Evidence on the Regularisation Properties of Maximum-Entropy Reinforcement Learning

Rémy Hosseinkhan Boucher, Onofrio Semeraro, Lionel Mathelin,
LISN, CNRS, Université Paris-Saclay

What is the bias introduced by **entropy regularisation**? Are **complexity measures** linked to noise robustness?

Background

Partially Observable Markov Decision Process (POMDP)

$$X_{h+1} = F\left(X_h, U_h\right)$$
$$Y_h = G\left(X_h\right) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_Y^2 I_d)$$

F : state operator
G : observation operator
ϵ : Gaussian noise

Maximum-Entropy Objective

$$J^{\pi,\epsilon} = \mathbb{E}^{\pi,\epsilon}\left[\sum_{h=0}^{H} \gamma^h c\left(X_h, U_h\right)\right] - \alpha\ \mathbb{E}^{\pi,\epsilon}\left[\sum_{h=0}^{H} \gamma^h \mathcal{H}(\pi(\cdot \mid X_h))\right]$$

α : entropy coefficient
$\mathcal{H}$: entropy

Notations

P^{π}_{ϵ} : trajectory probability with **observation noise** $\mathbb{E}^{\pi,\epsilon}$: expectation under P^{π}_{ϵ}
$P^{\pi} = P^{\pi}_{0}$: no **noise** $(\epsilon \equiv 0)$

Excess Risk Metrics

$$\mathcal{R}^{\pi} = \mathbb{E}^{\pi,\epsilon}\left[\sum_{h=0}^{H} \gamma^h c\left(X_h, U_h\right)\right] - \mathbb{E}^{\pi}\left[\sum_{h=0}^{H} \gamma^h c\left(X_h, U_h\right)\right] = J^{\pi,\epsilon} - J^{\pi}$$

$$\mathring{\mathcal{R}}^{\pi} = \frac{J^{\pi,\epsilon} - J^{\pi}}{J^{\pi}} = \frac{\mathcal{R}^{\pi}}{J^{\pi}}$$

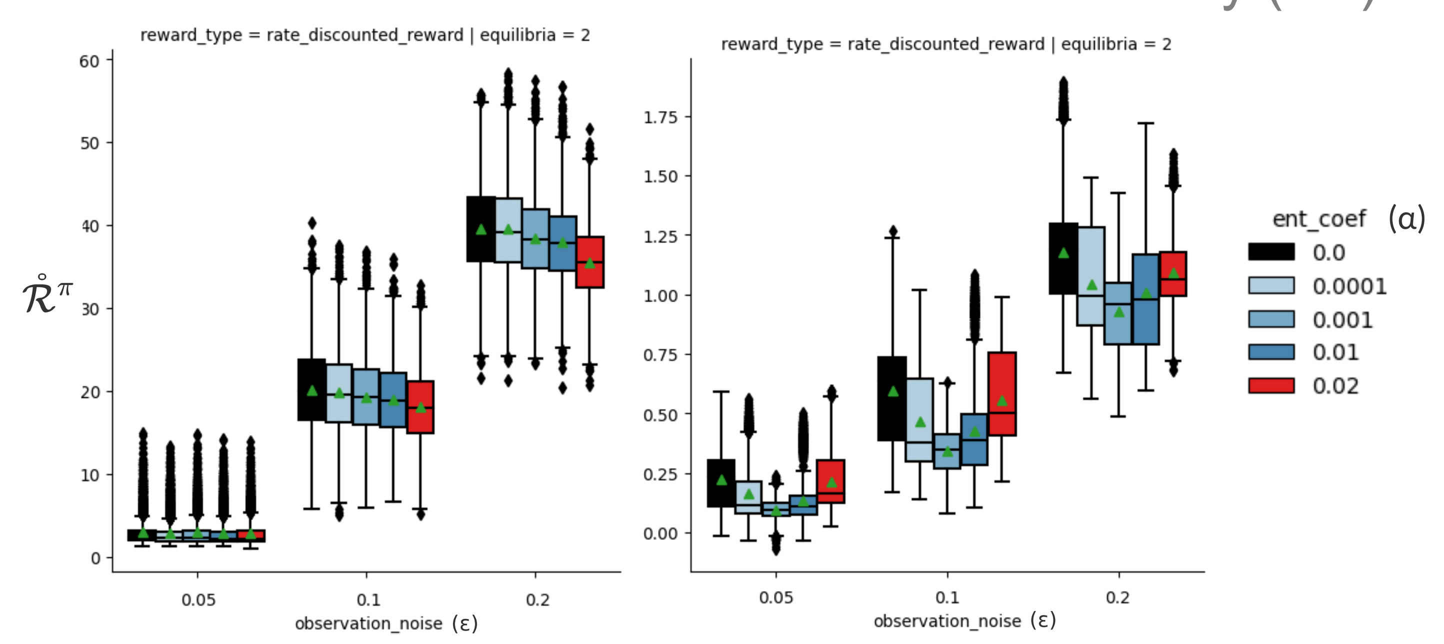
Goal

Evaluating the Robustness $\mathring{\mathcal{R}}^{\pi}$ of Max-Entropy Policies under Observation Noise $\epsilon \sim \mathcal{N}(0, \sigma_Y^2 I_d)$

Noise-free training with PPO $\longrightarrow$ $P^{\pi} = P^{\pi}_{0}$: no **noise** $(\epsilon \equiv 0)$
5 coeff. α (x 10 seeds) $\longrightarrow$ 5 x 10 policies $(\pi_{\theta^*_{\alpha}})$
Test $\pi_{\theta^*_{\alpha}}$ under different noise levels on Y

$\epsilon \nearrow \longrightarrow J^{\pi^*,\epsilon} \nearrow$ (noise impacts perf)
$\alpha > 0 \longrightarrow \mathring{\mathcal{R}}^{\pi,\alpha} \searrow$ (robustness)

RL Environments: Lorenz - Kuramoto-Sivashinsky (KS)



Variation $\frac{J^{\pi^*_{\alpha},\epsilon} - J^{\pi^*}}{J^{\pi^*}}$. Each **bar block** : noise intensity ϵ
Colors: $\alpha = 0$ (black), $\alpha > 0$ (blue), $\alpha_{\max}$ (red)

Find complexity measures $\mathcal{M}(\pi_\theta)$ controlling the Excess Risk $\mathring{\mathcal{R}}^{\pi}$

Parameterised Policy π_θ
Parameter Space $\theta \in \Theta$ $\longrightarrow$ **Complexity measure:** $\mathcal{M} : \Theta \to \mathbb{R}_+$
Complexity measures quantify model **complexity**
Regularisation $\longrightarrow$ Low complexity

Question

Which **complexity measures** indicate **noise robustness?** Why do **high entropy** policies learn **better final solutions?**

Contribution

Introduction of complexity measures from Statistical Learning

Norm-based Complexity Measures

$\pi_\theta(\cdot|X_k) \sim \mathcal{N}\left(\mu_{\theta_\mu}(X_k), \theta_{\sigma_\pi} I\right)$

If $\mu_\theta\left(x\right) = \left(\sigma_l \circ \sigma_{l-1} \circ \ldots \circ \sigma_1\right)(x)$, $Lips(\mu_\theta) \leq \prod_{i=1}^{l} Lips\left(\sigma_i\right) = \prod_{i=1}^{l} \|\theta_i\|$

Regularity - Lipshitz Continuity $\longrightarrow$ Lipshtitz bound (norm product)?

- $\mathcal{M}(\pi_\theta, \mathcal{D}) = \|\theta_\mu\|_p$
- $\mathcal{M}(\pi_\theta, \mathcal{D}) = \Pi_{i=1}^{l} \|\theta^i_\mu\|_p$ where θ^i_μ is the i^{th} layer of the network μ_{θ_μ}

Sharpness-based Complexity Measures

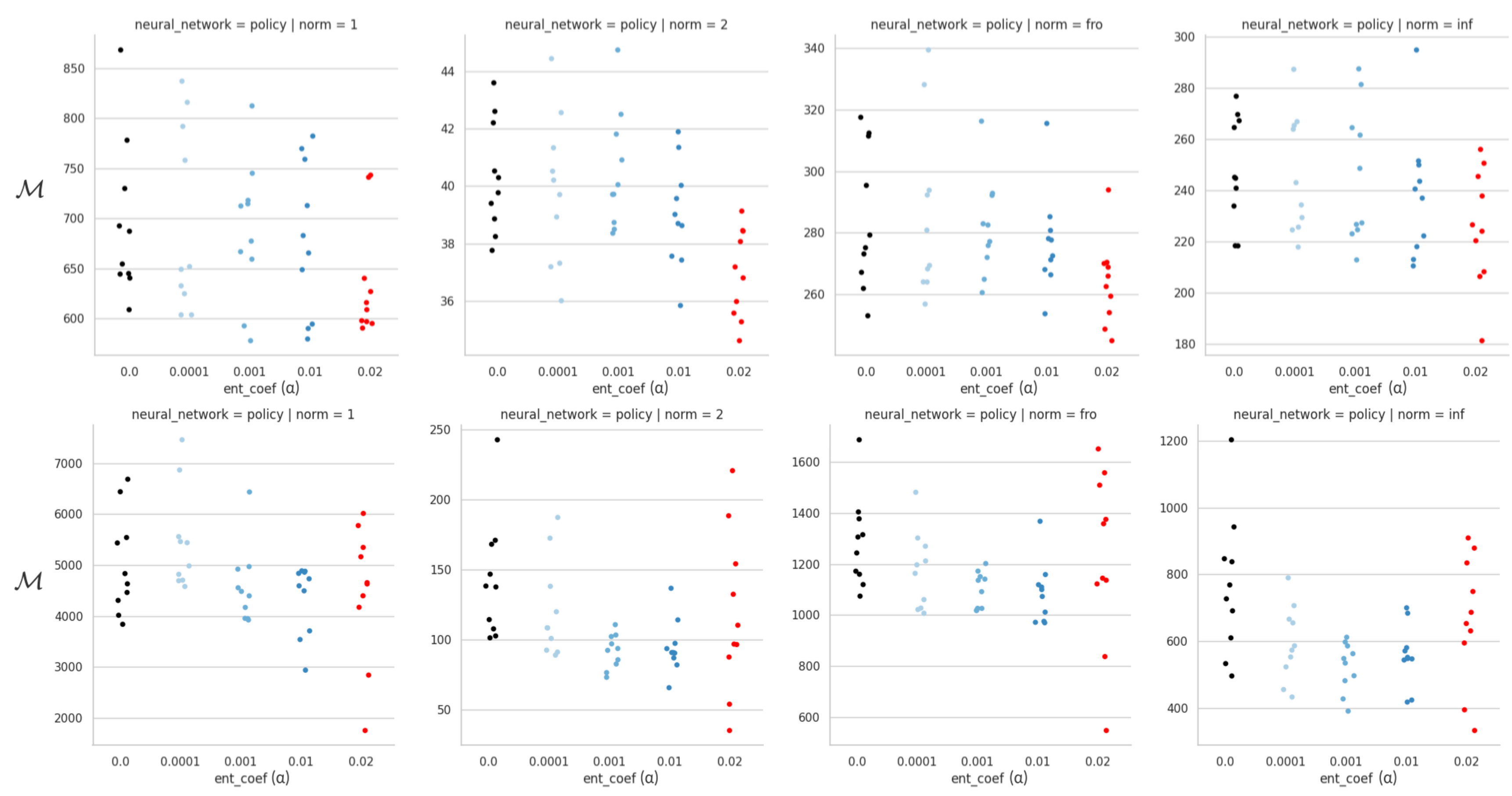
Curvature - Hessian $\longrightarrow$ Fisher Information?
Objective landscape - Robustness $\longrightarrow$ Flat minima?

$$\nabla^2_\theta J^{\pi_\theta} = \mathbb{E}^{\pi_\theta}\left[\sum_{h,i,j=0}^{H} c\left(X_h, U_h\right)\left(\nabla_\theta \log \pi_\theta\left(U_i \mid X_i\right) \nabla_\theta \log \pi_\theta\left(U_j \mid X_j\right)^T + \nabla^2_\theta \log \pi_\theta\left(U_i \mid X_i\right)\right)\right]$$

$$\mathcal{I}(\theta) = -\ \mathbb{E}^{X\sim\rho, U\sim\pi_\theta(\cdot|X)}\left[\nabla^2_\theta \log \pi_\theta(U \mid X)\right]$$

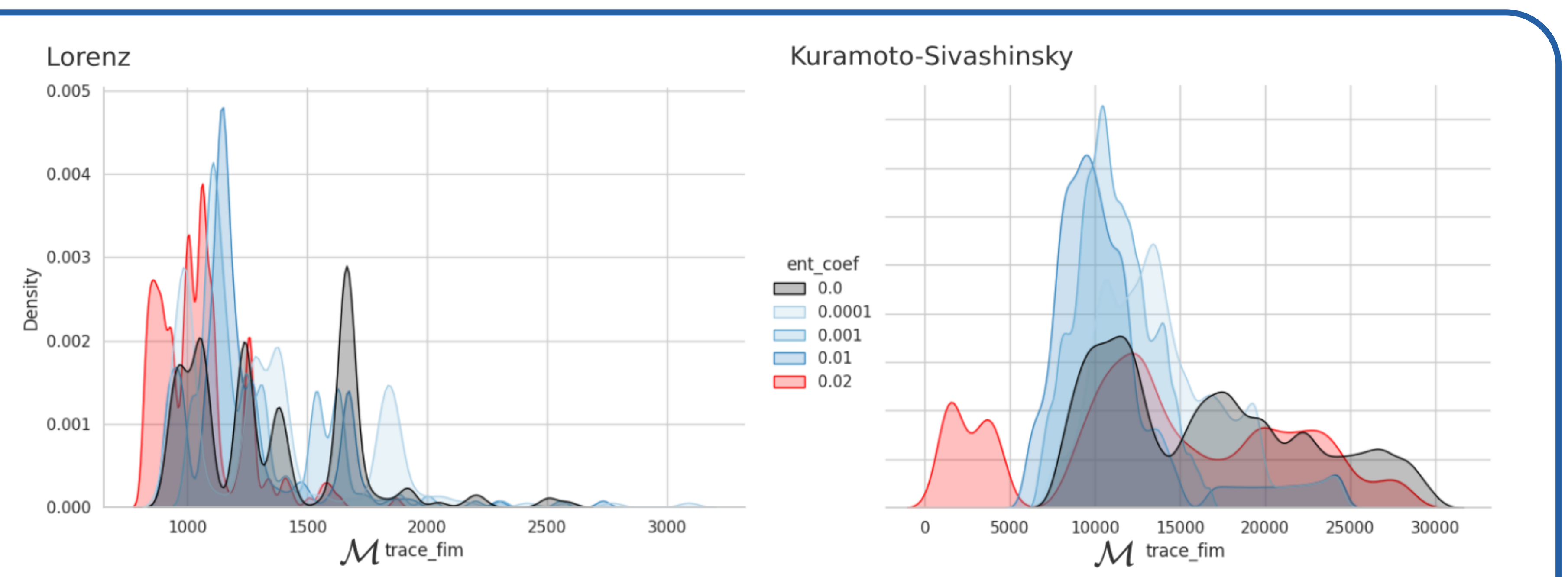
- $\mathcal{M}(\pi_\theta, \mathcal{D}) = Tr(\mathcal{I}\left(\theta_\mu\right)) = Tr(-\ \mathbb{E}^{X\sim\rho^\pi, U\sim\pi_\theta(\cdot|X)}\left[\nabla^2_{\theta_\mu} \log \pi_\theta(U \mid X)\right])$

Results



$\mathcal{M}\left(\pi^{\alpha}_{\theta}\right) = \prod_{i=1}^{l} \|\theta^{i\alpha}_{\mu}\|$
Colors: $\alpha = 0$, $\alpha > 0$, $\alpha_{\max}$
Top: Lorenz, Bottom: KS

✓ Low $\mathcal{M}\left(\pi^{\alpha}_{\theta_\mu}\right)$ corresponds to low $\mathring{\mathcal{R}}^{\pi}$



$\mathcal{M}\left(\pi^{\alpha}_{\theta}\right) = -Tr(\mathbb{E}^{X\sim\rho^{\pi_\theta}, U\sim\pi_\theta(\cdot|X)}\left[\nabla^2_{\theta_\mu} \log \pi_\theta\left(U|X\right)\right])$
Colors: $\alpha = 0$, $\alpha > 0$, $\alpha_{\max}$

✓ Measure $\mathcal{M}\left(\pi^{\alpha}_{\theta_\mu}\right)$ distribution with **fat-right tail (extremely large value)** for $\mathring{\mathcal{R}}$ large

✓ Complexity measures can explain noise robustness

Z. Ahmed et al. - Understanding the Impact of Entropy on Policy Optimization (ICML, 2019)
B. Neyshabur, S. Bhojanapalli, D. Mc Allester, N. Srebo - Exploring Generalization in Deep Learning (NIPS, 2017)