finance and operations research (Föllmer and Schied (2002)), where consecutive optimal decisions must be made. Optimal control problems share a common structure in which an agent seeks to shape the (stochastic) behaviour of a dynamical system by minimizing some notion of cost over a (finite) time horizon. These problems are typically addressed by means of dynamic programming, which decomposes the problem into several subproblems that are solved recursively. These subproblems, however, remain higly nonlinear and exact closed-form solutions are limited to but a few special cases (Kárnỳ (1996)). Motivated by this observation, over the last decades, many researchers investigated alternative problem formulations whose solutions closely resembles those of classical optimal control problems, however, exhibit more favourable computational properties.

Kárnỳ (1996) was amongst the first to acknowledge this problem and introduced the notion of probabilistic optimal control (Kárnỳ and Guy (2006)). Optimal decision making was formulated as a *density matching problem* using the Kullback-Leibler (KL) divergence as measure of discrepancy between densities. Rather than formulating the problem directly in terms of cost minimization, in this approach, the agent aims to shape the closed-loop density trajectories to match a desired target distribution as closely as possible. This perspective yields a solution that is structurally similar to dynamic programming; however, optimization operators are replaced by expectations over known densities, *improving the solution's tractability.*

This line of work is as an early predecessor of what is now known as Control as Inference (CaI) (Toussaint and Storkey (2006); Toussaint (2009); Levine (2018); Rawlik et al. (2012); Rawlik (2013); Neumann (2011)). In this formulation, the notion of cost is encoded in the probabilistic graphical model (PGM), corresponding to the dynamical system, through the introduction of auxiliary optimality variables. The probability of observing such an optimality variable is chosen to be proportional to the exponential of the negative control cost, thereby encoding low-cost trajectories as more probable. This formulation led to various productive ways of deriving policies (Levine and Koltun (2013a,b); Rawlik et al. (2012)), yet for a long time, it remained unclear how these were associated to classical optimal control policies exactly.

Drawing inspiration from the probabilistic perspective, one way of deriving a policy within CaI framework is to condition the action probability on the state and the (future) optimality variables. Levine (2018) showed that

these policies too are governed by a recursion that is structurally similar to dynamic programming yet shares similar tractability to Kárnỳ's approach. Alternatively, when the joint density, expressing the probability of states, actions and optimality variables, is interpreted as a target distribution, this framework also supports density matching. Neumann (2011) observed that different policies can be established through density matching by using either the forward (I-projection) or reverse divergence (M-projection).

This density-matching perspective, however, presents a conceptual departure from conventional optimal control design. Instead of *synthesizing* a control objective (e.g., an expected cost) whose minimization yields an optimal policy, with the resulting closed-loop trajectory distribution being a *byproduct*, this framework *starts* by postulating an ad-hoc optimal trajectory distribution and then *infers* a control policy to match it, using a divergence as the matching metric. This reversal of the design process naturally raises a fundamental question: *what interpretable, traditional control objective is actually being solved when one performs this density matching procedure?*

Rawlik et al. showed that the I-projection results into an entropy-regularized Stochastic Optimal Control (SOC) objective (Rawlik et al. (2012); Rawlik (2013)). Furthermore, when the I-projection is used to establish a fixed-point iteration, it was shown that the resulting sequence of policies converges to the corresponding SOC policy. Lefebvre (2024) demonstrated that the M-projection recovers the same policies as when one conditions the actions on the optimality variables. When a fixed-point iteration is established based on the M-projection, the resulting sequence of policies now converges to the Risk Seeking Optimal Control (RSOC), an alternative optimal control problem with an exponential cost establishing some notion of risk. Moreover, it was already known that the Maximum Likelihood Estimation (MLE) problem associated to the probabilistic graph model described above, which finds the policy that maximizes the probability of observing the optimality variables, is itself equivalent to a Risk Seeking Optimal Control problem (Toussaint (2009); Noorani and Baras (2022b); Watson et al. (2021); Watson and Peters (2021)). Lefebvre (2024) showed that the fixed-point iteration of the M-projection coincides with solving the MLE problem by means of Expectation Maximization.

Despite these many insights, the question of what interpretable control-theoretic objective is actually solved by the density-matching procedures of CaI, in particular the M-projection case, remained open until the recent work of Ito and Kashima (2024). They showed that performing variational inference with Rényi divergence (rather than KL) in the CaI framework yields a *log-probability regularized risk-sensitive control* problem, thereby providing a clear, interpretable objective for the otherwise opaque density-matching procedure. Our work offers a complementary and equally interpretable alternative interpretation. Still, two open questions remain. Firstly, it is unclear whether the various stochastic control formulations—the classical SOC/RSOC problems, the I-projection and the M-projection—can be unified into a single overarching mathematical structure. Secondly, KL-regularized control is closely linked to Distributionally Robust Control (DRC), where KL divergences define ambiguity sets for transition models, allowing the controller to hedge against model mismatches. Under certain conditions, the resulting DRC problem is equivalent to the Risk Averse Op-

timal Control problem formulation with an appropriate risk parameter (Nishimura et al. (2021); Zhang et al. (2024); Föllmer and Schied (2002); Noorani and Baras (2022a)). The relationship between DRC and the other KL-regularized formulations, such as those based on I- and M-projections, remains largely unexplored.

Closely related are also the path-integral formulations of Kappen (2005) and the linearly solvable MDP (LMDP) framework of Todorov (2006). Here the goal was not to regularize an optimal control problem but rather penalize the control cost through a KL divergence. It was shown that augmenting the objective with a KL penalty between controlled and passive dynamics results in optimal control problems whose Bellman equations are linear and can be evaluated by drawing samples from the passive system. These particular properties facilitate closed-form expressions for the policy beyond the Linear-Quadratic setting and inspired sampling-based solution methods (Theodorou et al. (2010); Williams et al. (2018)). Generalizing these properties to general optimal control problems turned out non-trivial. Lefebvre (2024) showed that similar properties are shared by the M-projection problem and are thus intimately linked to the point-of-view offered by the probabilistic graph model.

This paper addresses several of these gaps by showing that these different problems are unified by a broader class of KL-regularized optimal control problems. Our key observation is that prior formulations implicitly impose a *single* KL regularization term acting jointly on policies and transitions, or equivalently on entire trajectory distributions. In contrast, we consider a *strict generalization* in which (i) the KL penalties on policies and transitions are separated, and, (ii) each penalty is weighted independently. This formulation serves as a generative structure, from which classical SOC, RSOC, and soft-policy variants naturally emerge as special cases. Second, we show that the soft-policy formulations naturally majorize the original SOC and RSOC objectives through the Majorization–Minimization framework, therewith providing a principled foundation for designing iterative algorithms. Finally, we discuss the special case where the policy and transition KL weights are *identical*. In this synchronized setting several desirable properties emerge simultaneously. Our results therefore reveal that the structural harmony between policy and transition regularization, implicit in many earlier works, is not incidental but mathematically necessary for these special properties to hold.

2. NOTATION

We consider a discrete-time controlled stochastic dynamical system over a finite horizon, T. The behavior of the system is characterized by trajectories, where a trajectory is the sequence of states and actions, $\underline{\xi}_T = (x_0, u_0, \ldots, x_{T-1}, u_{T-1}, x_T)$.

One way to parametrize a distribution over a system's behavior/trajectories is via a sequence of policies, $\underline{\pi} = (\pi_0, \ldots, \pi_{T-1})$, and transition kernels, $\underline{\tau} = (\tau_0, \ldots, \tau_{T-1})$. The resulting trajectory distribution then is given by

$$p_{(\underline{\pi}, \underline{\tau})}(\underline{\xi}_T) = p(x_0) \prod_{t=0}^{T-1} \pi_t(u_t \mid x_t) \, \tau_t(x_{t+1} \mid \xi_t), \quad (1)$$

where $p(x_0)$ is initial state distribution and $\xi_t = (x_t, u_t)$.

For regularization purposes, we will also consider a baseline behavior, parametrized by a reference policy, $\underline{\rho}$, and

the system's transition kernels, $\underline{\iota}$. The corresponding baseline trajectory distribution then is given as

$$p_{(\underline{\rho},\underline{\iota})}(\underline{\xi}_T) = p(x_0) \prod_{t=0}^{T-1} \rho_t(u_t \mid x_t)\, \iota_t(x_{t+1} \mid \xi_t). \quad (2)$$

Along each trajectory, the system accumulates cost over time. The cumulative cost of a trajectory, $\underline{\xi}_T$, is defined as

$$\underline{c}_T(\underline{\xi}_T) = \sum_{t=0}^{T-1} c_t(\xi_t) + c_T(x_T), \quad (3)$$

where $c_t(\xi_t) \geq 0$ denotes the instantaneous cost at time t and $c_T(x_T) \geq 0$ is the terminal cost.

Further, we adopt a Risk-sensitive expectation operator with risk parameter, λ, using

$$\mathcal{R}_\mu^\lambda[f] := -\frac{1}{\lambda} \log \mathbb{E}_\mu\left[e^{-\lambda f}\right]. \quad (4)$$

Finally, the Kullback-Leibler divergence between two distributions, μ and σ, will be denoted by

$$\mathbb{D}_\sigma^\mu := \mathbb{E}_\mu[\log(\tfrac{\mu}{\sigma})]. \quad (5)$$

Throughout this paper, we adopt compact notation where function arguments are omitted for brevity. Instantaneous costs, c_t, policies, π_t, transition kernels, τ_t, value and state-action value functions (introduced subsequently) are to be interpreted with their appropriate domain variables.

3. PRELIMINARIES

In this section we recall two mathematical concepts that will serve as a foundation for developments later in the paper. First is the Risk Measures and their dual representations, which we repeatedly use to translate between the Entropic Risk Measure and a KL-regularized expectation-based form. The second is the Majorization–Minimization (MM) framework, which provides the conceptual foundation for later sections, where KL-regularized problems are used as tractable surrogates for classical stochastic control objectives.

3.1 Risk Measures and their Dual Representations

This section introduces risk measures, a theory which originated in mathematical finance (Föllmer and Schied (2002)), and provides a systematic framework for quantifying preferences beyond expected values and enable the encoding of risk-sensitive behavior in stochastic control. The dual representation of risk measures serves as a fundamental bridge connecting risk sensitivity with regularization.

Let $\mathcal{X}$ be a measurable space and $\mathcal{F}$ be the set of real-valued measurable functions on $\mathcal{X}$. A risk measure is a functional $\sigma : \mathcal{F} \to \mathbb{R}$. We work with distribution-dependent risk measures and write $\sigma_\rho(\cdot)$ to indicate dependence on a probability distribution $\rho \in \Delta(\mathcal{X})$.

Definition 1. A functional $\sigma_\rho : \mathcal{F} \to \mathbb{R}$ is called a *convex risk measure* if it satisfies the following properties for all $f, g \in \mathcal{F}$:

(1) **Monotonicity**: If $f(x) \leq g(x)$ for all $x \in \mathcal{X}$, then $\sigma_\rho(g) \leq \sigma_\rho(f)$.
(2) **Translation invariance**: For any $m \in \mathbb{R}$, $\sigma_\rho(f + m) = \sigma_\rho(f) - m$.
(3) **Convexity**: For any $\alpha \in [0,1]$, $\sigma_\rho(\alpha f + (1 - \alpha)g) \leq \alpha \sigma_\rho(f) + (1 - \alpha)\sigma_\rho(g)$.

Through standard duality theory, convex risk measures admit the following dual representation:

Theorem 2 (Dual Representation). A functional $\sigma_\rho : \mathcal{F} \to \mathbb{R}$ is a convex risk measure iff there exists a penalty function $\alpha_\rho : \Delta(\mathcal{X}) \to \mathbb{R}$ such that for all $f \in \mathcal{F}$:

$$-\sigma_\rho(f) = \inf_{\pi \ll \rho} \left\{ \mathbb{E}_\pi[f] + \alpha_\rho(\pi) \right\}, \quad (6)$$

where the infimum is taken over all probability measures π absolutely continuous with respect to ρ Föllmer and Schied (2002). The minimal penalty function is given by:

$$\alpha_\rho(\pi) = \sup_{f \in \mathcal{F}} \left\{ -\sigma_\rho(f) - \mathbb{E}_\pi[f] \right\}. \quad (7)$$

For our control applications, we will primarily use entropic risk measures with different signs of the risk parameter to encode risk-seeking and risk-averse behavior.

Example 3. (Entropic Risk Measures). We illustrate Theorem 2 using entropic risk measures that capture risk-seeking ($\lambda > 0$) and risk-averse ($\lambda < 0$) behavior.

Risk-Seeking Case ($\lambda > 0$): The entropic risk measure

$$\sigma_\rho^S(f) = \frac{1}{\lambda} \log \mathbb{E}_\rho\left[e^{-\lambda f}\right], \ \lambda > 0, \quad (8)$$

with penalty $\alpha_\rho(\pi) = \frac{1}{\lambda}\mathbb{D}(\pi\|\rho)$ admits the dual representation

$$-\sigma_\rho^S(f) = \mathcal{R}_\rho^\lambda[f] = \inf_{\pi \ll \rho} \left\{ \mathbb{E}_\pi[f] + \tfrac{1}{\lambda}\mathbb{D}(\pi\|\rho) \right\}, \quad (9)$$

with unique minimizer

$$\pi^* = \frac{\rho e^{-\lambda f}}{\mathbb{E}_\rho[e^{-\lambda f}]}. \quad (10)$$

Risk-Averse Case ($\lambda < 0$): Similarly, the entropic risk measure

$$\sigma_\rho^A(f) = -\frac{1}{\lambda} \log \mathbb{E}_\rho\left[e^{\lambda f}\right], \ \lambda < 0, \quad (11)$$

has the penalty function $\alpha_\rho(\pi) = -\frac{1}{\lambda}\mathbb{D}(\pi\|\rho)$ and, rewritten in sup-form, admits the dual representation

$$\sigma_\rho^A(-f) = \mathcal{R}_\rho^\lambda[f] = \sup_{\pi \ll \rho} \left\{ \mathbb{E}_\pi[f] + \tfrac{1}{\lambda}\mathbb{D}(\pi\|\rho) \right\}, \quad (12)$$

with the same extremal distribution (10).

We therefore have for $\lambda \in \mathbb{R} \setminus \{0\}$:

$$\mathcal{R}_\rho^\lambda[f] = \begin{cases} \inf_{\pi \ll \rho} \left\{ \mathbb{E}_\pi[f] + \tfrac{1}{\lambda}\mathbb{D}(\pi\|\rho) \right\}, \ \lambda > 0 \ \text{(risk-seeking)}, \\ \sup_{\pi \ll \rho} \left\{ \mathbb{E}_\pi[f] + \tfrac{1}{\lambda}\mathbb{D}(\pi\|\rho) \right\}, \ \lambda < 0 \ \text{(risk-averse)}. \end{cases}$$
$$(13)$$

The extremal distribution in both cases is given by:

$$\pi^* = \frac{\rho e^{-\lambda f}}{\mathbb{E}_\rho[e^{-\lambda f}]}. \quad (14)$$

The dual representations (13) provide a crucial link between risk sensitivity and regularization: the risk assessment of a cost function considers alternative distributions π, where deviations from ρ are penalized by the KL divergence, along with the direction of optimization (minimization vs. maximization) determining the risk attitude. This perspective also establishes a connection to DRO with KL ambiguity sets. Here a saddle point problem is solved to calculate the exact Lagrangian multiplier that corresponds with a predefined KL discrepancy. In that sense, Risk Sensitivity can be interpreted as a soft DRO problem where the Lagrangian multiplier is predefined instead.

In conclusion, to capture both risk-averse and -seeking behaviour in a unified notation, we introduce a single extremization operator that uses minimization and maximization depending on the sign of the risk parameter, λ:

$$\mathop{\mathrm{opt}}_{q \in \mathcal{Q}}^{\lambda} J(q) := \begin{cases} \min_{q \in \mathcal{Q}} J(q), & \lambda > 0 \text{ (risk-seeking)}, \\ \max_{q \in \mathcal{Q}} J(q), & \lambda < 0 \text{ (risk-averse)}, \end{cases}$$

and $\arg \mathrm{opt}^{\lambda}$ denotes the corresponding extremizer.

3.2 Majorization–Minimization (MM) Framework

The Majorization–Minimization framework addresses optimization problems of the form $\min_x F(x)$ particularly when F is difficult to optimize directly. MM algorithms iteratively minimize tractable surrogate functions that upper-bound F.

Given an iterate x^k, a function $G(\cdot \,|\, x^k)$ *majorizes* F at x^k if

$$G(x^k \,|\, x^k) = F(x^k), \tag{15a}$$

$$G(x \,|\, x^k) \geq F(x) \,\forall x, \tag{15b}$$

and the next iterate is defined by $x^k = \arg\min_x G(x \,|\, x^k)$.

Two features make MM attractive:

(1) **Descent guarantee:** Since minimization of G yields $G(x^{k+1}|x^k) \leq G(x^k|x^k)$, majorization implies that $F(x^{k+1}) \leq F(x^k)$.
(2) **Surrogate flexibility:** The surrogate G may be chosen to be a simpler tractable objective, allowing efficient updates.

In later sections, we will see that certain KL-regularized control problems naturally provide such for classical optimal control problems. This facilitates algorithms in which each iteration solves a tractable KL-regularized subproblem while guaranteeing descent on the original objective.

4. A UNIFYING KL-REGULARIZED CONTROL FORMULATION

In this section, we introduce a central KL-regularized control problem (C-KLR-CP) that serves as a generative central structure. By treating both the control policy and the transition dynamics as decision variables and by allowing their KL-based deviation penalties from specified baseline behaviors to be modulated, the central formulation provides a flexible template that subsumes classical problems and their regularized counterparts.

We consider the problem of jointly optimizing a sequence of control policies, $\underline{\pi}$, and a sequence of artificial transition kernels, $\underline{\tau}$, to minimize the expected cumulative trajectory cost while simultaneously penalizing deviations from a baseline behaviour specified by a reference policy sequence, $\underline{\rho}$, and the system's true transition kernel sequence, $\underline{\iota}$. Unlike standard optimal control formulations, the key idea here is to extend the decision variable to include not only the policy but also the transition kernels. The extension endows the control design with the capacity to reason about alternative system evolutions, optimistic or pessimistic in terms of potential cost accumulation. Clearly this encodes the notion of risk-sensitive behaviour.

The central KL-regularized control problem is defined as:

$$\min_{\underline{\pi}} \mathop{\mathrm{opt}}_{\underline{\tau}}^{\lambda^S} \mathbb{E}_{p_{(\underline{\pi},\underline{\tau})}} \Big[\underline{c}_T + \frac{1}{\lambda^P} \mathbb{D}_{\underline{\rho}}^{\underline{\pi}} + \frac{1}{\lambda^S} \mathbb{D}_{\underline{\iota}}^{\underline{\tau}} \Big], \tag{16}$$

where $\lambda^P > 0$ and $\lambda^S \in \mathbb{R} \setminus \{0\}$ control the strength and direction of regularization. The parameter λ^P governs deviations of the control policy from the baseline $\underline{\rho}$, and λ^S the deviations of the transition kernels from the baseline $\underline{\iota}$. The sign of λ^S determines whether these deviations are optimistic (risk-seeking, $\lambda^S > 0$) or pessimistic (risk-averse, $\lambda^S < 0$). In the most general case, the KL weights can be made time-dependent (λ_t^P, λ_t^S), allowing fine-grained control over baseline adherence and risk attitude throughout the horizon. While this time-varying extension enriches the family of tractable stochastic control problems, we defer its detailed discussion to future work and concentrate here on the time-homogeneous case.

Theorem 4 (Optimal Solution for C-KLR-CP). For the problem defined in (16), define the terminal value function $V_T = c_T$. Then, for $t = T-1, \ldots, 0$, the optimal value function V_t, action-value function Q_t, policy π_t^*, and transition kernel τ_t^* satisfy the following recursions:

$$V_t = \min_{\pi_t} \mathbb{E}_{\pi_t}[Q_t] + \frac{1}{\lambda^P} \mathbb{D}_{\rho_t}^{\pi_t} = \mathcal{R}_{\rho_t}^{\lambda^P}[Q_t], \tag{17a}$$

$$Q_t = \mathop{\mathrm{opt}}_{\tau_t}^{\lambda^S} \Big\{ \mathbb{E}_{\tau_t}[c_t + V_t] + \frac{1}{\lambda^S} \mathbb{D}_{\iota_t}^{\tau_t} \Big\} = \mathcal{R}_{\iota_t}^{\lambda^S}[c_t + V_t], \tag{17b}$$

$$\pi_t^* = \arg\min_{\pi_t} \mathbb{E}_{\pi_t}[Q_t] + \frac{1}{\lambda^P} \mathbb{D}_{\rho_t}^{\pi_t} = \rho_t \frac{e^{-\lambda^P Q_t}}{e^{-\lambda^P V_t}}, \tag{17c}$$

$$\tau_t^* = \arg\mathop{\mathrm{opt}}_{\tau_t}^{\lambda^S} \Big\{ \mathbb{E}_{\tau_t}[V_{t+1}] + \frac{1}{\lambda^S} \mathbb{D}_{\iota_t}^{\tau_t} \Big\} = \iota_t \frac{e^{-\lambda^S(c_t + V_t)}}{e^{-\lambda^S Q_t}}. \tag{17d}$$

Proof. See Appendix A.

Remark 5. This formulation represents a departure from the classical Probabilistic Control Design (PCD) paradigm Kárný (1996), where one begins with a prescribed optimal trajectory distribution and seeks policies that *match it*. In our approach, we do not presuppose the form of the optimal trajectory distribution; instead, it emerges organically as the solution to the optimization problem (16).

5. RECOVERING STANDARD STOCHASTIC CONTROL FORMULATIONS

The central KL-regularized control problem (16) serves as an expressive umbrella under which one can systematically represent, relate, and compare several classical stochastic control formulations by toggling (i) whether the control policy is regularized against a baseline reference $\underline{\rho}$, and (ii) whether the transition kernel sequence is fixed to the true dynamics $\underline{\iota}$ or left free as an optimization variable. We illustrate these formulations in the subsequent subsections. A comprehensive summary of the dynamic programming recursion of these formulations is provide in Table A.1.

5.1 No Policy Regularization: $\underline{\rho} = \underline{\pi}$

When no meaningful baseline policy is available (e.g., when designing a policy from scratch), the policy regularization term can be removed by setting the reference policy equal to the optimizing policy, yielding $\mathbb{D}_{\rho_t}^{\pi_t} = 0$. Two cases then

arise, depending on whether the transition kernels are fixed or free.

5.1.1. Stochastic Optimal Control (SOC)

Setting $\underline{\rho} = \underline{\pi}$ and constraining $\underline{\tau} = \underline{\iota}$, the central viewpoint yields the good old SOC problem

$$\min_{\underline{\pi}} \ \mathbb{E}_{p_{(\underline{\pi}, \underline{\iota})}}[\underline{c}_T]. \tag{18}$$

5.1.2. Risk-Sensitive Optimal Control (RSOC)

By setting $\underline{\rho} = \underline{\pi}$, while leaving $\underline{\tau}$ free, the central problem becomes

$$\min_{\underline{\pi}} \mathop{\mathrm{opt}}_{\underline{\tau}}^{\lambda^S} \mathbb{E}_{p_{(\underline{\pi}, \underline{\tau})}}[\underline{c}_T] + \frac{1}{\lambda^S} \mathbb{D}_{p_{(\underline{\pi}, \underline{\iota})}}^{p_{(\underline{\pi}, \underline{\tau})}}. \tag{19}$$

Equation (19) admits an equivalent two-step representation:

$$\min_{\underline{\pi}} \mathop{\mathrm{opt}}_{\underline{\tau}}^{\lambda^S} \mathbb{E}_{p_{(\underline{\pi}, \underline{\tau})}}[\underline{c}_T] + \frac{1}{\lambda^S} \mathbb{D}_{p_{(\underline{\pi}, \underline{\iota})}}^{p_{(\underline{\pi}, \underline{\tau})}}$$

$$:= \min_{\underline{\pi}} \left\{ \mathop{\mathrm{opt}}_{p(\underline{\xi}_T)}^{\lambda^S} \ \mathbb{E}_{p(\underline{\xi}_T)}[\underline{c}_T] + \frac{1}{\lambda^S} \mathbb{D}_{p_{(\underline{\pi}, \underline{\iota})}}^{p(\underline{\xi}_T)} \right\} \tag{20a}$$

$$:= \min_{\underline{\pi}} \left[-\frac{1}{\lambda^S} \log \mathbb{E}_{p_{(\underline{\pi}, \underline{\iota})}} \left[e^{-\lambda^S \underline{c}_T} \right] \right]. \tag{20b}$$

Equation (20b) is the standard risk-sensitive optimal control formulation. When $\lambda^S > 0$, it yields risk-seeking behavior, where the agent exhibits optimistic behavior by optimizing against favorable transitions. When $\lambda^S < 0$, it yields risk-averse behavior, where the agent exhibits pessimistic behavior by optimizing against worst-case transitions.

5.2 With Policy Regularization: $\underline{\rho} \neq \underline{\pi}$

When a baseline policy is available (e.g., from expert demonstrations, safe priors, or a stabilizing controller), regularization toward $\underline{\rho}$ is meaningful. Fixing or freeing $\underline{\tau}$ again yields two distinct formulations. We call the resulting policies *soft* because they appear as smoothed, exponentially tilted versions of the baseline that favor cost-minimizing actions.

5.2.1. Soft-Policy SOC (SP-SOC)

Fixing $\underline{\tau} = \underline{\iota}$ and taking $\underline{\rho}$ as a baseline policy, the central problem becomes

$$\min_{\underline{\pi}} \mathbb{E}_{p_{(\underline{\pi}, \underline{\iota})}}[\underline{c}_T] + \frac{1}{\lambda^P} \mathbb{D}_{p_{(\underline{\rho}, \underline{\iota})}}^{p_{(\underline{\pi}, \underline{\iota})}}. \tag{21}$$

Remark 6. This formulation can be interpreted as a density-matching problem. Specifically, it is equivalent to the I-projection problem Neumann (2011); Lefebvre (2024):

$$\min_{\underline{\pi}} \mathbb{D}_{p^*}^{p_{(\underline{\pi}, \underline{\iota})}} \tag{22}$$

where $p^* \propto p_{(\underline{\rho}, \underline{\iota})} e^{-\lambda^P \underline{c}_T}$ is the target distribution.

Note that when $\underline{\rho}$ is uniform, the KL term reduces to policy entropy, recovering the maximum entropy SOC formulation.

5.2.2. Soft-Policy Risk-Sensitive Control (SP-RSOC)

Allowing both $\underline{\pi}$ and $\underline{\tau}$ to be optimized with regularization parameters $\lambda^P > 0$ and $\lambda^S \neq 0$ produces:

$$\min_{\underline{\pi}} \mathop{\mathrm{opt}}_{\underline{\tau}}^{\lambda^S} \mathbb{E}_{p_{(\underline{\pi}, \underline{\tau})}}[\underline{c}_T] + \frac{1}{\lambda^P} \mathbb{D}_{\underline{\rho}}^{\underline{\pi}} + \frac{1}{\lambda^S} \mathbb{D}_{\underline{\iota}}^{\underline{\tau}}. \tag{23}$$

Here $\lambda^S > 0$ a risk-seeking behavior, while $\lambda^S < 0$ induces risk-averse behavior.

When $\lambda^P = |\lambda^S|$ with $\lambda^S \neq 0$, we get:

$$\min_{\underline{\pi}} \mathop{\mathrm{opt}}_{\underline{\tau}}^{\lambda^S} \mathbb{E}_{p_{(\underline{\pi}, \underline{\tau})}}[\underline{c}_T] + \frac{1}{|\lambda^S|} \mathbb{D}_{\underline{\rho}}^{\underline{\pi}} + \frac{1}{\lambda^S} \mathbb{D}_{\underline{\iota}}^{\underline{\tau}}. \tag{24}$$

again, where $\lambda^S > 0$ yields risk-seeking and $\lambda^S < 0$ yields risk-averse behavior. We denote these specific cases as *Synchronized Risk-Seeking SP-RSOC* (SRS-SP-RSOC) and *Synchronized Risk-Averse SP-RSOC* (SRA-SP-RSOC) respectively. The SRS-SP-RSOC formulation exhibits several remarkable properties which we detail in Section 7.

Remark 7. The SRS-SP-RSOC problem (24) admits the same solution as a density-matching problem based on the *M-projection* Neumann (2011); Lefebvre (2024):

$$\min_{\underline{\pi}} \mathbb{D}_{p_{(\underline{\pi}, \underline{\iota})}}^{p^*}, \tag{25}$$

where $p^* \propto p_{(\underline{\rho}, \underline{\iota})} e^{-\lambda^P \underline{c}_T}$, is the target distribution.

Remark 8. Ito and Kashima (2024) addressed the question of what performance index is implicitly optimized by the optimal (or target) trajectory distribution within the CaI framework. They approached this by performing variational inference using the Rényi divergence instead of the standard KL divergence. Their analysis revealed that the solution to a *log-probability regularized risk-sensitive control problem* with exponential utility coincides with the solution obtained from the CaI procedure, thereby providing a clear and interpretable control-theoretic objective for what was originally an inference-based design.

In using Renyi divergence based variational inference, Ito and Kashima (2024) also effectively extended the CaI framework to a *Risk-Sensitive Control as Inference* (RCaI) framework. While they provide a log-probability regularized risk-sensitive optimal control interpretation for the RCaI framework, our synchronized SP-RSOC formulation (24) offers a complementary and equally interpretable alternative interpretation: that of a regularized risk-sensitive optimal control problem where both the policy *and* the transition dynamics are treated as design variables, penalized via explicit KL divergences from their respective baselines.

5.3 Equivalences for Deterministic Dynamics

The relationships among SOC, RSOC, SP-SOC, and SP-RSOC become particularly simple when the baseline dynamics $\underline{\iota}$ are *deterministic*—each ι_t is a Dirac measure concentrated at the next state prescribed by the dynamics.

Recall the structural distinctions: for the central problem, one may either constrain the transition kernels to match the baseline (SOC/SP-SOC) or allow the transition kernels to be free design variables (RSOC/SP-RSOC).

When the baseline dynamics are deterministic, the RSOC and SP-RSOC formulations lose the expressive freedom associated with the auxiliary transition kernel. Because

the optimization over τ_t is always taken over distributions absolutely continuous with respect to ι_t, and ι_t is a Dirac measure, absolute continuity forces each feasible τ_t to *coincide* with ι_t. Thus the extremization over $\{\tau_t\}$ collapses to the same constraint that defines SOC and SP-SOC.

Consequently, for deterministic baseline dynamics, we have:

$$\text{SOC} \equiv \text{RSOC}, \quad \text{SP-SOC} \equiv \text{SP-RSOC}.$$

In both cases, the deterministic nature of the baseline removes the possibility of modifying transitions through τ_t, leaving only control over π_t. Hence RSOC adds no additional expressiveness beyond SOC, and likewise SP-RSOC adds none beyond SP-SOC. We refer to these deterministic special cases as Deterministic Optimal Control (DOC) and Soft-Policy DOC (SP-DOC).

6. MAJORIZATION OF STANDARD CONTROL FORMULATIONS

We now show that the KL-regularized formulations SP-SOC and SP-RSOC naturally serve as majorizers for their classical counterparts, SOC and RSOC respectively. This provides a principled interpretation: KL-regularized problems are not merely relaxations, but tractable surrogate objectives guaranteeing descent on the corresponding control objectives.

6.1 SP-SOC majorizes SOC.

We have the SOC objective $J_{\text{SOC}}(\underline{\pi}) = \mathbb{E}_{p_{(\underline{\pi},\underline{\iota})}}[\underline{c}_T]$. Given a current iterate $\underline{\pi}^k$, we can construct the surrogate with $\lambda^P > 0$:

$$J_{\text{SP-SOC}}(\underline{\pi}|\underline{\pi}^k) := \mathbb{E}_{p_{(\underline{\pi},\underline{\iota})}}[\underline{c}_T] + \frac{1}{\lambda^P}\mathbb{D}_{\underline{\pi}^k}^{\underline{\pi}}. \quad (26)$$

This is exactly the SP-SOC objective with the baseline $\underline{\rho}$ replaced by $\underline{\pi}^k$ and rewritten differently by cancelling the transition kernel $\underline{\iota}$ inside the KL divergence term. Since the KL divergence is nonnegative and vanishes only when $\pi_t = \pi_t^k$, this surrogate satisfies the MM conditions:

$$J_{\text{SP-SOC}}(\underline{\pi}^k \,|\, \underline{\pi}^k) = J_{\text{SOC}}(\underline{\pi}^k), \quad (27a)$$

$$J_{\text{SP-SOC}}(\underline{\pi} \,|\, \underline{\pi}^k) \geq J_{\text{SOC}}(\underline{\pi}) \quad \forall \underline{\pi}. \quad (27b)$$

Thus SP-SOC is a valid majorizer of SOC. Minimizing this surrogate yields the folllowing policy update that guarantees descent on the original SOC objective.

$$\underline{\pi}^k \quad \leftarrow \quad \arg\min_{\underline{\pi}} J_{\text{SP-SOC}}(\underline{\pi}|\underline{\pi}^k). \quad (28)$$

6.2 SP-RSOC majorizes RSOC.

We start from the RSOC objective (19). Rewriting it differently by canceling the policy $\underline{\pi}$ inside the KL term gives the equivalent form

$$J_{\text{RSOC}}(\underline{\pi},\underline{\tau}) := \mathbb{E}_{p_{(\underline{\pi},\underline{\tau})}}[\underline{c}_T] + \frac{1}{\lambda^S}\mathbb{D}_{\underline{\iota}}^{\underline{\tau}}. \quad (29)$$

Given a current policy iterate $\underline{\pi}^k$, we define the KL-regularized surrogate with $\lambda^P > 0$:

$$J_{\text{SP-RSOC}}(\underline{\pi},\underline{\tau}|\underline{\pi}^k) := \mathbb{E}_{p_{(\underline{\pi},\underline{\tau})}}[\underline{c}_T] + \frac{1}{\lambda^P}\mathbb{D}_{\underline{\pi}^k}^{\underline{\pi}} + \frac{1}{\lambda^S}\mathbb{D}_{\underline{\iota}}^{\underline{\tau}}. \quad (30)$$

This surrogate is exactly the SP-RSOC problem (23), with the substitution $\underline{\rho} = \underline{\pi}^k$. Because the additional term $\frac{1}{\lambda^P}\mathbb{D}_{\underline{\pi}^k}^{\underline{\pi}} \geq 0$, with equality iff $\underline{\pi} = \underline{\pi}^k$, the surrogate satisfies

$$J_{\text{SP-RSOC}}(\underline{\pi}^k,\underline{\tau}^k|\underline{\pi}^k) = J_{\text{RSOC}}(\underline{\pi}^k,\underline{\tau}^k), \quad (31a)$$

$$J_{\text{SP-RSOC}}(\underline{\pi},\underline{\tau}|\underline{\pi}^k) \geq J_{\text{RSOC}}(\underline{\pi},\underline{\tau}) \quad \forall(\underline{\pi},\underline{\tau}). \quad (31b)$$

Hence SP-RSOC majorizes RSOC, yielding the folllowing policy update that guarantees descent on the RSOC objective:

$$(\underline{\pi}^k,\underline{\tau}^k) \quad \leftarrow \quad \arg\min_{\underline{\pi},\underline{\tau}} J_{\text{SP-RSOC}}(\underline{\pi},\underline{\tau}|\underline{\pi}^k). \quad (32)$$

7. SPECIAL PROPERTIES OF SRS-SP-RSOC

The SRS-SP-RSOC formulation exhibits several remarkable structural properties that distinguish it from other classical stochastic control formulations. These features not only provide computational advantages but also enable powerful extensions such as compositional control design.

Note that these special properties hold only when the policy and transition regularization weights are equal ($\lambda^P = \lambda^S = \lambda > 0$).

These properties have previously been identified in case of, continuous-time Path Integral Control Kappen (2005) and in Linear Markov Games and SP-DOC case Dvijotham and Todorov (2012). However, our analysis reveals that these features extend to the broader class of SRS-SP-RSOC.

7.1 Linear Bellman Operator

A key property of SRS-SP-RSOC is that the Bellman equation characterizing the optimal value function becomes linear. From (17) , for SRS-SP-RSOC we have:

$$V_t = \mathcal{R}_{\rho_t}^{\lambda}[Q_t] = -\frac{1}{\lambda}\log\mathbb{E}_{\rho_t}\left[e^{-\lambda Q_t}\right], \quad (33a)$$

$$Q_t = \mathcal{R}_{\iota_t}^{\lambda}[c_t + V_{t+1}] = -\frac{1}{\lambda}\log\mathbb{E}_{\iota_t}\left[e^{-\lambda(c_t+V_{t+1})}\right]. \quad (33b)$$

Defining the *desirability function* $z_t := e^{-\lambda V_t}$ and the *instantaneous reward* $r_t := e^{-\lambda c_t}$, the SRS-SP-RSOC Bellman equation admits a linear form:

$$z_t = \mathbb{E}_{\rho_t}\left[r_t\,\mathbb{E}_{\iota_t}[z_{t+1}]\right]. \quad (34)$$

This linearity arises from the multiplicative structure of the exponential transform and represents a significant simplification compared to the nonlinear Bellman equations of standard stochastic optimal control.

7.2 Path Integral Solution

The linear Bellman equation (34) enables a *path integral* representation of the value function. Unlike backward dynamic programming, the desirability function can be expressed as an expectation over trajectories generated by the baseline policy and dynamics:

$$z_t = \mathbb{E}_{p_{(\underline{\rho}_t,\underline{\iota}_t|x_t)}}\left[e^{-\lambda\left(\sum_{k=t}^{T-1}c_k+c_T\right)}\right], \quad (35)$$

where $p_{(\underline{\rho}_t,\underline{\iota}_t|x_t)}$ denotes the trajectory distribution obtained by composing the baseline policy $\underline{\rho}_t = (\rho_t,\ldots,\rho_{T-1})$

and baseline transitions $\underline{\iota}_t = (\iota_t, \ldots, \iota_{T-1})$ from time t onward, conditioned on x_t.

This path integral representation has several implications:

- **Forward Simulation**: The value function can be estimated via forward sampling of trajectories, bypassing the need for backward dynamic programming.
- **Parallelization**: Multiple trajectories can be simulated independently, enabling efficient parallel computation.
- **Model-Free Estimation**: When the baseline policy is known but the value function is unknown, Monte Carlo estimation can provide approximate solutions.

The optimal policy in this representation takes an especially elegant form:

$$\pi_t^* = \rho_t \frac{r_t \, \mathbb{E}_{\iota_t}[z_{t+1}]}{z_t}, \tag{36}$$

which can be interpreted as the baseline policy ρ_t being reweighted by the likelihood given by the product of immediate reward r_t and expected future desirability $\mathbb{E}_{\iota_t}[z_{t+1}]$.

7.3 Compositionality of Value Functions and Policies

Perhaps the most powerful feature of SRS-SP-RSOC is its inherent compositionality: the ability to combine solutions to simpler subproblems into solutions for more complex problems. This property emerges directly from the linear structure of the desirability Bellman equation.

Consider a terminal cost function that decomposes as a weighted combination of N component costs:

$$e^{-\lambda c_T} = \sum_{n=1}^{N} \gamma_n e^{-\lambda c_T^{(n)}}, \tag{37}$$

where $\gamma_n > 0$ are arbitrary positive weights. This represents a flexible way to combine multiple terminal objectives.

Define component desirability functions recursively:

$$z_T^{(n)} := \gamma_n e^{-\lambda c_T^{(n)}}, \tag{38a}$$

$$z_t^{(n)} := \mathbb{E}_{\rho_t}\left[r_t \, \mathbb{E}_{\iota_t}\left[z_{t+1}^{(n)}\right]\right], \quad t = T-1, \ldots, 0. \tag{38b}$$

Theorem 9 (Compositionality of SRS-SP-RSOC). If the terminal desirability decomposes as $z_T = \sum_{n=1}^{N} z_T^{(n)}$ with weights γ_n, then for all $t = 0, \ldots, T-1$, the desirability function maintains the decomposition:

$$z_t = \sum_{n=1}^{N} z_t^{(n)}, \tag{39}$$

and the optimal policy decomposes as a mixture:

$$\pi_t^* = \sum_{n=1}^{N} w_t^{(n)} \pi_t^{(n)}, \tag{40}$$

where the mixture weights are given by $w_t^{(n)} = \frac{z_t^{(n)}}{z_t}$ with $\sum_{n=1}^{N} w_t^{(n)} = 1$, and each component policy is given by:

$$\pi_t^{(n)} = \rho_t \frac{r_t \, \mathbb{E}_{\iota_t}[z_{t+1}^{(n)}]}{z_t^{(n)}}. \tag{41}$$

Proof. See Appendix B.

7.4 Maximum Likelihood Estimation on a PGM

We conclude by establishing a final connection with CaI framework Toussaint and Storkey (2006); Levine (2018);

Lefebvre (2024). Central to Control as Inference is a probabilistic graphical model that introduces a set of binary *optimality variables* $\mathcal{O}_t \in \{0, 1\}$, representing whether the state-action pair ξ_t is optimal. These optimality variables give rise to a joint probability model over states, inputs, and optimality variables:

$$p(\underline{\xi}_T, \underline{\mathcal{O}}_T; \underline{\rho}, \underline{\iota}) = p_{\underline{\rho}, \underline{\iota}} p(\underline{\mathcal{O}}_T | \underline{\xi}_T) \tag{42}$$

where the likelihood term encodes optimality via:

$$p(\underline{\mathcal{O}}_T = \underline{1}_T | \underline{\xi}_T) \propto e^{-\lambda \underline{c}_T}, \; \lambda > 0 \tag{43}$$

There are several ways to connect this probabilistic model to the body of work covered so far, which we describe ahead.

Posterior Equivalence: Note that for SRS-SP-RSOC setting with $\lambda^P = \lambda^S = \lambda > 0$, the optimization problem becomes:

$$\min_{\underline{\pi}, \underline{\tau}} \mathbb{E}_{p_{(\underline{\pi}, \underline{\tau})}}[\underline{c}_T] + \frac{1}{\lambda} \mathbb{D}_{p_{(\underline{\rho}, \underline{\iota})}}^{p_{(\underline{\pi}, \underline{\tau})}}, \tag{44}$$

Denoting $q := p_{(\underline{\pi}, \underline{\tau})}$ and $p_b := p_{(\underline{\rho}, \underline{\iota})}$, (44) can be rewritten as:

$$\min_{q \ll p_b} \mathbb{E}_q[\underline{c}_T] + \frac{1}{\lambda} \mathbb{D}_{p_b}^q. \tag{45}$$

Using (14), the unique minimizer q^* is given by:

$$q^* = \frac{p_b e^{-\lambda \underline{c}_T}}{\mathbb{E}_{p_b}[e^{-\lambda \underline{c}_T}]}. \tag{46}$$

This is exactly the posterior distribution of states and actions given optimality in the CaI model:

$$q^* = p(\underline{\xi}_T | \underline{\mathcal{O}}_T = 1; \underline{\rho}, \underline{\iota}), \tag{47}$$

Density matching: The density q^* (or the CaI posterior) can be used as a *target distribution* for density matching. The forward KL minimization of $\mathbb{D}_{q^*}^{p_{(\underline{\pi}, \underline{\iota})}}$ recovers SP-SOC, while reverse KL minimization of $\mathbb{D}_{p_{(\underline{\pi}, \underline{\iota})}}^{q^*}$ recovers SRS-SP-RSOC.

MLE interpretation: The equivalence above gives rise to an alternative interpretation of the RSOC problem as a *Maximum Likelihood Estimation* (MLE) problem. Maximizing the likelihood of observing optimality under policy $\underline{\pi}$, i.e.:

$$\max_{\underline{\pi}} \log p(\underline{\mathcal{O}} = 1; \underline{\pi}, \underline{\iota}), \tag{48}$$

can easily be shown to be equivalent to the risk-seeking RSOC problem (20b). This MLE problem can be solved efficiently via the Expectation-Maximization algorithm, where the E-step computes the posterior $q^k(\underline{\xi}_T) = p(\underline{\xi}_T | \underline{\mathcal{O}}_T; \underline{\pi}^k)$ and the M-step updates the policy to match this posterior $\underline{\pi}^{k+1} = \arg\max_{\underline{\pi}} \mathbb{E}_{q^k}[\log p(\underline{\xi}_T, \underline{\mathcal{O}}_T; \underline{\pi})]$.

Conditional policy form: Finally, the optimal SRS-SP-RSOC policy coincides precisely with the policy obtained by conditioning the input on the current state and future optimality in the CaI setting:

$$\pi_t^{\text{SRS-SP-RSOC}}(u_t | x_t) = p(u_t | x_t, \underline{\mathcal{O}}_T = 1; \underline{\rho}, \underline{\iota}). \tag{49}$$

This conditional distribution can be computed via Bayesian smoothing on the PGM, revealing why the SRS-SP-RSOC dynamic programming recursions take the form of backward message-passing recursions.

The connection with exact inference on a PGM explains why the synchronized SRS-SP-RSOC setting enjoys particularly favorable properties: linear Bellman equations, path integral solutions, and compositionality. This provides a unified probabilistic foundation for understanding the relationships between risk-sensitive control, density matching, and Bayesian inference.

8. CONCLUSION

This paper has presented a unified perspective on stochastic optimal control by introducing a central KL-regularized problem that serves as a generative core for various control formulations. By treating both policies and transition kernels as decision variables, and by separating their KL penalties with independently chosen weights, we obtained a strict generalization of the standard trajectory-level KL-regularization commonly used in probabilistic, entropy-regularized, and risk-sensitive control.

Within this general formulation, classical problems such as SOC and RSOC, along with their policy-regularized counterparts SP-SOC and SP-RSOC, emerge from different structural restrictions. The Majorization-Minimization interpretation provides a principled foundation for viewing these regularized formulations as tractable surrogates that majorize the original control objectives, guaranteeing descent.

A key insight from our analysis is that regularization of the transition kernels and policies need not be equivalent. Put differently the convergence rate of the fixed-point iteration and risk parameter of the corresponding RSOC problem need not be synchronised. Our analysis also shows that in case they are synchronised, the associated Risk Seeking Soft-Policy RSOC problem exhibits favourable computational properties, including linear Bellman equations, path integral solutions, and compositionality.

In future work we will explore how allowing the regularization weights to vary with time yields new problem classes, enriching the family of standard stochastic control formulations beyond the usual time-homogeneous cases. These time-varying formulations provide finer control over policy and transition regularization throughout the horizon, enabling more flexible and adaptive control strategies while maintaining the structural benefits of the unified framework. Other directions could include (i) shifting from soft regularization to hard regularization by introduction of a hard constraint on the KL policy discrepancy, and, (ii) integration of the notion of partial observability into the general framework. Future practical work could explore how to exploit these insights for improved algorithmic development in many applications areas that require efficient solution of optimal control problems.

REFERENCES

Dvijotham, K. and Todorov, E. (2012). Linearly solvable optimal control. *Reinforcement learning and approximate dynamic programming for feedback control*, 119–141.

Föllmer, H. and Schied, A. (2002). Convex measures of risk and trading constraints. *Finance and stochastics*, 6(4), 429–447.

Ito, K. and Kashima, K. (2024). Risk-sensitive control as inference with rényi divergence. *Advances in Neural Information Processing Systems*, 37, 71381–71413.

Kappen, H.J. (2005). Linear theory for control of nonlinear stochastic systems. *Physical review letters*, 95(20), 200201.

Kárný, M. (1996). Towards fully probabilistic control design. *Automatica*, 32(12), 1719–1722.

Kárný, M. and Guy, T.V. (2006). Fully probabilistic control design. *Systems & Control Letters*, 55(4), 259–265.

Lefebvre, T. (2024). Probabilistic control and majorisation of optimal control. *Systems & Control Letters*, 190, 105837.

Levine, S. (2018). Reinforcement learning and control as probabilistic inference. *arXiv:1805.00909*.

Levine, S. and Koltun, V. (2013a). Guided policy search. In *Int. Conf. Mach. Learn.*, 1–9.

Levine, S. and Koltun, V. (2013b). Variational policy search via trajectory optimization. *Advances in neural information processing systems*, 26.

Neumann, G. (2011). Variational inference for policy search in changing situations. In *International conference on machine learning*, 817–824.

Nishimura, H., Mehr, N., Gaidon, A., and Schwager, M. (2021). Rat ilqr: A risk auto-tuning controller. *IEEE Robotics and Automation Letters*, 6(2), 763–770.

Noorani, E. and Baras, J.S. (2022a). Embracing risk in reinforcement learning: Connection between risk-sensitive exponential and distributionally robust criteria. In *Amer. Control Conf. (ACC)*, 2703–2708.

Noorani, E. and Baras, J.S. (2022b). A probabilistic perspective on risk-sensitive reinforcement learning. In *Amer. Control Conf. (ACC)*, 2697–2702.

Rawlik, K., Toussaint, M., and Vijayakumar, S. (2012). Stochastic optimal control by approximate inference. *Robotics: Science and Systems*.

Rawlik, K.C. (2013). Probabilistic inference approaches to stochastic optimal control.

Theodorou, E., Buchli, J., and Schaal, S. (2010). Generalized path integral control. *Journal of Machine Learning Research*, 11, 3137–3181.

Todorov, E. (2006). Linearly-solvable markov decision problems. *Advances in neural information processing systems*, 19.

Toussaint, M. (2009). Robot trajectory optimization via approximate inference. In *International conference on machine learning*, 1049–1056.

Toussaint, M. and Storkey, A. (2006). Probabilistic inference for mdps. In *international conference on machine learning*, 945–952.

Watson, J., Abdulsamad, H., Findeisen, R., and Peters, J. (2021). Efficient stochastic optimal control through bayesian input inference. *arXiv:2105.07693*.

Watson, J. and Peters, J. (2021). Advancing trajectory optimization via approximate inference. In *Amer. Control Conf. (ACC)*, 1231–1236.

Williams, G., Drews, P., Goldfain, B., Rehg, J.M., and Theodorou, E.A. (2018). Information-theoretic mpc. *IEEE Transactions on Robotics*, 34(6), 1603–1622.

Zhang, R., Hu, Y., and Li, N. (2024). Soft robust mdps and risk-sensitive mdps. In *International Conference on Learning Representations*.

Appendix A. PROOF OF THEOREM

We prove the dynamic programming recursion for the central KL-regularized control problem (16) by backward induction on the time index t.

Notation: For $t = 0, 1, \ldots, T-1$, let:

- $\underline{\pi}_t = (\pi_t, \pi_{t+1}, \ldots, \pi_{T-1})$ denote the policy sequence from time t onward
- $\underline{\tau}_t = (\tau_t, \tau_{t+1}, \ldots, \tau_{T-1})$ denote the transition kernel sequence from time t onward
- $p_{(\underline{\pi}_t, \underline{\tau}_t)}$ denote the conditional distribution of the trajectory segment from t to T, given x_t

Define the cost-to-go from state x_t under policy and transition sequences $(\underline{\pi}_t, \underline{\tau}_t)$ as:

$$J_t(x_t; \underline{\pi}_t, \underline{\tau}_t) = \mathbb{E}_{p_{(\underline{\pi}_t, \underline{\tau}_t)}} \left[c_T + \sum_{k=t}^{T-1} \left(c_t + \frac{1}{\lambda^P} \mathbb{D}_{\rho_t}^{\pi_t} + \frac{1}{\lambda^S} \mathbb{D}_{\iota_t}^{\tau_t} \right) \right]$$
(A.1)

The optimal cost-to-go function is:

$$V_t(x_t) = \min_{\underline{\pi}_t} \underset{\underline{\tau}_t}{\overset{\lambda^S}{\mathrm{opt}}} \, J_t(x_t; \underline{\pi}_t, \underline{\tau}_t),$$
(A.2)

Proof Structure: We will prove by backward induction that V_t satisfies the recursions in (17). The proof proceeds in three stages:

(1) **Base Case** $(t = T)$: Define the terminal condition.
(2) **Inductive Hypothesis**: Assume the recursion holds at time $t+1$.
(3) **Inductive Step**: Prove the recursion holds at time t.

Base Case $(t = T)$: At the terminal time T, there are no decisions to make. By definition:

$$V_T(x_T) = c_T(x_T).$$
(A.3)

This serves as the initial condition for the backward recursion.

Inductive Hypothesis: For the inductive proof, we assume that for time $t+1$ (where $0 \leq t+1 \leq T-1$), the function V_{t+1} exists and satisfies the recursions (17). Specifically, we assume:

$$V_{t+1} = \min_{\pi_{t+1}} \left[\mathbb{E}_{\pi_{t+1}}[Q_{t+1}] + \frac{1}{\lambda^P} \mathbb{D}_{\rho_{t+1}}^{\pi_{t+1}} \right] = \mathcal{R}_{\rho_{t+1}}^{\lambda^P}[Q_{t+1}],$$
(A.4a)

$$Q_{t+1} = \underset{\tau_{t+1}}{\overset{\lambda^S}{\mathrm{opt}}} \left\{ \mathbb{E}_{\tau_{t+1}}[c_{t+1} + V_{t+2}] + \frac{1}{\lambda^S} \mathbb{D}_{\iota_{t+1}}^{\tau_{t+1}} \right\}$$

$$= \mathcal{R}_{\iota_{t+1}}^{\lambda^S}[c_{t+1} + V_{t+2}],$$
(A.4b)

$$\pi_{t+1}^* = \rho_{t+1} \frac{e^{-\lambda^P Q_{t+1}}}{e^{-\lambda^P V_{t+1}}},$$
(A.4c)

$$\tau_{t+1}^* = \iota_{t+1} \frac{e^{-\lambda^S(c_{t+1} + V_{t+2})}}{e^{-\lambda^S Q_{t+1}}}.$$
(A.4d)

This hypothesis is trivially true for $t+1 = T-1$ when we take V_T as defined in base case, and use dual representations (13) and the extremal distribution (14).

Inductive Step: We now prove that if the inductive hypothesis holds for $t+1$, then the recursions (17) hold for t.

By the principle of optimality (Bellman's principle), an optimal policy from time t onward must consist of an optimal immediate decision at time t followed by optimal decisions from time $t+1$ onward. Therefore:

$$V_t(x_t) = \min_{\pi_t} \underset{\tau_t}{\overset{\lambda^S}{\mathrm{opt}}} \, \mathbb{E}_{\pi_t, \tau_t} \left[c_t + \frac{1}{\lambda^P} \mathbb{D}_{\rho_t}^{\pi_t} \right.$$

$$\left. + \frac{1}{\lambda^S} \mathbb{D}_{\iota_t}^{\tau_t} + \underbrace{\min_{\underline{\pi}_{t+1}} \underset{\underline{\tau}_{t+1}}{\overset{\lambda^S}{\mathrm{opt}}} \, J_{t+1}(x_{t+1}; \underline{\pi}_{t+1}, \underline{\tau}_{t+1})}_{=V_{t+1}(x_{t+1})} \right].$$
(A.5)

For fixed (x_t, u_t), define $f(x_{t+1}) = c_t(x_t, u_t) + V_{t+1}(x_{t+1})$. The term involving τ_t in (A.5) is:

$$\underset{\tau_t}{\overset{\lambda^S}{\mathrm{opt}}} \left\{ \mathbb{E}_{\tau_t}[f] + \frac{1}{\lambda^S} \mathbb{D}_{\iota_t}^{\tau_t} \right\}.$$
(A.6)

This is exactly of the form in the dual representation of the entropic risk measure (13). Therefore, using this dual representation:

$$\underset{\tau_t}{\overset{\lambda^S}{\mathrm{opt}}} \left\{ \mathbb{E}_{\tau_t}[f] + \frac{1}{\lambda^S} \mathbb{D}_{\iota_t}^{\tau_t} \right\} = \mathcal{R}_{\iota_t}^{\lambda^S}[f].$$
(A.7)

Denote this optimal value by:

$$Q_t(x_t, u_t) = \mathcal{R}_{\iota_t}^{\lambda^S}[c_t + V_{t+1}] = -\frac{1}{\lambda^S} \log \mathbb{E}_{\iota_t} \left[e^{-\lambda^S(c_t + V_{t+1})} \right].$$
(A.8)

Moreover, the optimal transition kernel τ_t^* achieving this extremum is given by the exponential tilting formula from (14):

$$\tau_t^* = \iota_t \frac{e^{-\lambda^S(c_t + V_{t+1})}}{\mathbb{E}_{\iota_t}[e^{-\lambda^S(c_t + V_{t+1})}]} = \iota_t \frac{e^{-\lambda^S(c_t + V_{t+1})}}{e^{-\lambda^S Q_t}}$$
(A.9)

Now the expression for V_t simplifies to:

$$V_t(x_t) = \min_{\pi_t} \left[\mathbb{E}_{\pi_t}[Q_t] + \frac{1}{\lambda^P} \mathbb{D}_{\rho_t}^{\pi_t} \right].$$
(A.10)

Applying the dual representation (13):

$$V_t(x_t) = \mathcal{R}_{\rho_t}^{\lambda^P}[Q_t] = -\frac{1}{\lambda^P} \log \mathbb{E}_{\rho_t} \left[e^{-\lambda^P Q_t} \right].$$
(A.11)

The optimal policy achieving this minimum is given by the exponential tilting formula:

$$\pi_t^* = \rho_t \frac{e^{-\lambda^P Q_t}}{\mathbb{E}_{\rho_t}[e^{-\lambda^P Q_t}]} = \rho_t \frac{e^{-\lambda^P Q_t}}{e^{-\lambda^P V_t}}.$$
(A.12)

We therefore have:

Table A.1. Dynamic programming recursions for stochastic control formulations

	SOC	SP-SOC	RSOC	SP-RSOC	DOC	SP-DOC
V_t	$\min_{\pi_t} \mathbb{E}_{\pi_t}[Q_t]$	$\min_{\pi_t} \mathbb{E}_{\pi_t}[Q_t] + \frac{1}{\lambda^P}\mathbb{D}_{\rho_t}^{\pi_t}$ $= \mathcal{R}_{\rho_t}^{\lambda^P}[Q_t]$	$\min_{\pi_t} \mathbb{E}_{\pi_t}[Q_t]$	$\min_{\pi_t} \mathbb{E}_{\pi_t}[Q_t] + \frac{1}{\lambda^S}\mathbb{D}_{\rho_t}^{\pi_t}$ $= \mathcal{R}_{\rho_t}^{\lambda^S}[Q_t]$	$\min_{\pi_t} \mathbb{E}_{\pi_t}[Q_t]$	$\min_{\pi_t} \mathbb{E}_{\pi_t}[Q_t] + \frac{1}{\lambda^P}\mathbb{D}_{\rho_t}^{\pi_t}$ $= \mathcal{R}_{\rho_t}^{\lambda^P}[Q_t]$
Q_t	$c_t + \mathbb{E}_{\iota_t}[V_{t+1}]$	$c_t + \mathbb{E}_{\iota_t}[V_{t+1}]$	$\underset{\tau_t}{\overset{\lambda^S}{\mathrm{opt}}}\, \mathbb{E}_{\tau_t}[c_t + V_{t+1}] + \frac{1}{\lambda^S}\mathbb{D}_{\iota_t}^{\tau_t}$ $= \mathcal{R}_{\iota_t}^{\lambda^S}[c_t + V_{t+1}]$	$\underset{\tau_t}{\overset{\lambda^S}{\mathrm{opt}}}\, \mathbb{E}_{\tau_t}[c_t + V_{t+1}] + \frac{1}{\lambda^S}\mathbb{D}_{\iota_t}^{\tau_t}$ $= \mathcal{R}_{\iota_t}^{\lambda^S}[c_t + V_{t+1}]$	$c_t + V_{t+1}$	$c_t + V_{t+1}$
π_t^*	$\arg\min_{\pi_t} \mathbb{E}_{\pi_t}[Q_t]$	$\rho_t \dfrac{e^{-\lambda^P Q_t}}{e^{-\lambda^P V_t}}$	$\arg\min_{\pi_t} \mathbb{E}_{\pi_t}[Q_t]$	$\rho_t \dfrac{e^{-\lambda^S Q_t}}{e^{-\lambda^S V_t}}$	$\arg\min_{\pi_t} \mathbb{E}_{\pi_t}[Q_t]$	$\rho_t \dfrac{e^{-\lambda^P Q_t}}{e^{-\lambda^P V_t}}$
τ_t^*	$:= \iota_t$	$:= \iota_t$	$\iota_t \dfrac{e^{-\lambda^S (c_t + V_{t+1})}}{e^{-\lambda^S Q_t}}$	$\iota_t \dfrac{e^{-\lambda^S (c_t + V_{t+1})}}{e^{-\lambda^S Q_t}}$	ι_t	ι_t

$$V_t = \min_{\pi_t}\left[\mathbb{E}_{\pi_t}[Q_t] + \frac{1}{\lambda^P}\mathbb{D}_{\rho_t}^{\pi_t}\right] = \mathcal{R}_{\rho_t}^{\lambda^P}[Q_t], \qquad \text{(A.13a)}$$

$$Q_t = \underset{\tau_t}{\overset{\lambda^S}{\mathrm{opt}}}\left\{\mathbb{E}_{\tau_t}[c_t + V_{t+1}] + \frac{1}{\lambda^S}\mathbb{D}_{\iota_t}^{\tau_t}\right\} = \mathcal{R}_{\iota_t}^{\lambda^S}[c_t + V_{t+1}], \qquad \text{(A.13b)}$$

$$\pi_t^* = \rho_t \frac{e^{-\lambda^P Q_t}}{e^{-\lambda^P V_t}}, \qquad \text{(A.13c)}$$

$$\tau_t^* = \iota_t \frac{e^{-\lambda^S (c_t + V_{t+1})}}{e^{-\lambda^S Q_t}}. \qquad \text{(A.13d)}$$

These are exactly the recursions in (17).

We have therefore shown:

(1) A terminal condition: $V_T = c_T$.
(2) If V_{t+1} satisfies the recursions (17), then V_t also satisfies them.
(3) The recursions (17) hold for V_{T-1}.

Therefore, by backward induction, the optimal solution of the central KL-regularized problem (16) is given by the recursions in Theorem 4.

Appendix B. PROOF OF THEOREM

The proof proceeds by induction. The base case $t = T$ holds by definition. Assume the decomposition holds at time $t + 1$, so $z_{t+1} = \sum_{n=1}^{N} z_{t+1}^{(n)}$. Then:

$$z_t = \mathbb{E}_{\rho_t}[r_t \mathbb{E}_{\iota_t}[z_{t+1}]]$$

$$= \mathbb{E}_{\rho_t}\left[r_t \mathbb{E}_{\iota_t}\left[\sum_{n=1}^{N} z_{t+1}^{(n)}\right]\right]$$

$$= \sum_{n=1}^{N} \mathbb{E}_{\rho_t}\left[r_t \mathbb{E}_{\iota_t}\left[z_{t+1}^{(n)}\right]\right]$$

$$= \sum_{n=1}^{N} z_t^{(n)}.$$

For the policy decomposition we get:

$$\pi_t^* = \rho_t \frac{r_t \mathbb{E}_{\iota_t}[z_{t+1}]}{z_t}$$

$$= \rho_t \frac{r_t \mathbb{E}_{\iota_t}\left[\sum_{n=1}^{N} z_{t+1}^{(n)}\right]}{\sum_{m=1}^{N} z_t^{(m)}}$$

$$= \sum_{n=1}^{N} \frac{z_t^{(n)}}{\sum_{m=1}^{N} z_t^{(m)}} \cdot \rho_t \frac{r_t \mathbb{E}_{\iota_t}[z_{t+1}^{(n)}]}{z_t^{(n)}}$$

$$= \sum_{n=1}^{N} w_t^{(n)} \pi_t^{(n)},$$

where $w_t^{(n)} = \frac{z_t^{(n)}}{z_t}$ and $\sum_{n=1}^{N} w_t^{(n)} = 1$ since $z_t = \sum_{n=1}^{N} z_t^{(n)}$.