Carnegie Mellon University

Improving Model-Based Reinforcement Learning by Converging to Flatter Minima

Shrinivas Ramasubramanian, Benjamin Freed, Alexandre Capone, Jeff Schneider

Auton Lab

STACK

NEURAL INFORMATION PROCESSING SYSTEMS 2025

Problem and Key Idea

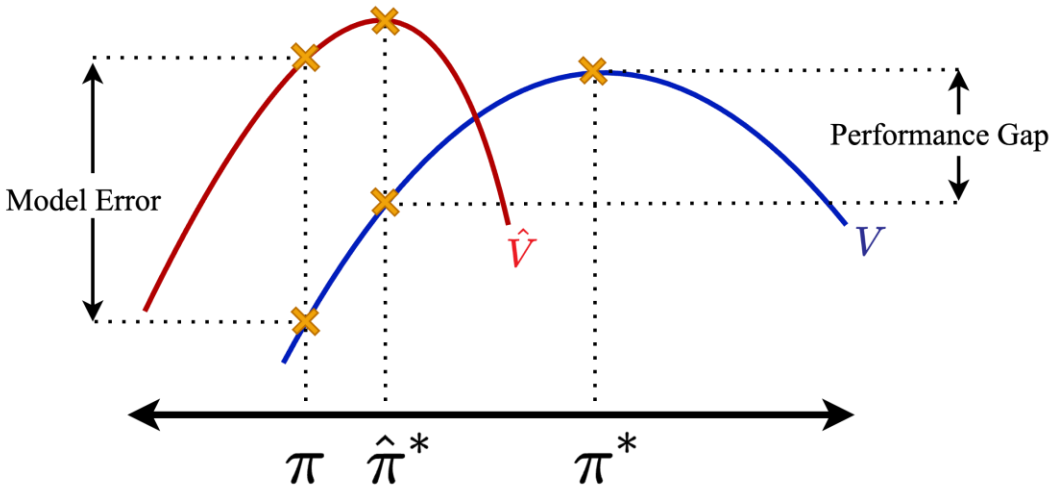
Model-Based Reinforcement Learning works by learning a world-model of env. Dynamics

Policy is learnt using either imagined rollouts or value-based methods

Error in imagined rollouts can lead to compounding errors leading to brittle policy learning

These errors are exacerbated due to policy state-action visitation distribution drift.

Can enforce a regularization with strong PAC-Bayesian guarantees?



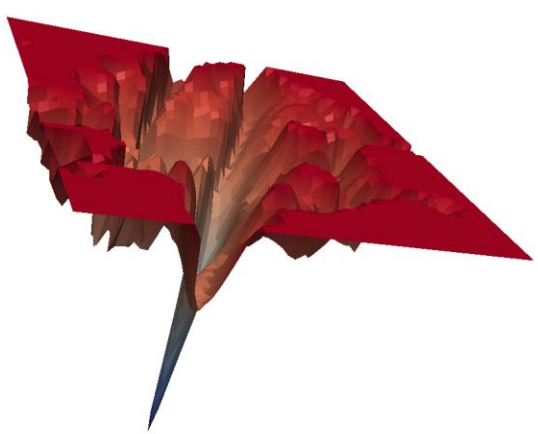
Sharpness-Aware Minimization

Models converging to flat minima exhibit lower test error and robustness to distribution shifts.

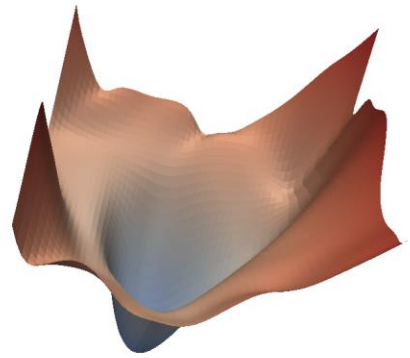
$$\mathcal{L}_{\mathcal{S}}(\theta) = \frac{1}{|\mathcal{S}|} \sum_{(s,a,s') \in \mathcal{S}} \|\hat{s}'_\theta(s,a) - s'\|^2$$

$$\max_{\|\epsilon\|_2 \leq \rho} \mathcal{L}_{\mathcal{S}}(\theta^\star + \epsilon) \approx \mathcal{L}_{\mathcal{S}}(\theta^\star)$$

$$\min_\theta \underbrace{\max_{\|\epsilon\|_2 \leq \rho} \mathcal{L}_{\mathcal{S}}(\theta + \epsilon)}_{=: L_{\mathcal{S}}^{\mathrm{SAM}}(\theta)} + \lambda \|\theta\|_2^2$$



Sharp minima

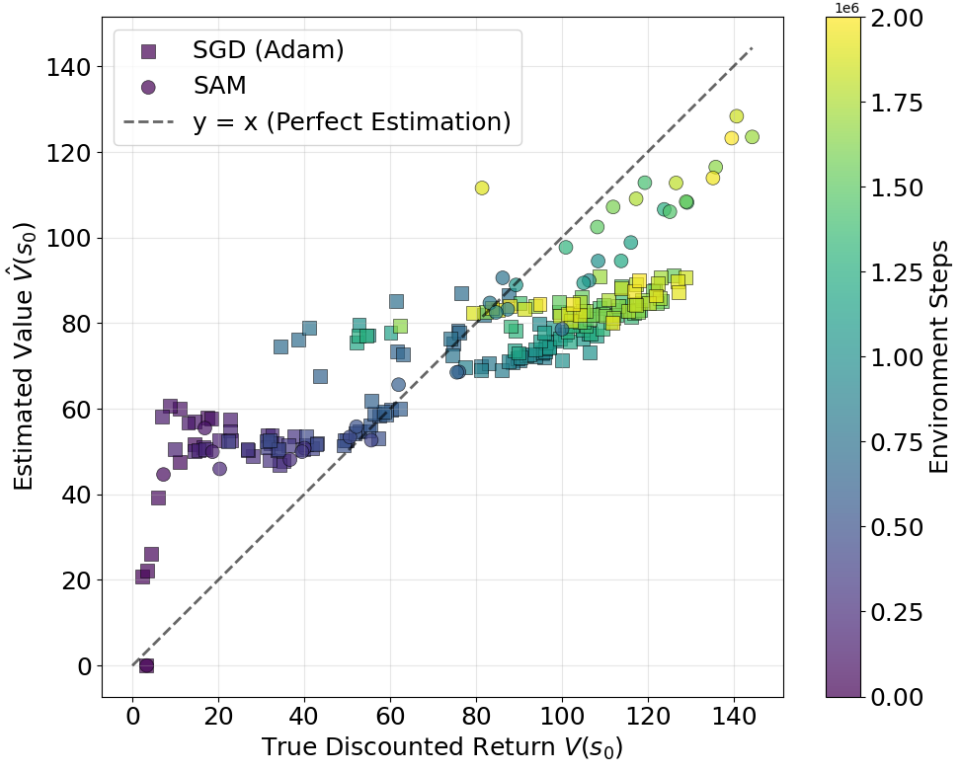


Flat minima

Our Theoretical Contribution

Theorem 1 (Return-estimation gap). *For any policy π and discount $\gamma \in [0, 1)$. Let $\hat{L}(\theta; \pi)$ be the empirical model error under π computed from n i.i.d. samples, and assume the per-sample loss is bounded by M_{loss} and the first-order sharpness of minimum $R_\rho^{(1)}(\theta; \pi)$, for any initial state s_0 we have, with probability at least $1 - \delta$,*

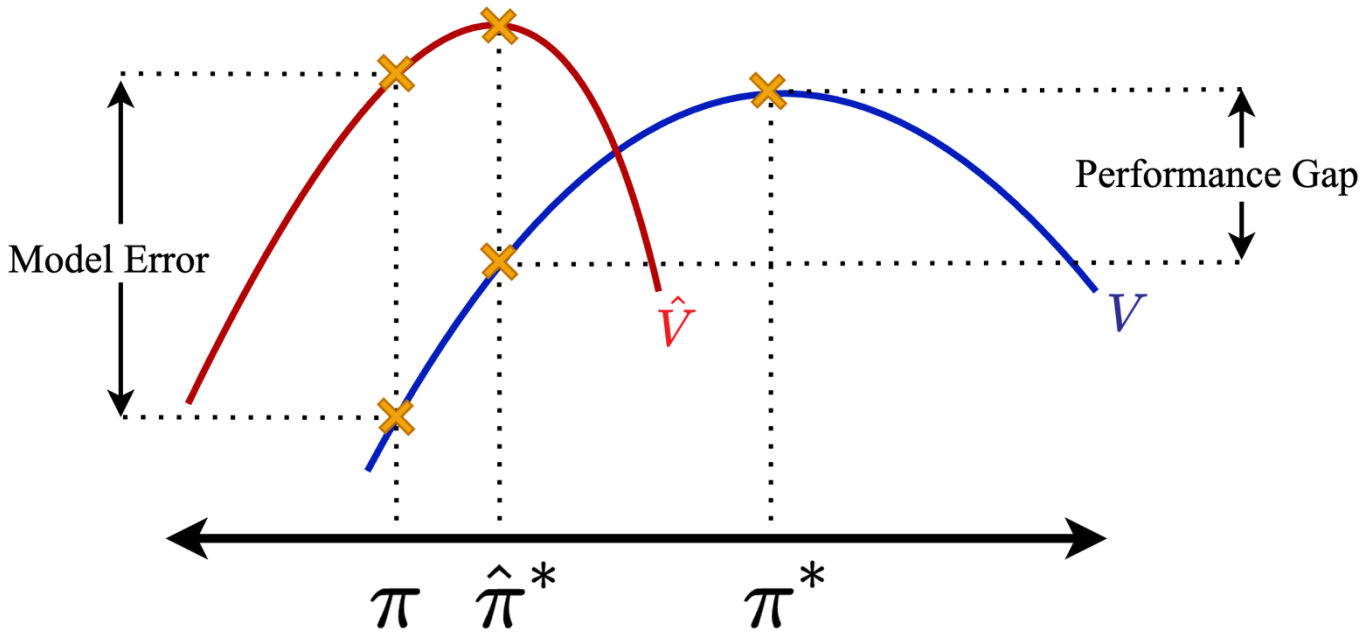
$$\left|V^\pi(s_0) - \hat{V}^\pi(s_0)\right| \leq \frac{\gamma}{(1-\gamma)^2}\left[\hat{L}(\theta;\pi) + R_\rho^{(1)}(\theta;\pi) + \sqrt{M_{\text{loss}}/n} + \Omega(d, n, \rho, \delta)\right].$$



Our Theoretical Contribution

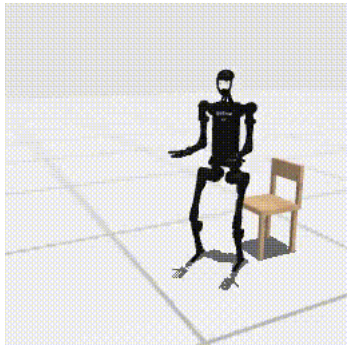
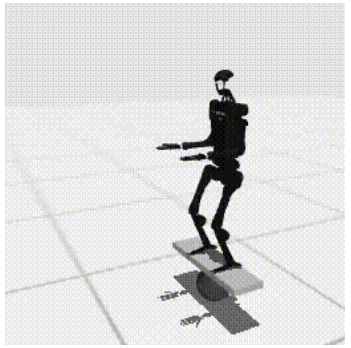
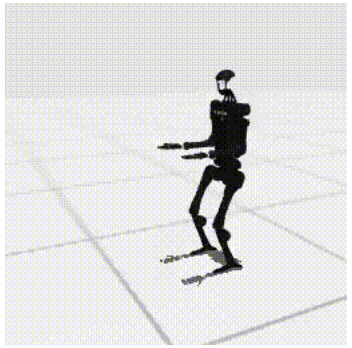
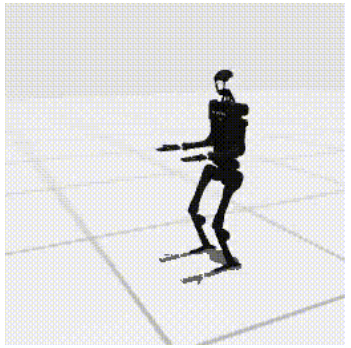
Theorem 2 (Performance Gap). *Let π^* be the optimal policy in the true environment and $\hat{\pi}^*$ be the policy that is optimal according to the learned model $\hat{M}_\theta$. If $C = \max_{(s,a)} \frac{d^{\pi^*}(s,a)}{d^{\hat{\pi}^*}(s,a)}$ is the concentrability coefficient measuring the distribution mismatch between π^* and $\hat{\pi}^*$:*

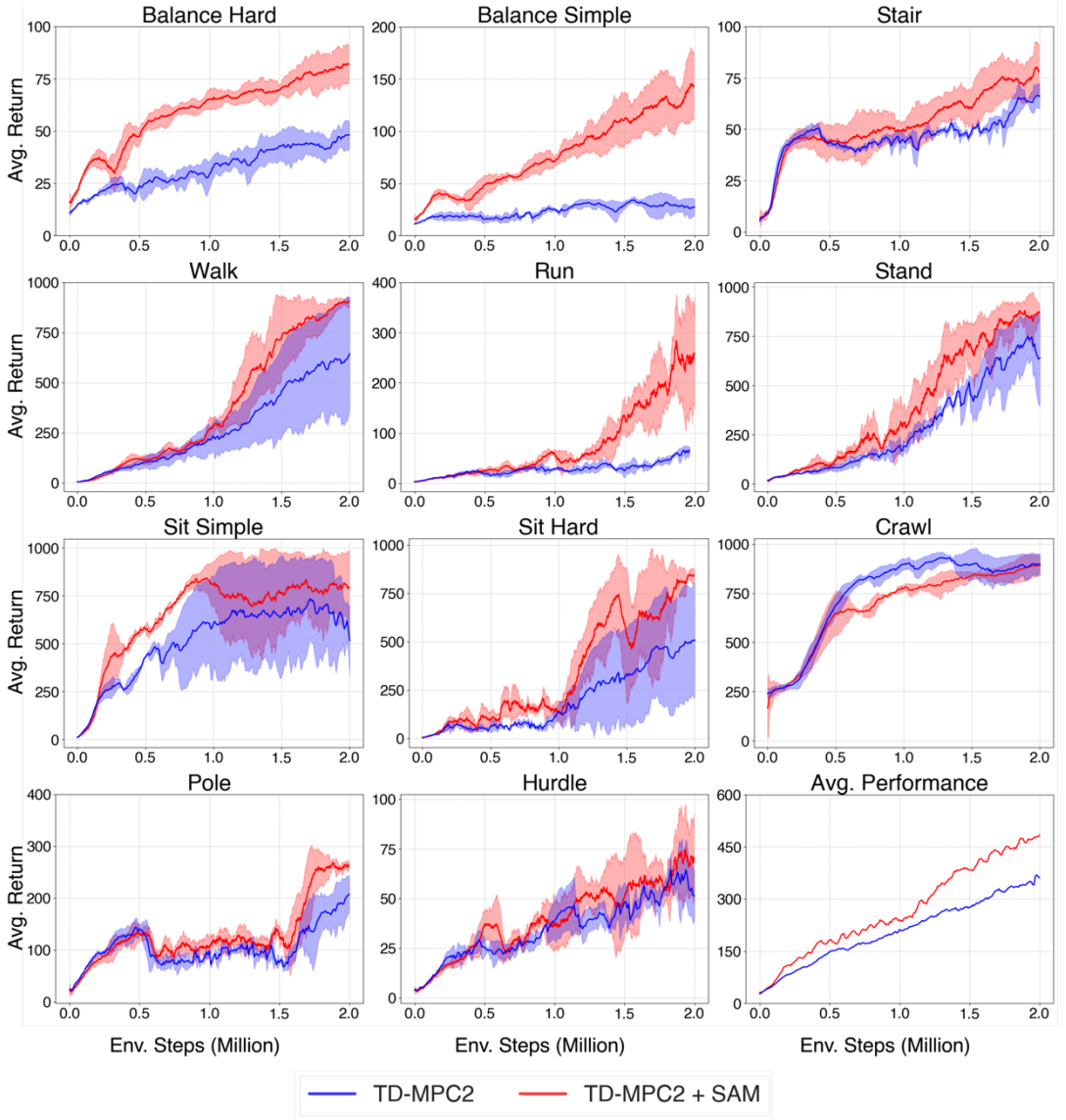
$$\left|V^{\pi^*}(s_0) - V^{\hat{\pi}^*}(s_0)\right| \leq \frac{\gamma(1+C)}{(1-\gamma)^2}\left[\hat{L}(\theta;\hat{\pi}^*) + R_\rho^{(1)}(\theta;\hat{\pi}^*) + \sqrt{\frac{M_{loss}}{n}} + \Omega(d,n,\rho,\delta)\right].$$



Empirical Results

We evaluate TD-MPC2 w/ SAM on HumanoidBench,
Which is a high dof env. (21) degree continuous action space



Empirical Results

On Atari-100k with the TWISTER world-model, adding SAM boosts mean human-normalized score by **27.6%**, with standout gains on **Frostbite (+315%)**, **Gopher (+97%)**, **BattleZone (+88%)**, and **Road Runner (+58%)**.

SAM helps most when applied to the **dynamics loss** (≫ policy) and is **harmful on reward/value heads** because it can corrupt targets.

Limitations: Results are **simulation-only**, use **4–5 seeds**, are **ρ-sensitive**, and add ~**1.7×** wall-clock. **Bottom line:** a one-line optimizer swap on the **world-model** measurably flattens the loss landscape and improves control across modalities **without** touching the planner or policy.

Table 1: Performance comparison across 20 Atari games. TWISTER and its SAM-augmented variant are benchmarked against prior model-based methods (TWIM, IRIS, Dreamer v3) and baseline agents (Random, Human, Simple). TWISTER w/ SAM consistently improves performance in challenging environments, demonstrating the effectiveness of sharpness-aware optimization in MBRL

Game	Random	Human	Simple	TWIM	IRIS	Dreamer v3	TWISTER	TWISTER w/ SAM
Alien	228	7128	617	675	420	959	823	947
Amidar	6	1720	74	122	143	139	172	172
Assault	222	742	527	683	1524	706	777	1102
Asterix	210	8503	1128	1116	854	932	1132	1030
Bank Heist	14	753	34	467	53	649	673	886
Battle Zone	2360	37188	4031	5068	13074	12250	5452	10230
Boxing	0	12	8	78	70	78	80	86
Demon Attack	152	1971	208	350	2034	303	286	293
Freeway	0	30	17	24	31	0	0	0
Frostbite	65	4335	237	1476	259	909	388	1610
Gopher	258	2412	597	1675	2236	3730	2078	4099
Hero	1027	30826	2657	7254	7037	11161	9836	12320
James Bond	29	303	100	362	463	445	354	426
Kangaroo	52	3035	51	1240	838	4098	1349	1555
Ms Pacman	307	6952	1480	1588	999	1327	2319	2409
Pong	-21	15	13	19	15	18	20	20
Road Runner	12	7845	5641	9107	9615	15565	9811	15532
Seaquest	68	42055	683	774	661	618	434	426
Up N Down	533	11693	3350	15982	3546	7600	4761	6857

Table 7: Where to apply SAM in TD-MPC2 (episode return; mean ± SEM).

Variant	Apply SAM to	Episode return
TD-MPC2 (baseline)	—	301 ± 11
TD-MPC2 + SAM (ours)	**Dynamics**	$\mathbf{484 \pm 19}$
TD-MPC2 + SAM	Reward & Value	10 ± 3
TD-MPC2 + SAM	Policy	463 ± 49