Stochastic Decision Horizons for Constrained Reinforcement Learning

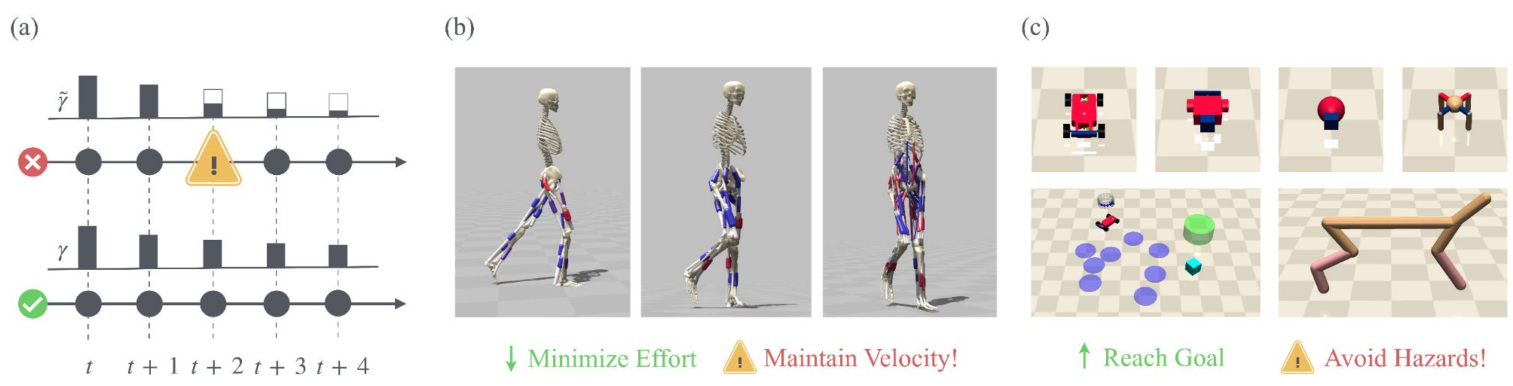
Nikola Milosevic, Leonard Franz, Daniel Häufle, Georg Martius, Nico Scherf*, Pavel Kolev*

MAX PLANCK INSTITUTE FOR HUMAN COGNITIVE AND BRAIN SCIENCES · ScaDS.AI DRESDEN LEIPZIG · EBERHARD KARLS UNIVERSITÄT TÜBINGEN · MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS · Hertie Institute for Clinical Brain Research · imprs-is

Abstract

We extend Control as Inference (CaI) [1,2] to constrained RL using *stochastic decision horizons*, where constraint violations reduce continuation probabilities, attenuating rewards and shortening the effective planning horizon. The resulting survival-weighted objective remains replay-compatible for off-policy learning and yields SAC/MPO-style updates under *absorbing* or *virtual* termination semantics. Experiments show improved sample efficiency and strong return-violation trade-offs, scaling to high-dimensional musculoskeletal control.

Highlights



Problem Formulation

Constrained RL is commonly modeled as a CMDP: an infinite-horizon discounted MDP with per-step violation signals. The objective maximizes expected return subject to bounds on the expected discounted cumulative violation cost, or via a chance constraint limiting the probability of ever violating safety.

$$\max_{\pi} \; \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t\, r(s_t,a_t)\right] \quad \text{s.t.} \quad \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t\, c(s_t,a_t)\right] \le d,$$

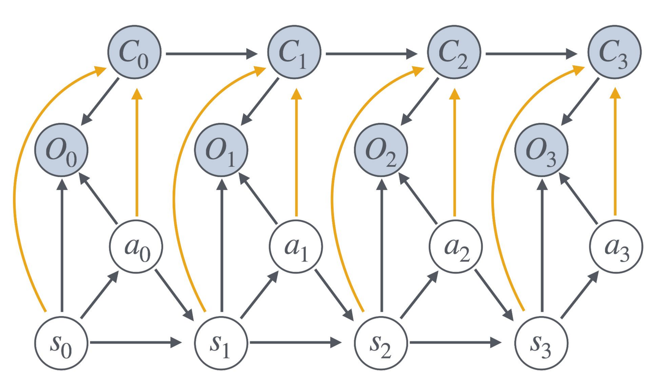
Challenge: Most practical CMDP algorithms rely on *Lagrangian or primal-dual methods*, which (1) are typically *on-policy*, (2) require careful tuning of dual variables, and (3) integrate poorly with modern *off-policy actor-critic methods*. At the same time, *infeasible experience is often informative* and *unavoidable* during exploration.

Our Approach

Instead of enforcing constraints via dual variables or hard feasibility, using CaI we model safety as state-action-dependent survival: violations reduce the continuation probability $\alpha(s,a)$, shortening the effective horizon $\tilde{\gamma}(s,a) := \gamma\,\alpha(s,a)$ and attenuating rewards $\tilde{r}(s,a) := \alpha(s,a)r(s,a)$ This generalizes termination style relaxations such as CaT [5] by treating the continuation model as a flexible mapping from violation signals. The resulting survival-weighted objectives are replay-compatible and induce off-policy schemes with SAC/MPO-style updates.

Probabilistic Graphical Model

$$C_t \sim \text{Bernoulli}(\alpha(s_t,a_t)),$$
$$\text{where} \quad \alpha: \mathcal{S}\times\mathcal{A} \to [0,1]$$



We distinguish two termination semantics: **virtual termination** (VT), where the agent continues acting after a violation, and **absorbing state** (AS), where a violation ends the decision process. Both share the same survival-weighted return, but KL-regularization is discounted differently: standard in VT, survival-weighted in AS, leading to different policy updates.

Main Theorem

$$J_{\text{surv}}(\pi) := \mathbb{E}_{\tau\sim\pi}\Big[\sum_{t\ge 0} u_t\,\tilde{r}(s_t,a_t)\Big], \quad u_t := \prod_{k=0}^{t-1}\tilde{\gamma}(s_k,a_k).$$

VT-ELBO

$$\mathcal{J}_{\text{VT}}(\pi) = J_{\text{surv}}(\pi) - \kappa\,\mathbb{E}_{\tau\sim\pi}\Big[\sum_{t\ge 0}\gamma^t\,\text{KL}_t\Big]$$

AS-ELBO

$$\mathcal{J}_{\text{AS}}(\pi) = J_{\text{surv}}(\pi) - \kappa\,\mathbb{E}_{\tau\sim\pi}\Big[\sum_{t\ge 0}u_t\,\text{KL}_t\Big]$$

Scalable Off-policy Algorithms

VT-MPO — Maximizes the VT-ELBO using policy updates similar to Maximum-a-posteriori Policy Optimization (MPO) [3]

AS-SAC — Maximizes the AS-ELBO using policy updates similar to Soft Actor Critic (SAC) [4]

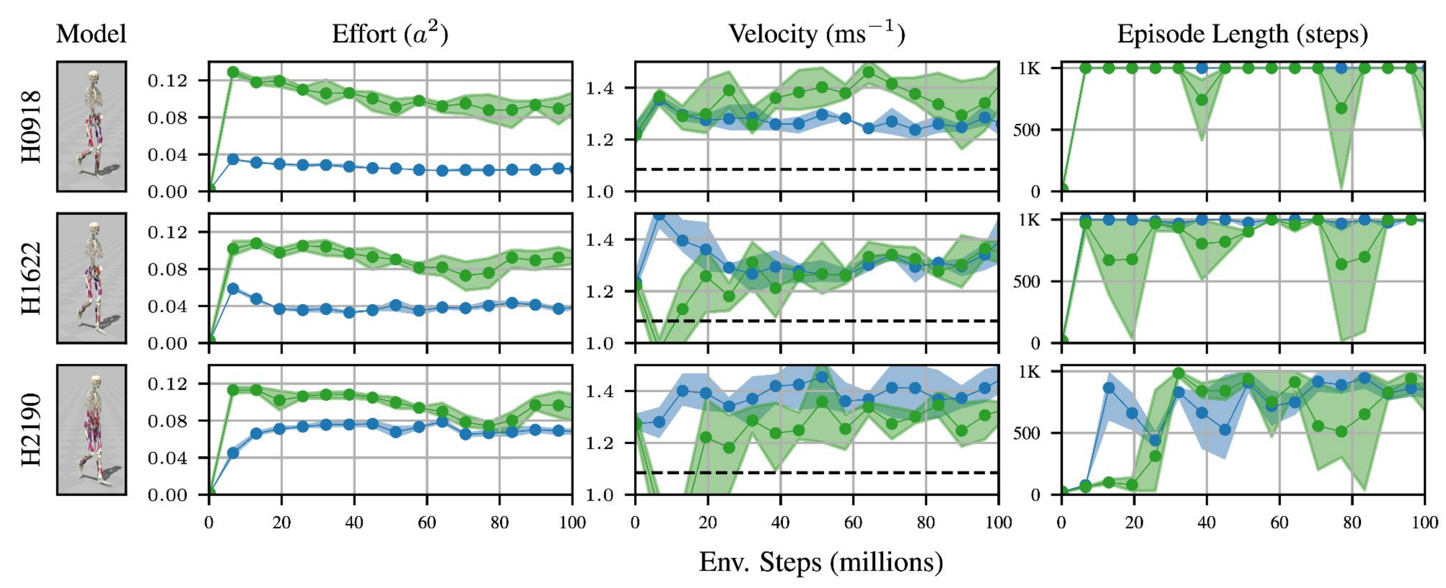
Remark (Critic). AS and VT share the same survival-shaped critic: Bellman backups use $(\tilde{r},\tilde{\gamma})$ and remain a contraction $(\sup_{s,a}\tilde{\gamma}(s,a)\le\gamma)$, enabling stable off-policy replay.

Remark (AS subtlety). Under AS, regularization is survival-weighted, inducing a non-constant "*living cost*". To address this, we propose a two-critic decoupling that enables principled off-policy *temperature* tuning.

Experiments

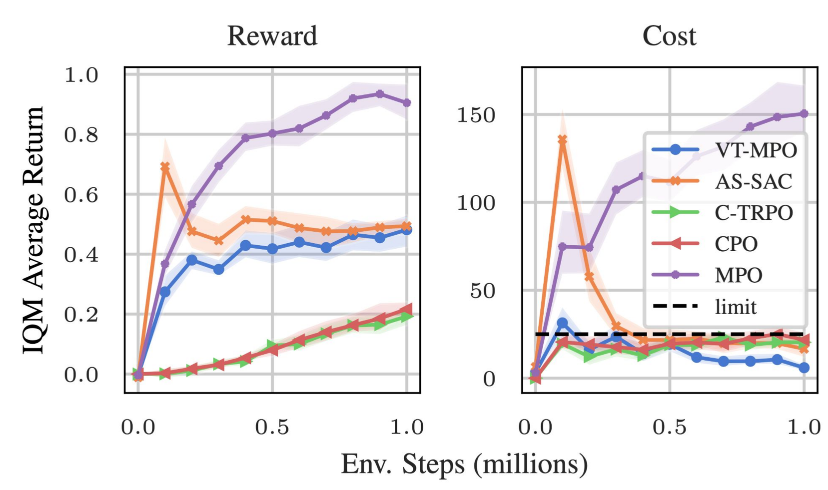
Hyfydy

In Hyfydy musculoskeletal control environments, **VT-MPO** scales in our minimal effort-velocity constrained formulation, using a single continuation schedule shared across tasks. It learns low-effort gaits that meet the target velocity without Lagrange multipliers or adaptive dual tuning, and trains robustly across seeds. Under the same objective and protocol, **VT-MPO** improves the effort-velocity trade-off over **EWA**, a state-of-the-art adaptive effort-weight baseline.



Safety-Gymnasium

On Safety Gymnasium, **VT-MPO** and **AS-SAC** substantially reduce violations relative to *unconstrained* MPO while preserving reward, yielding smooth return-violation trade-offs without Lagrange multipliers. **VT-MPO** is robust across tasks and seeds with stable learning dynamics, whereas **AS-SAC** is more return-seeking, often reaching higher asymptotic reward when costs are driven low, but shows higher variance when violations persist.



Future Directions

1. Unify SDH theory (AS/VT) as a regularized MDP
2. Learn *continuation* to automatically match target violation levels
3. Risk + new regimes (distributional / offline / model-based).

Paper · Website

nmilosevic@cbs.mpg.de
pavel.kolev@uni-tuebingen.de

References

[1] Rawlik, K., Toussaint, M., and Vijayakumar, S. On stochastic optimal control and reinforcement learning by approximate inference. Proceedings of Robotics: Science and Systems VIII, 2012.
[2] Levine, S. Reinforcement learning and control as probabilistic inference: Tutorial and review. CoRR, abs/1805.00909, 2018.
[3] Abdolmaleki, A., Springenberg, J. T., Tassa, Y., Munos, R., Heess, N., and Riedmiller, M. A. Maximum a posteriori policy optimisation. In International Conference on Learning Representations. OpenReview.net, 2018b.
[4] Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In International conference on machine learning, pp. 1861–1870. Pmlr, 2018.
[5] Chane-Sane, E., Leziart, P.-A., Flayols, T., Stasse, O., Soueres, P., and Mansard, N. CaT: Constraints as terminations for legged locomotion reinforcement learning. In 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2024