Trading-off Reward Maximization and Stability in Sequential Decision Making

Federico Corso, Marco Mussi, and Alberto Maria Metelli
{federico.corso, marco.mussi, albertomaria.metelli}@polimi.it

Motivation

RL objective: maximize expected *reward*.

Limitation: Reward maximization alone is often *insufficient* in real-world, safety-critical domains. Properties beyond pure *performance* are difficult to encode in the reward.

Our idea: *Enforce stability within RL*, inspired by control theory, where stability is fundamental for systems' robustness, safety, and reliability.

Objective

Find a control policy that *trades off* between maximizing reward and ensuring a stable behavior.

Setting: Undiscounted MDPs

We consider infinite-horizon undiscounted MDPs:

$$\mathcal{M} \coloneqq \langle \mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{r}, \boldsymbol{\eta}_0 \rangle$$

Assumptions:

- *Finite* state and action spaces.
- *Ergodicity:* for any $\boldsymbol{\pi} \in \Pi$, the induced chain admits a unique stationary distribution $\boldsymbol{\eta}^{\boldsymbol{\pi}}$:

$$\lim_{t \to \infty} \boldsymbol{\eta}_t^{\boldsymbol{\pi}} = \boldsymbol{\eta}^{\boldsymbol{\pi}}$$

A Measure of Stability

$$\mathcal{M} + \boldsymbol{\pi} \quad \Rightarrow \quad \mathcal{M}^{\boldsymbol{\pi}} \coloneqq \langle \mathcal{S}, \mathbf{P}^{\boldsymbol{\pi}}, \boldsymbol{\eta}_0 \rangle$$

A policy $\boldsymbol{\pi}$ is *more stable* if the Markov Chain (MC) $\mathcal{M}^{\boldsymbol{\pi}}$ induced converges faster to its stationary distribution $\boldsymbol{\eta}^{\boldsymbol{\pi}}$.

The asymptotic convergence rate is governed by the *Second Largest Eigenvalue Modulus (SLEM)*:

$$\lim_{t \to \infty} \left(\max_{s \in \mathcal{S}} \left\| \mathbf{p}_t^{\boldsymbol{\pi}}(\cdot | s) - \boldsymbol{\eta}^{\boldsymbol{\pi}} \right\|_{\mathrm{TV}} \right)^{1/t} = \mu(\mathbf{P}^{\boldsymbol{\pi}})$$

with $\mu(\mathbf{P}^{\boldsymbol{\pi}}) \coloneqq \max_{i \geq 2} |\lambda_i|$.

Single objective: *reward* and *stability*

$$\max_{\pi \in \Pi} \; (\boldsymbol{\eta}^{\boldsymbol{\pi}})^\top \mathbf{r} \; - \; \mu(\mathbf{P}^{\boldsymbol{\pi}})$$

An Explicit Form of the Objective: Bilinearity Issues

$$\underset{\boldsymbol{\Pi}, \boldsymbol{\eta}}{\text{maximize}} \quad \underbrace{\boldsymbol{\eta}^\top (\boldsymbol{\Pi} \mathbf{r})}_{\text{Bilinear in } (\boldsymbol{\eta}, \boldsymbol{\Pi})} - \underbrace{\left\| \mathbf{D}_{\boldsymbol{\eta}}^{1/2} (\boldsymbol{\Pi} \mathbf{P}) \mathbf{D}_{\boldsymbol{\eta}}^{-1/2} - \sqrt{\boldsymbol{\eta}} \sqrt{\boldsymbol{\eta}}^\top \right\|_2}_{\mu(\mathbf{P}^{\boldsymbol{\pi}}),\ \text{Non-convex}}$$

$$\text{subject to} \quad \boldsymbol{\eta} = \underbrace{(\boldsymbol{\Pi} \mathbf{P})^\top \boldsymbol{\eta}}_{\text{Bilinear in } (\boldsymbol{\eta}, \boldsymbol{\Pi})} \rightarrow \text{Stationary constraint}$$

The problem is **not jointly-convex** in $(\boldsymbol{\eta}, \boldsymbol{\Pi})$ and the stationary constraints implies fixing $\boldsymbol{\eta}$ once $\boldsymbol{\Pi}$ is fixed

A Surrogate Objective in $\mathcal{S} \times \mathcal{A}$: Bypassing Bilinearity

(1) Introduce a New MC over $\mathcal{S} \times \mathcal{A}$

$$\mathcal{T} \coloneqq \langle \mathcal{S} \times \mathcal{A}, \mathbf{T}, \mathbf{x}_0 \rangle$$

where:

- $\mathbf{T} = \mathbf{P}\boldsymbol{\Pi}$ (state–action transition matrix)
- $\mathbf{x} = \boldsymbol{\eta}^\top \boldsymbol{\Pi}$ (state–action *stationary* distribution)

Eigenvalue relationship between $\mathbf{P}\boldsymbol{\Pi}$ and $\boldsymbol{\Pi}\mathbf{P}$

$$\Lambda(\mathbf{P}\boldsymbol{\Pi}) \; = \; \Lambda(\boldsymbol{\Pi}\mathbf{P}) \; \cup \; \{0\}^{|\mathcal{S}||\mathcal{A}| - |\mathcal{S}|}$$

(2) Rewrite the Objective in $\mathcal{S} \times \mathcal{A}$

$$\begin{array}{ccc}
\boldsymbol{\eta}^\top (\boldsymbol{\Pi}\, \mathbf{r}) - \overbrace{\left\| \mathbf{D}_{\boldsymbol{\eta}}^{1/2} (\boldsymbol{\Pi}\mathbf{P}) \mathbf{D}_{\boldsymbol{\eta}}^{-1/2} - \sqrt{\boldsymbol{\eta}} \sqrt{\boldsymbol{\eta}}^\top \right\|_2}^{\mu(\boldsymbol{\Pi}\mathbf{P})} \\
\| \qquad\qquad /\!\!\backslash \\
\mathbf{x}^\top \mathbf{r} - \underbrace{\left\| \mathbf{D}_{\mathbf{x}}^{1/2} \mathbf{P}\boldsymbol{\Pi} \mathbf{D}_{\mathbf{x}}^{-1/2} - \sqrt{\mathbf{x}} \sqrt{\mathbf{x}}^\top \right\|_2}_{\sigma_2(\mathbf{P}\boldsymbol{\Pi})}
\end{array}$$

Lifting removes the $\boldsymbol{\eta}^\top \boldsymbol{\Pi}$ bilinearity in the reward term

(3) Bilinearity Resolution

- Introduce the auxiliary variable $\mathbf{X} = \mathbf{D}_{\mathbf{x}} \mathbf{T}$
- Rewrite the problem in $(\mathbf{x}, \mathbf{X})$

$$\begin{array}{ll}
\underset{\mathbf{X}, \mathbf{x}}{\text{maximize}} & \underbrace{(\mathbf{X}\mathbf{1})^\top \mathbf{r}}_{\text{Linear in } (\mathbf{x}, \mathbf{X})} - \left\| \mathbf{D}_{\mathbf{x}}^{-1/2} \mathbf{X} \mathbf{D}_{\mathbf{x}}^{-1/2} - \sqrt{\mathbf{x}} \sqrt{\mathbf{x}}^\top \right\|_2 \\
\text{subject to} & \mathbf{X}\mathbf{1} = \mathbf{x} \rightarrow \text{stationarity; linear but no } \textit{d.o.f.}, \\
& \mathbf{X} = f(\mathbf{P}) \rightarrow \text{compatible with transitions of } \mathcal{M}
\end{array}$$

(4) The Final Problem

$$\begin{array}{ll}
\underset{\mathbf{X} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}}{\text{maximize}} & (\mathbf{X}\mathbf{1})^\top \mathbf{r} \; - \; \left\| \mathbf{D}_{\mathbf{x}^\star}^{-1/2} \mathbf{X} \mathbf{D}_{\mathbf{x}^\star}^{-1/2} - \sqrt{\mathbf{x}^\star} \sqrt{\mathbf{x}^\star}^\top \right\|_2 \\
\text{subject to} & \mathbf{1}^\top \mathbf{X} \mathbf{1} = 1, \\
& \|\mathbf{X}\mathbf{1} - \mathbf{x}^\star\|_2^2 \leq \delta^2 \rightarrow \textit{Relaxed} \text{ stationary constraints}
\end{array}$$

It is **convex** in $\mathbf{X}$

$\mathbf{x}^\star$: solution of the standard undiscounted MDP

Recovering the Policy

Given:

- $\mathbf{X}^\dagger$: Solution to the final problem.
- $\mathbf{x}^\dagger = \mathbf{X}^\dagger \mathbf{1}$: Stationary distribution associated to the MC with transition $\mathbf{T}^\dagger = \mathbf{D}_{\mathbf{x}^\dagger}^{-1} \mathbf{X}^\dagger$.

We can compute the optimal policy:

$$\pi^\dagger(a \mid s) \coloneqq \frac{x^\dagger(s,a)}{\sum_{a \in \mathcal{A}} x^\dagger(s,a)}$$

Let $\boldsymbol{\pi}^\dagger \in \Pi$ be a solution, then it holds:

$$\left(\mathbf{x}^\star - \mathbf{x}^\dagger\right)^\top \mathbf{r} \leq \delta R_{\max} \sqrt{|\mathcal{S}||\mathcal{A}|},$$

and:

$$\sigma_2(\mathbf{D}_{\mathbf{x}^\star}^{1/2} \mathbf{P}\boldsymbol{\Pi}^\dagger \mathbf{D}_{\mathbf{x}^\star}^{-1/2}) \leq \sigma_2(\mathbf{D}_{\mathbf{x}^\star}^{1/2} \mathbf{P}\boldsymbol{\Pi}^\star \mathbf{D}_{\mathbf{x}^\star}^{-1/2}).$$

Main Limitation

Improvement on the upper-bound of the SLEM is **not guaranteed**:

$$\sigma_2(\mathbf{D}_{\mathbf{x}^\dagger}^{1/2} \mathbf{T}^\dagger \mathbf{D}_{\mathbf{x}^\dagger}^{-1/2}) \leq \sigma_2(\mathbf{D}_{\mathbf{x}^\star}^{1/2} \mathbf{T}^\star \mathbf{D}_{\mathbf{x}^\star}^{-1/2})$$

Future Works

- Guarantee monotonic improvement of the SLEM upper-bound
- Experimental evaluation
- Consider unknown model setting
- Extend to continuous state and action spaces

References

Stephen P. Boyd, Persi Diaconis, and Lin Xiao. Fastest mixing markov chain on a graph. *SIAM Review*, 2004.

Ilse C. F. Ipsen and Teresa M. Selee. Ergodicity coefficients defined by vector norms. *SIAM Journal on Matrix Analysis and Applications*, 2011.

Mirco Mutti and Marcello Restelli. An intrinsically-motivated approach for learning highly exploring and fast mixing policies. In *AAAI Conference on Artificial Intelligence*, 2020.

Jean Tarbouriech and Alessandro Lazaric. Active exploration in markov decision processes. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.