

FAST MIXING STEADY-STATE CONTROL IN MARKOV DECISION PROCESSES

FEDERICO CORSO, MARCO MUSSI, AND ALBERTO MARIA METELLI
{federico.corso, marco.mussi, albertomaria.metelli}@polimi.it



MOTIVATION

RL Objective: Optimality of learned policy w.r.t. a reward signal

Limitation: Insufficient in safety-critical domains. Requirements beyond optimality difficult to encode

Control Theory: Steady-State Control/Stability, a fundamental "safety" requirement

Research Question: How to find the "best" policy inducing a specific Steady-State in an MDP?

GOAL – FAST MIXING STEADY STATE

Synthesize a policy converging at the *fastest possible rate* to a target steady-state distribution

SETTING

Infinite-horizon, reward-free, finite MDP:

$$\mathcal{M} := \langle \mathcal{S}, \mathcal{A}, \mathbf{P}, \mu_0 \rangle$$

Ergodicity: for any $\pi \in \Pi^{\text{MD}}$, the induced chain admits a steady-state distribution η^π : $\lim_{t \rightarrow \infty} \mu_t^\pi = \eta^\pi$

MDP STEADY STATE CONVERGENCE

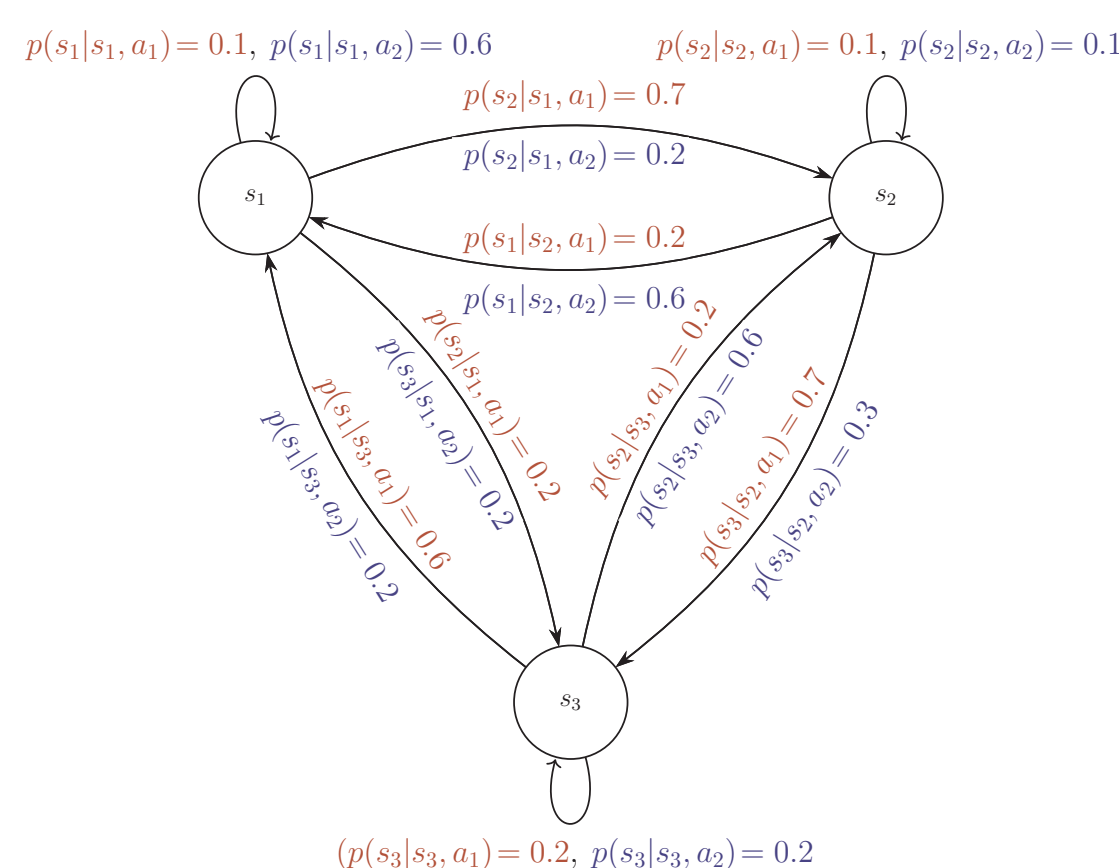
$$\mathcal{M} + \pi \Rightarrow \mathcal{M}^\pi := \langle \mathcal{S}, \mathbf{P}^\pi, \mu_0 \rangle$$

Distance from Steady-State is governed by *Second Largest Eigenvalue Modulus* (SLEM):

$$\lim_{t \rightarrow \infty} \left(\max_{s \in \mathcal{S}} \left\| \mathbf{P}_t^\pi(\cdot|s) - \eta^\pi \right\|_{\text{TV}} \right)^{1/t} = \nu(\mathbf{P}^\pi)$$

$d(t; \mathbf{P}^\pi)$

NON-NORMALITY AND SECOND LARGEST SINGULAR VALUE (SLSV) $\sigma_2(\mathbf{P}^\pi)$



PROBLEM FORMULATION

(1) Projection Problem $\text{PROJ}(\tilde{\eta}, \mathbf{P})$

$$\tilde{\mathbf{x}} \in \arg \min_{\mathbf{x} \in \mathcal{X}_{\mathbf{P}}} \mathcal{D}(\tilde{\eta} \| \mathbf{K}\mathbf{x})$$

$\mathcal{X}_{\mathbf{P}} := \{ \mathbf{x} \in \Delta(\mathcal{S} \times \mathcal{A}) \mid \mathbf{K}\mathbf{x} = \mathbf{P}^\top \mathbf{x} \}$
Feasible state-action stationary distributions

$\mathcal{D}(\cdot \| \cdot)$
convex w.r.t. second argument, defined also for $\tilde{\eta} \geq 0$

$\tilde{\eta} = \mathbf{K}\tilde{\mathbf{x}}$
Feasible target distribution on $\mathcal{S}$

(2) SLSV Minimization $\text{FMSS}(\tilde{\eta}, \mathbf{P})$

$$\min_{\pi \in \Pi^{\text{MR}}} \left\| \tilde{\mathbf{A}}_{\tilde{\eta}}^\pi \right\|_2 = \sigma_2(\mathbf{P}^\pi)$$

s.t. $\tilde{\eta} = \mathbf{P}^\top \tilde{\eta}$ (Stationarity)

$\mathbf{D}_{\tilde{\eta}}^{1/2} \mathbf{P}^\pi \mathbf{D}_{\tilde{\eta}}^{-1/2} - \sqrt{\tilde{\eta}} \sqrt{\tilde{\eta}}^\top$

(3) Feasible Policy Space

For a fixed, realizable $\tilde{\eta}$ there can be an infinite number of $\pi \in \Pi^{\text{MR}}$ realizing it

SAMPLE-BASED FMSS PROBLEM

Algorithm: FMSS through Singular Values

Input: Target $\tilde{\eta} \in \Delta(\mathcal{S})$,
Trajectory $\tau_b = (S_t, A_t, S_{t+1})_{t=0}^{T-1}$,
Parameter $v \in [0, \varepsilon/2\sqrt{|\mathcal{S}||\mathcal{A}|}]$.

Output: Policy $\hat{\pi}^*$ inducing $\eta^{\hat{\pi}^*} \simeq \tilde{\eta}$.

for $(s, a) \in \mathcal{S} \times \mathcal{A}$ **do**
 for $s' \in \mathcal{S}$ **do**
 Estimate transition $\check{p}(s' | s, a)$ from samples
 Ensure ergodicity:
 $\hat{p}(s' | s, a) \leftarrow (1 - v)\check{p}(s' | s, a) + v/|\mathcal{S}|$
 end
end

end
Solve for feasible stationary distributions:

$\hat{\tilde{\mathbf{x}}} \in \arg \min_{\mathbf{x} \in \mathcal{X}_{\hat{\mathbf{P}}}} \mathcal{D}(\tilde{\eta} \| \mathbf{K}\mathbf{x})$
 $\hat{\tilde{\eta}} \leftarrow \mathbf{K}\hat{\tilde{\mathbf{x}}}$
return $\hat{\pi}^* \in \arg \min_{\pi \in \Pi^{\text{MR}}} \|\hat{\tilde{\mathbf{A}}}_{\hat{\tilde{\eta}}}^\pi\|_2$ s.t. $\hat{\tilde{\eta}} = (\hat{\mathbf{P}}^\pi)^\top \hat{\tilde{\eta}}$

(1) Sample Complexity

With probability $1 - \delta$, FMSS-SV achieves $\|\hat{\mathbf{P}} - \mathbf{P}\|_2 \leq 2\varepsilon$ with trajectory length

$$T \leq \tilde{O} \left(\frac{(1 + t_{\text{mix}}^b) |\mathcal{S}| |\mathcal{A}| \log(1/\delta)}{\eta_{\min}^b \xi \varepsilon^2} (\log(1/\delta) + |\mathcal{S}|) \right)$$

$$t_{\text{mix}}^b = \min \{ t \geq 0 : d(t; \mathbf{P}^\pi b) \leq 1/8 \} \quad \min_{s \in \mathcal{S}} \eta^{\pi^b}(s) \quad \pi^b(a|s) \geq \xi > 0 \forall (s, a)$$

(2) Error Propagation

Let $\Delta_P := \|\hat{\mathbf{P}} - \mathbf{P}\|_2$ and let L, C_σ be problem-dependent constants.

Projection Error

$$\left| \mathcal{D}(\tilde{\eta} \| \tilde{\eta}) - \mathcal{D}(\tilde{\eta} \| \hat{\tilde{\eta}}) \right| \leq L\sqrt{\mathcal{S}} \sup_{\pi \in \Pi^{\text{MR}}} \eta_{\min}^{\pi, -1/2} / (1 - \sigma_2(\mathbf{P}^\pi)) \Delta_P$$

Steady-state Error

$$\left\| \hat{\tilde{\eta}} - \eta^{\hat{\pi}^*} \right\|_2 \leq \eta_{\min}^{\hat{\pi}^*, -1/2} / (1 - \sigma_2(\mathbf{P}^{\hat{\pi}^*})) \Delta_P$$

SLSV Error

$$\left| \sigma_2(\mathbf{P}^{\pi^*}) - \sigma_2(\mathbf{P}^{\hat{\pi}^*}) \right| \leq C_\sigma \left(\Delta_P + \left\| \hat{\tilde{\eta}} - \tilde{\eta} \right\|_2^{1/2} \right)$$

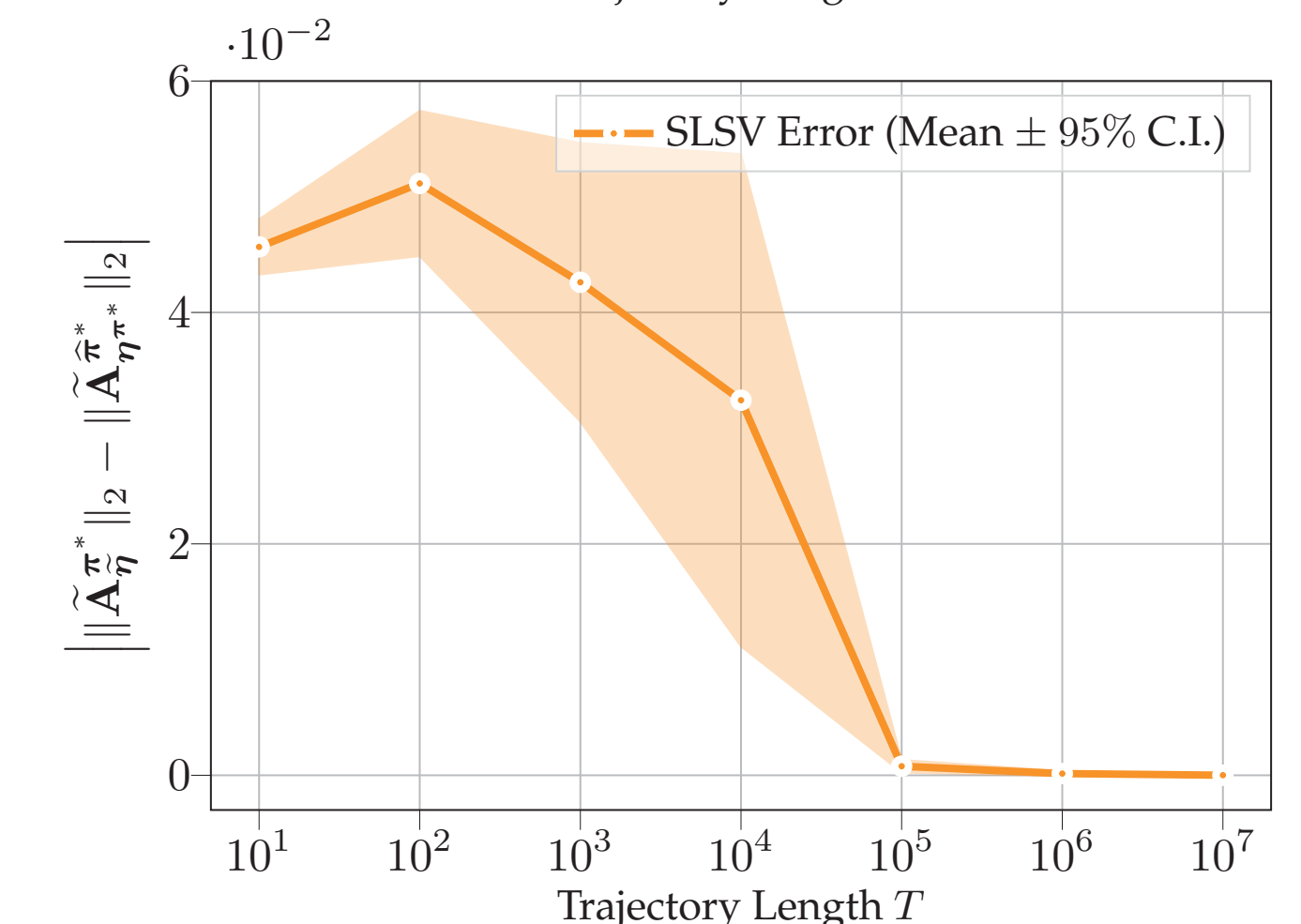
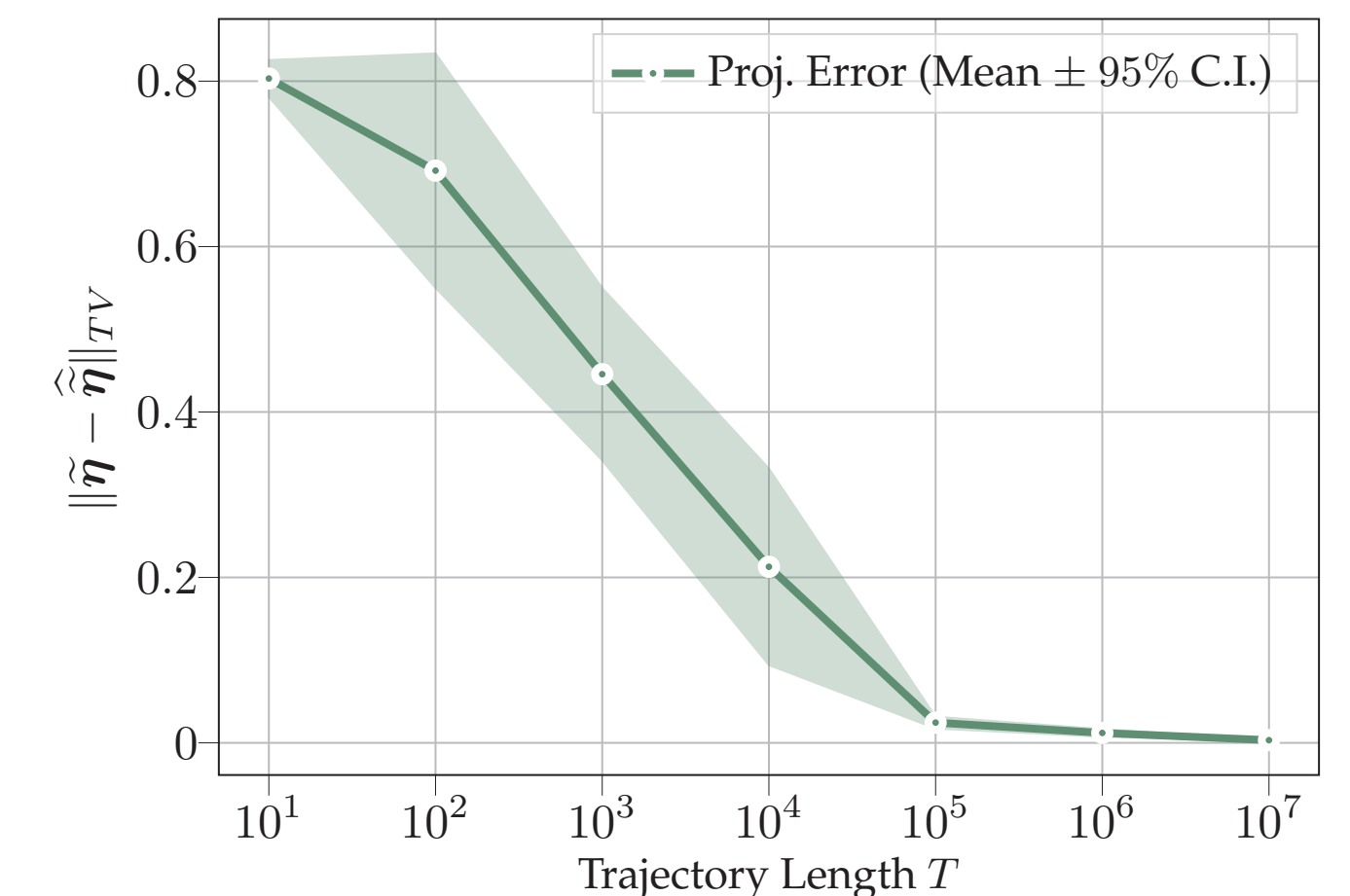
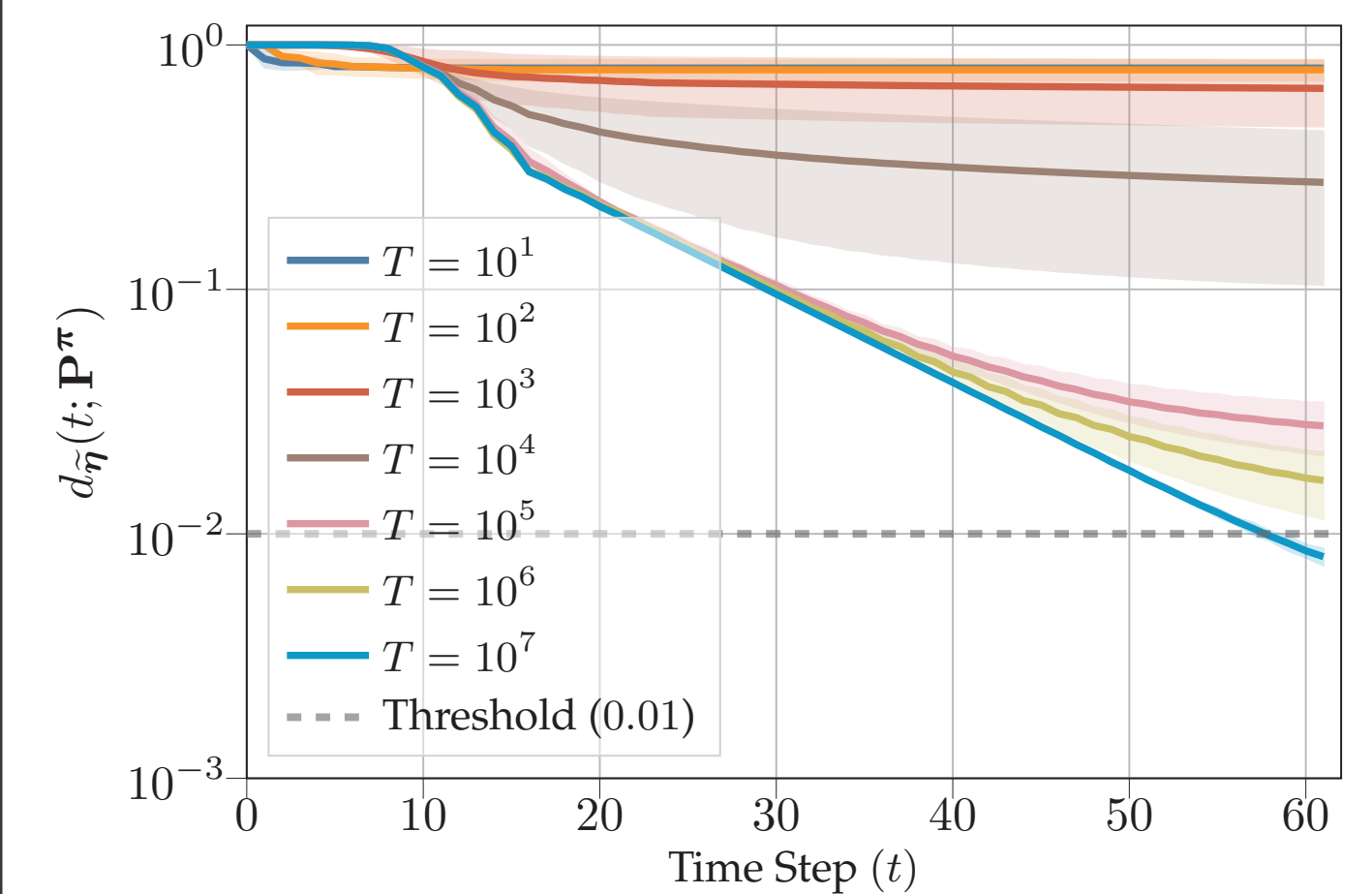
SLEM Error

$$\left| \nu(\mathbf{P}^{\pi^*}) - \nu(\mathbf{P}^{\hat{\pi}^*}) \right| \leq |\sigma_2(\mathbf{P}^{\pi^*}) - \sigma_2(\mathbf{P}^{\hat{\pi}^*})| + \max\{\|\mathbf{N}^{\pi^*}\|_2, \|\mathbf{N}^{\hat{\pi}^*}\|_2\}$$

REFERENCES

- S. Akshay, N. Bertrand, S. Haddad, and L. Helouet. The steady-state control problem for Markov decision processes. In *International Conference on Quantitative Evaluation of Systems*, 2013.
- S. P. Boyd, P. Diaconis, and L. Xiao. Fastest mixing markov chain on a graph. *SIAM Review*, 2004.

8x8 GRID-WORLD EXPERIMENT



LIMITATIONS

- $\sigma_2(\mathbf{P}^\pi)$ is exact for normal Markov chains, but only a surrogate for the SLEM $\nu(\mathbf{P}^\pi)$ in general
- Solving FMSS in general MDPs may require explicitly inducing non-normal dynamics

Future Works

- Extend FMSS to continuous spaces and study the trade-off between asymptotic convergence and finite-time mixing
- Connect distributional stability with other stability notions