Tractable Optimism: From Structural Insights in Probabilistic Optimal Control Towards Efficient Exploration in Model-Based RL

Ajinkya Bhole, Mohammad Mahmoudi Filabadi, Guillaume Crevecoeur and Tom Lefebvre

***Abstract*— A key challenge in model-based reinforcement learning (MBRL) is the exploration–exploitation tradeoff: an agent must collect informative data to improve its dynamics model while simultaneously minimizing cumulative cost. A principled approach to provably efficient exploration is the optimistic strategy, which selects policies that are optimal under the most favorable plausible model. While theoretically sound, this strategy requires solving a joint optimization over policies and model classes that is generally intractable. In this paper, we outline how insights into the connections of various paradigms within probabilistic optimal control can be brought to bear on this challenge. The key observation is that the optimistic exploration problem can be relaxed into a fully regularized probabilistic control formulation that naturally admits tractable analytical solutions. We survey its main properties, connect them to the regret minimization for efficient MBRL, and identify open questions for future investigation.**

I. Introduction

Model-based reinforcement learning (MBRL) offers a promising route to sample-efficient decision making by maintaining an explicit dynamics model that is refined as new data becomes available [1], [2]. A central challenge in MBRL is the *exploration–exploitation tradeoff*: the agent must decide whether to exploit its current model to minimize cost, or to deliberately seek out informative data that will improve the model for future episodes.

A principled approach to provably efficient exploration is *optimism in the face of uncertainty*, formalized through the Upper Confidence Reinforcement Learning (UCRL) framework [3]. The idea is simple yet powerful: maintain a set $\mathcal{T}_n$ of dynamical models consistent with the data collected so far, then select the policy that is optimal under the *most favorable* model in this set. This optimistic strategy leads to sublinear cumulative regret, i.e., $\mathcal{R}_N = O(\sqrt{N})$, guaranteeing that the agent's performance converges to the true optimum.

Despite this theoretical elegance, implementing optimistic exploration in practice is challenging. The joint optimization over policies and plausible models is generally intractable for nonlinear dynamics with continuous state and action spaces. Within the episodic regret framework, two dominant strategies have emerged: *optimistic exploration* [1], [4], which selects the most favorable plausible model and plans against it so that uncertain regions are visited as a byproduct of optimism, and *intrinsic reward* [5], [6], which keeps the model fixed and instead subtracts an exploration bonus proportional to model uncertainty from the cost. The exploration problem has also been studied from adjacent perspectives: system identification that designs maximally informative experiments [7], [8], optimality gap minimization that weights experiment design by downstream task impact rather than pure estimation accuracy [9]–[11], and curiosity-driven approaches that reward learning progress rather than uncertainty reduction [12].

In parallel, our recent work [13] has developed a consolidated map of probabilistic optimal control, revealing deep structural connections between seemingly disparate formulations: including Stochastic Optimal Control (SOC), Risk-Sensitive Optimal Control (RSOC), and Distributionally Robust Control (DRC), among others. The map is anchored in a central control problem that jointly optimizes over policies and auxiliary transition kernels, with independently weighted divergence penalties. By toggling structural decisions within this formulation, one navigates between these paradigms. This central perspective clarifies how entropy regularization acts across various frameworks as a fundamental mathematical tool to bound information processing and tame computational complexity. Ultimately, establishing these connections bring the notions of risk, regularization, and robustness to the same table.

In this paper, we bring these structural insights to bear on the exploration problem in MBRL. We show that the traditionally intractable optimistic exploration problem can be relaxed into a risk-seeking instance of the central formulation from [13], directly unlocking its tractable path integral solutions and softmax policies for exploration. Moreover, a second-order analysis of this risk-seeking formulation reveals that it naturally bridges optimism and intrinsic reward within a single structural template. We survey these results and identify key open questions for future investigation.

II. A Central Control Formulation

We consider a discrete-time, finite-horizon stochastic control problem. The system state evolves according to transition kernels $\underline{\tau} = (\tau_0, \dots, \tau_{T-1})$, and the agent selects actions according to a policy sequence $\underline{\pi} = (\pi_0, \dots, \pi_{T-1})$. The resulting trajectory distribution is $p_{(\underline{\pi},\underline{\tau})}(\underline{\xi}_T) = p(x_0) \prod_{t=0}^{T-1} \pi_t(u_t|x_t)\, \tau_t(x_{t+1}|\xi_t)$, and the cumulative cost is $\underline{c}_T = \sum_{t=0}^{T-1} c_t(\xi_t) + c_T(x_T)$. The central control problem introduced in [13] jointly optimizes over policies and auxiliary transition kernels, allowing to distinguish between the agent's informational uncertainty regarding action selection and its anticipation of deviations in the system dynamics:

This work was supported by the Research Foundation Flanders (FWO) under SBO grant no. S007723N.

The authors are with Dept. EMSME, Ghent Univ. & Core Lab MIRO, Flanders Make. (e-mail: ajinkya.bhole@ugent.be).

$$\min_{\underline{\pi}} \overset{\lambda^S}{\underset{\underline{\tau}}{\operatorname{opt}}} \mathbb{E}_{p_{(\underline{\pi},\underline{\tau})}}\left[\underline{c}_T + \frac{1}{\lambda^P}\mathbb{D}^{\underline{\pi}}_{\underline{\rho}} + \frac{1}{\lambda^S}\mathbb{D}^{\underline{\tau}}_{\underline{\iota}}\right], \tag{1}$$

where $\underline{\rho}$ is a baseline policy, $\underline{\iota}$ is the nominal dynamics, $\mathbb{D}$ denotes the per-step KL divergence summed over time, and $\lambda^P > 0, \lambda^S \in \mathbb{R} \setminus \{0\}$ independently weight the policy and transition penalties. The operator opt toggles between $\max$ (when $\lambda^S < 0$, leading to risk-averse behavior) and $\min$ (when $\lambda^S > 0$, leading to risk-seeking behavior).

This formulation acts as a *structural template*: by toggling two binary decisions (keeping policy regularization ON or OFF, and keeping auxiliary transition optimization ON or OFF) one navigates between different paradigms. These connections are visualized in Fig. 1.

A. Key Toggling Outcomes

Auxiliary transition optimization OFF and policy regularization OFF recovers the traditional Stochastic Optimal Control formulation

$$\min_{\underline{\pi}} \mathbb{E}_{p_{(\underline{\pi},\underline{\tau})}}\left[\underline{c}_T\right] = J[\underline{\pi}, \underline{\iota}]$$

Auxiliary transition optimization OFF and policy regularization ON yields the Soft-Policy Stochastic Optimal Control (SP-SOC) objective (see Fig. 1), which after rearranging reveals equivalence to a Forward KL (Information projection) matching problem $\min_{\underline{\pi}} \mathbb{D}^{p_{(\underline{\pi},\underline{\iota})}}_{p^*}$, where the target distribution is defined as $p^* \propto p_{(\underline{\rho},\underline{\iota})} e^{-\lambda^P \underline{c}_T}$.

Auxiliary transition optimization ON and policy regularization OFF. By retaining the opt operator over the auxiliary kernel $\underline{\tau}$ in (1), the problem inherently evaluates the underlying cost distribution beyond its mean. The choice of operator toggles the risk attitude, admitting a game-theoretic interpretation where opt $\equiv \max$ models a demonic nature, evaluating worst-case transitions and yielding risk-averse behavior, while opt $\equiv \min$ models an angelic nature, yielding an optimistic, risk-seeking policy. From Risk-Sensitive Control, the transition toggles shift from the penalty space to the constraint space via Lagrangian duality, yielding (pessimistic) Distributionally Robust Control (DRC) [14]:

$$\min_{\underline{\pi}} \max_{\substack{\underline{\tau}\in\underline{\mathcal{U}} \\ \mathcal{T}_t=\{\tau_t \mid \mathbb{D}^{\tau_t}_{\iota_t} \le \eta_t\}}} \mathbb{E}_{p_{(\underline{\pi},\underline{\tau})}}\left[\underline{c}_T + \frac{1}{\lambda^P}\mathbb{D}^{\underline{\pi}}_{\underline{\rho}}\right].$$

The scalar penalty $\frac{1}{\lambda^S}$ on the transition divergence morphs into a hard ambiguity budget ($\mathbb{D}^{\tau_t}_{\iota_t} \le \eta_t$).

Synchronizing the regularization weights, i.e., setting $\lambda^P = \lambda^S = \lambda > 0$, yields the Synchronized Risk-Seeking Soft-Policy RSOC (SRS-SP-RSOC). This amounts to the M-projection (Reverse KL) matching problem $\min_{\underline{\pi}} \mathbb{D}^{p^*}_{p_{(\underline{\pi},\underline{\iota})}}$, where the target distribution is $p^* \propto p_{(\underline{\rho},\underline{\iota})} e^{-\lambda \underline{c}_T}$.

III. Structural Properties of the Central Formulation

The map of connections revealed in [13] exposes several properties that provide a consolidated compass for algorithm design.

A. Majorization–Minimization Iterations

The soft-policy formulations (SP-SOC and SP-RSOC) naturally serve as *majorizers* for their classical counterparts (SOC and RSOC). Since the KL penalty $\frac{1}{\lambda^P}\mathbb{D}^{\underline{\pi}}_{\underline{\pi}^k} \ge 0$ with equality at $\underline{\pi} = \underline{\pi}^k$, the surrogate

$$\underline{\pi}^{k+1} = \arg\min_{\underline{\pi}} \mathbb{E}_{p_{(\underline{\pi},\underline{\iota})}}\left[\underline{c}_T + \frac{1}{\lambda^P}\mathbb{D}^{\underline{\pi}}_{\underline{\pi}^k}\right] \tag{2}$$

guarantees monotone descent on the SOC objective. This means that repeatedly solving the *tractable* regularized problem yields a sequence of policies that converges to the classical, unregularized optimum. An analogous result holds for SP-RSOC majorizing RSOC. In other words, the regularized formulations are not mere relaxations: they provide *tractable surrogates* that can be iteratively solved to recover the original, unregularized optimal policies. This establishes a bridge from the computationally favorable regularized world back to classical optimal control.

B. Deterministic Collapse

The map also reveals the geometric boundaries of different paradigms. When the nominal dynamics $\underline{\iota}$ are purely deterministic, the auxiliary transition optimization loses its expressive freedom because the transition kernel is locked to a deterministic Dirac delta. Consequently, the theoretical boundaries dissolve, yielding SOC $\equiv$ RSOC and SP-SOC $\equiv$ SP-RSOC.

C. Features of SRS-SP-RSOC

In the synchronized case ($\lambda^P = \lambda^S = \lambda > 0$), the SRS-SP-RSOC Bellman recursion simplifies dramatically. Defining the *desirability function* $z_t := e^{-\lambda V_t}$ and setting $r_t := e^{-\lambda c_t}$, the nonlinear Bellman equations transform into a linear recursion [13]:

$$z_t = \mathbb{E}_{\rho_t}[r_t \, \mathbb{E}_{\iota_t}[z_{t+1}]]\,. \tag{3}$$

This linearity enables a *path integral solution*: the value function can be computed as an expectation over forward-sampled trajectories under the baseline policy $\underline{\rho}$ and nominal dynamics $\underline{\iota}$, with future cost $\bar{c}_t = \sum_{k=t}^{T-1} c_t + c_T$:

$$z_t = \mathbb{E}_{p_{(\underline{\rho}_t,\underline{\iota}_t \mid x_t)}}[\exp(-\lambda \bar{c}_t)]\,, \tag{4}$$

bypassing backward dynamic programming entirely. The optimal policy takes the closed-form softmax structure

$$\pi^*_t = \rho_t \, \frac{r_t \, \mathbb{E}_{\iota_t}[z_{t+1}]}{z_t}, \tag{5}$$

which reweights the baseline policy proportionally to the expected future desirability. These properties, previously known only in the restricted path integral control setting [15], are shown in [13] to extend to the broader SRS-SP-RSOC class. Furthermore, this problem also enables *compositionality* for modular control design.

IV. Towards Efficient Exploration in MBRL

We now describe how the central formulation connects to the exploration–exploitation tradeoff in MBRL, outlining the direction of our ongoing and future work.

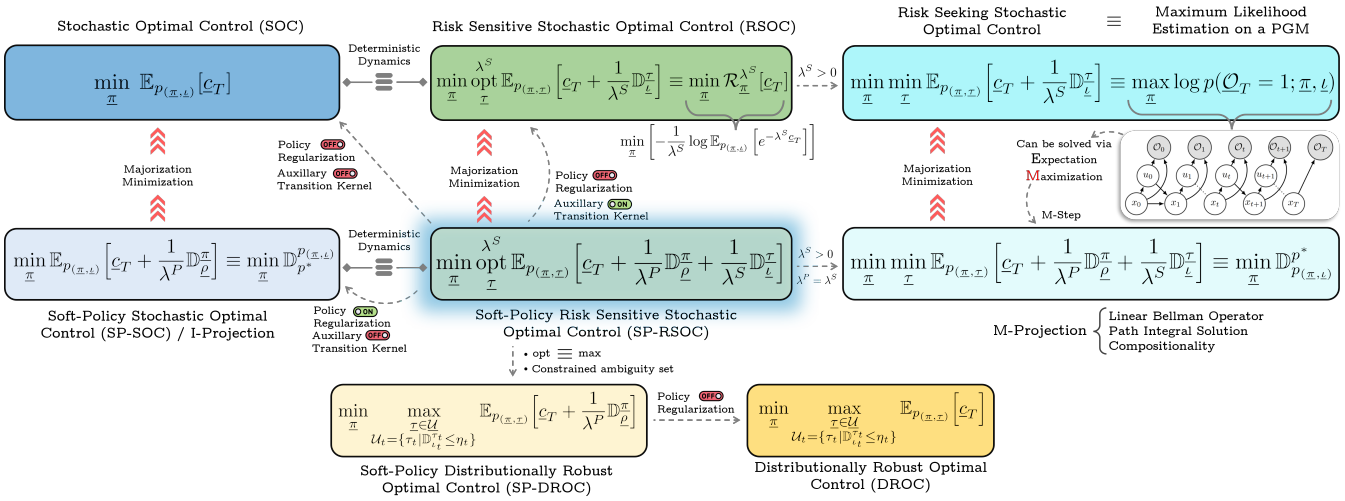

Fig. 1. A visual map of the probabilistic optimal control landscape, illustrating the structural toggles that connect distinct problem formulations.

A. The Optimistic Exploration Problem

Consider an episodic MBRL setting. In each episode n, the agent selects a policy $\underline{\pi}_n$ based on accumulated data $\mathcal{D}_n$ and collects new transition data. Let $\mathcal{T}_n$ be the set of dynamical models consistent with $\mathcal{D}_n$, and define the regret of playing $\underline{\pi}_n$ instead of the true optimal policy $\underline{\pi}^*$ as $R[\underline{\pi}_n] = J[\underline{\pi}_n, \underline{\tau}^*] - J[\underline{\pi}^*, \underline{\tau}^*]$, where again $J[\underline{\pi}, \underline{\tau}] = \mathbb{E}_{p_{(\underline{\pi},\underline{\tau})}}[\underline{c}_T]$. The agent's goal is to minimize the cumulative regret $\mathcal{R}_N = \sum_{n=1}^{N} R[\underline{\pi}_n]$.

The *optimistic exploration strategy* selects the policy that is optimal under the most favorable plausible model:

$$\underline{\pi}_n = \arg\min_{\underline{\pi}} \min_{\underline{\tau}\in\mathcal{T}_n} J[\underline{\pi}, \underline{\tau}]. \tag{6}$$

Under the assumption that the true model $\underline{\tau}^* \in \mathcal{T}_n$ (well-calibrated confidence), this yields a *lower bound* on the true optimal cost: $J[\underline{\pi}_n, \underline{\tau}_n] \leq J[\underline{\pi}^*, \underline{\tau}^*]$. It can then be shown that the per-episode regret is bounded by the model prediction error, ultimately yielding $\mathcal{R}_N = O(\sqrt{N})$.

Unfortunately, this joint optimization is generally intractable. However, one can impose structure on the confidence set $\mathcal{T}_n$ to gain analytical leverage. If $\mathcal{T}_n$ is defined as a KL ball, $\mathcal{T}_n = \{\underline{\tau} : \mathbb{D}^{\underline{\tau}}_{\bar{\underline{\tau}}_n} \leq \Delta^{\tau}_n\}$, the inner optimization boils down to an optimistic DRC problem. While this constrained problem is slightly more tractable, it remains challenging for general nonlinear systems.

To overcome this, one can additionally restrict the policy to a trust region $\Pi_n = \{\underline{\pi} : \mathbb{D}^{\underline{\pi}}_{\underline{\pi}_{n-1}} \leq \Delta^{\pi}_n\}$, yielding the constrained problem:

$$\underline{\pi}_n = \arg\min_{\underline{\pi}\in\Pi_n} \min_{\underline{\tau}\in\mathcal{T}_n} J[\underline{\pi}, \underline{\tau}]. \tag{7}$$

B. Connection to the Central Formulation

By applying Lagrangian duality to both the dynamics and policy constraints, the constrained problem (7) is relaxed into an unconstrained, fully regularized problem:

$$\underline{\pi}_n = \arg\min_{\underline{\pi}} \min_{\underline{\tau}} \mathbb{E}_{p_{(\underline{\pi},\underline{\tau})}}\left[\underline{c}_T + \frac{1}{\lambda}\mathbb{D}^{\underline{\pi}}_{\underline{\pi}_{n-1}} + \frac{1}{\lambda}\mathbb{D}^{\underline{\tau}}_{\bar{\underline{\tau}}_n}\right], \tag{8}$$

where $\lambda > 0$ is the shared Lagrange multiplier dual to the constraint radii.

This is precisely the central formulation (1) applied episodically, with baseline policy $\underline{\rho} = \underline{\pi}_{n-1}$ (previous episode's policy), nominal dynamics $\underline{\iota} = \bar{\underline{\tau}}_n$ (current model estimate), and synchronized weights $\lambda^P = \lambda^S = \lambda > 0$. Crucially, the sign $\lambda > 0$ in the transition penalty corresponds to $\lambda^S > 0$, i.e., the *risk-seeking* setting. This is the mathematical encoding of optimism: the agent seeks the most favorable transition model within the plausible set implicitly defined by the Lagrange multiplier λ.

This structural alignment is paramount: by relaxing the constrained problem, we place (8) directly in the SRS-SP-RSOC regime, avoiding constrained optimization entirely and unlocking the linear Bellman equations (3), path integral solutions (4), and softmax policies (5).

C. Unifying Existing Approaches

Within the episodic regret framework, the existing literature has addressed the intractability of (6) through two broad strategies:

1) **Optimistic exploration** [1], [4]: Select the most favorable plausible model and plan against it, so that uncertain regions are visited as a byproduct of optimism.
2) **Intrinsic reward** [5], [6], [16]: Keep the model fixed and subtract an exploration bonus proportional to model uncertainty from the cost.

The risk-seeking formulation (8) unifies both perspectives. Consider the parametric setting where transition dynamics are characterized by a parameter θ with current belief q (e.g., a Gaussian). Solving the inner optimization over $\underline{\tau}$ yields the risk-seeking cost $J^{\mathrm{RS}}[\underline{\pi}] = -\frac{1}{\lambda}\log\mathbb{E}_q[\exp(-\lambda\, J[\underline{\pi}, \theta])]$, whose second-order cumulant expansion gives:

$$J^{\mathrm{RS}}[\underline{\pi}] \approx \underbrace{\mathbb{E}_q\big[J[\underline{\pi}, \theta]\big]}_{\text{exploitation}} - \underbrace{\frac{\lambda}{2}\mathbb{V}_q\big[J[\underline{\pi}, \theta]\big]}_{\text{exploration bonus}}. \tag{9}$$

Risk-seeking control thus naturally combines exploitation (minimizing the expected cost) with exploration via an *intrinsic reward shaped by the cost*: the variance bonus favors policies whose cost is most sensitive to model uncertainty. When $q = \mathcal{N}(\hat{\theta}_n, \Sigma_n)$ is a concentrated density, $\mathbb{V}_q[J] \approx \nabla_\theta J|_{\hat{\theta}_n}^\top \Sigma_n \nabla_\theta J|_{\hat{\theta}_n}$, recovering the structure of the economic optimal experiment design criterion [9], – where Σ_n is replaced with an estimate based on the states to be visited such as the Fisher Information Matrix – which quantifies the expected loss of optimality from parameter estimation error. This suggests that optimality-gap-based exploration [10], [11] and optimistic MBRL can be connected through the proposed formulation.

V. Open Questions and Future Work

The connection outlined above raises several important open questions that require future investigation:

1) **Regret under regularization.** Does the sublinear regret guarantee $\mathcal{R}_N = O(\sqrt{N})$ survive under regularization? Under regularization, not every policy or model is reachable, and optimism may need to be re-established under a controlled approximation error. Quantifying how much worse it can get is essential.
2) **Scheduling the regularization.** The shared Lagrange multiplier λ (or equivalently, the trust region radii Δ_n^π, Δ_n^τ) controls the exploratory behavior. Designing an optimal schedule λ_n that balances convergence speed with regret minimization is an important open problem.
3) **Beyond KL divergences.** The current formulation relies exclusively on the KL divergence. Extending the map (Fig. 1) to other statistical divergences could yield alternative formulations with different exploration–exploitation characteristics.
4) **Comparison with existing MBRL approaches.** As discussed in Section IV-C, existing literature addresses the intractability of optimistic exploration through approximations such as hallucinated controls or cost modification. A key future direction is to theoretically and empirically compare these methods against our proposed formulation (8).
5) **Safe exploration.** Optimistic exploration steers toward uncertain regions, which may be unsafe. Recent work [17] addresses this by combining worst-case robustness through pessimistic DRC with exploration through maximum diffusion [18]. Since the central formulation (1) accommodates both risk-averse and risk-seeking modes, a natural question is how to systematically balance safety with exploration.

VI. Conclusion

In this paper, we described how structural insights gained from mapping the landscape of probabilistic optimal control led us to a new, tractable perspective on efficient exploration in model-based reinforcement learning. By showing that the traditionally intractable optimistic exploration problem can be relaxed into a fully regularized central control formulation, we unlock powerful analytical properties that avoid complex optimization. While several open questions remain, particularly regarding the formal preservation of sublinear regret guarantees under this regularization, this structural alignment establishes a promising direction for developing scalable and provably efficient exploration algorithms.

References

[1] S. Curi, F. Berkenkamp, and A. Krause, "Efficient model-based reinforcement learning through optimistic policy search and planning," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 14 156–14 170.

[2] L. Treven, J. Hübotter, B. Sukhija, F. Dörfler, and A. Krause, "Efficient exploration in continuous-time model-based reinforcement learning," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[3] T. Jaksch, R. Ortner, and P. Auer, "Near-optimal regret bounds for reinforcement learning," *Journal of Machine Learning Research*, vol. 11, pp. 1563–1600, 2010.

[4] B. O'Donoghue, "Efficient exploration via epistemic-risk-seeking policy optimization," *arXiv preprint arXiv:2302.09339*, 2023.

[5] M. Bartos, B. D. Lee, L. Treven, A. Krause, F. Dörfler, and M. N. Zeilinger, "Optimistic online LQR via intrinsic rewards," *IEEE Control Systems Letters*, 2025.

[6] B. Sukhija, L. Treven, C. Sferrazza, F. Dörfler, P. Abbeel, and A. Krause, "Sombrl: Scalable and optimistic model-based rl," *arXiv preprint arXiv:2511.20066*, 2025.

[7] N. Sobanbabu, G. He, T. He, Y. Yang, and G. Shi, "Sampling-based system identification with active exploration for legged sim2real learning," in *9th Annual Conference on Robot Learning*, 2025.

[8] B. Zhang, Z. Zhou, and R. Vasudevan, "Provably-safe, online system identification," *arXiv preprint arXiv:2504.21486*, 2025.

[9] B. Houska, D. Telen, F. Logist, M. Diehl, and J. F. Van Impe, "An economic objective for the optimal experiment design of nonlinear dynamic processes," *Automatica*, vol. 51, pp. 98–103, 2015.

[10] R. R. Jackson, D. Varagnolo, and S. Knorn, "A dual control approach to solve exploration versus exploitation trade-offs in the design of personalized physical exercise sessions," *IEEE Control Systems Letters*, vol. 7, pp. 2383–2388, 2023.

[11] X. Feng and Y. Jiang, "A scheme for simultaneous optimal tracking control and experiment design," *IEEE Access*, vol. 8, pp. 25 364–25 371, 2020.

[12] J. Schmidhuber, "Formal theory of creativity, fun, and intrinsic motivation (1990–2010)," *IEEE transactions on autonomous mental development*, vol. 2, no. 3, pp. 230–247, 2010.

[13] A. Bhole, M. M. Filabadi, G. Crevecoeur, and T. Lefebvre, "Unifying entropy regularization in optimal control: From and back to classical objectives via iterated soft policies and path integral solutions," 2025. [Online]. Available: https://arxiv.org/abs/2512.06109

[14] A. Nilim and L. El Ghaoui, "Robust control of Markov decision processes with uncertain transition matrices," *Operations Research*, vol. 53, no. 5, pp. 780–798, 2005.

[15] H. J. Kappen, "Linear theory for control of nonlinear stochastic systems," *Physical Review Letters*, vol. 95, no. 20, p. 200201, 2005.

[16] J. Boedecker, J. T. Springenberg, J. Wülfing, and M. Riedmiller, "Approximate real-time optimal control based on sparse gaussian process models," in *2014 IEEE symposium on adaptive dynamic programming and reinforcement learning (ADPRL)*. IEEE, 2014, pp. 1–8.

[17] H. Jesawada, G. Russo, A. Swikir, and F. Abu-Dakka, "Learning-based robust control: Unifying exploration and distributional robustness for reliable robotics via free energy," *arXiv preprint arXiv:2603.06831*, 2026.

[18] T. A. Berrueta, A. Pinosky, and T. D. Murphey, "Maximum diffusion reinforcement learning," *Nature Machine Intelligence*, vol. 6, no. 5, pp. 504–514, 2024.