Evidence on the regularization properties of Maximum-Entropy Reinforcement Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

The generalisation and robustness properties of policies learnt through Maximum-Entropy Reinforcement Learning are investigated on chaotic dynamical systems with Gaussian noise on the observable. First, the robustness under noise contamination of the agent's observation of entropy regularised policies is observed. Second, notions of statistical learning theory, such as complexity measures on the learnt model, are borrowed to explain and predict the phenomenon. Results show the existence of a relationship between entropy-regularised policy optimisation and robustness to noise, which can be described by the chosen complexity measures.

1 Introduction

Maximum-Entropy Reinforcement Learning [Williams et al., 1991] aims to solve the problem of learning a policy which optimises a chosen utility criterion while promoting the entropy of the policy. The standard way to account for the constraint is to add a Lagrangian term to the objective function. This entropy-augmented objective is commonly referred to as the soft objective.

There are multiple advantages in solving the soft objective over the standard objective. For instance, favouring stochastic policies over deterministic ones allows learning multi-modal distributions [Haarnoja et al., 2017]. In addition, agent stochasticity is a suitable way to deal with uncertainty induced by Partially Observable Markov Decision Processes (PO-MDP). Indeed, there are PO-MDP such that the best stochastic adapted policy can be arbitrarily better than the best deterministic adapted policy [Sigaud and Buffet, 2010][1].
Furthermore, several important works highlight both theoretical and experimental *robustness* of those policies under noisy dynamics and rewards [Eysenbach and Levine, 2022].

Related to the latter notion of robustness, the maximum-entropy principle exhibits non-trivial generalisation capabilities, which are desired in real-world applications [Haarnoja et al., 2018].

However, the reasons for such robustness properties are not yet well understood. Thus, further investigations are needed to grasp the potential of the approach and to design endowed algorithms. A clear connection between Maximum-Entropy RL and their robustness properties is important and intriguing.

Meanwhile, recent work in the deep learning community discusses how some complexity measures on the neural network model are related to generalisation, and explain typically observed phenomena [Neyshabur et al., 2017]. In fact, these complexity measures are derived from the learnt model,

[1] In this context, the term "stochastic adapted policy" is a conditional distribution on the control space $\mathcal{U}$ given the observation space $\mathcal{Y}$ since this type of policy is "adapted" from Markovian policies in fully observable MDPs.

bound the PAC-Bayes generalisation error, and are meant to identify which of the local minima generalise well.

As a matter of fact, a relatively recent trend in statistical learning suggests generalisation is not only favored by the regularisation techniques (*e.g.*, dropout) but mainly because of the flatness of the local minima [Hochreiter and Schmidhuber, 1997, Dinh et al., 2017, Keskar et al., 2017]. The reasons for such regularity properties remain an open problem. This work aims to address these points in the context of Reinforcement Learning, and addresses the following questions:

What is the bias introduced by entropy regularisation? Are the aforementioned complexity measures also related to the robustness of the learnt solutions in the context of Reinforcement Learning?

In that respect, by defining a notion of robustness against noisy contamination of the observable, a study on the impact of the entropy regularisation on the robustness of the learnt policies is first conducted. After explaining the rationale behind the choice of the complexity measures, a numerical study is performed to validate the hypothesis that some measures of complexity are good robustness predictors. Finally, a link between the entropy regularisation and the flatness of the local minima is treated through the information geometry notion of Fisher Information.

The paper is organised as follows. Section 2 introduces the background and related work, Section 3 presents the problem setting. Section 4 is the core contribution of this paper. This section introduces the rationale behind the studied complexity measures from a learning theory perspective, as well as their expected relation to robustness. Lastly, Section 5 presents the experiments related to the policy robustness as well as their complexity, while Section 6 examines the results obtained. Finally, Section 7 concludes the paper.

2 Related work

Maximum Entropy Policy Optimisation In Haarnoja et al. [2018], the generalisation capabilities of entropy-based policies are observed where multimodal policies lead to optimal solutions. It is suggested that maximum entropy solutions aim to learn all the possible ways to solve a task. Hence, transfer learning to more challenging objectives is made easier, as demonstrated in their experiment. This study investigates the impact of adopting policies with greater randomness on their robustness. The impact of the entropy regularisation on the loss landscape has been recently studied in Ahmed et al. [2019]. They provide experimental evidence about the smoothing effect of entropy on the optimisation landscape. The present study aims specifically to answer the question in Section 3.2.4 of their paper: *Why do high entropy policies learn better final solutions?* This paper extends their results from a complexity measure point of view. Recently, Neu et al. [2017], Derman et al. [2021] studied the equivalence between robustness and entropy regularisation on regularised MDP.

Flat minima and Regularity The notion of local minima flatness was first introduced in the context of supervised learning by Hochreiter and Schmidhuber [1997] through the Gibbs formalism [Haussler and Opper, 1997]. Progressively, different authors stated the concept with geometric tools such as first order (gradient) or second order (Hessian) regularity measures [Zhao et al., 2022, Keskar et al., 2017, Sagun et al., 2017, Yoshida and Miyato, 2017, Dinh et al., 2017]. In a similar fashion, Chaudhari et al. [2019] uses the concept of local entropy to smooth the objective function.
In the scope of Reinforcement Learning, Ahmed et al. [2019] observed that flat minima characterise maximum entropy solutions, and entropy regularisation has a smoothing effect on the loss landscape, reducing the number of local optima. A central objective of this present study is to investigate this latter property further and relate it to the field of research on robust optimisation. Lastly, among the few recent studies on the learning and optimisation aspects of RL, Gogianu et al. [2021] shows how a well-chosen regularisation can be very effective for deep RL. Indeed, they explain that constraining the Lipschitz constant of only one neural network layer is enough to compete with state-of-the-art performances on a standard benchmark.

Robust Reinforcement Learning A branch of research related to this work is the study of robustness with respect to the uncertainty of the dynamics, namely *Robust Reinforcement Learning* (Robust RL), which dates back to the 1970s [Satia and Lave, 1973]. Correspondingly, in the field of control theory, echoes the notion of robust control and especially H_∞ control [Zhou et al., 1996], which also appeared in the mid-1970s after observing Linear Quadratic Regulator (LQR) solutions are very sensitive to perturbations while not giving consistent enough guarantees [Doyle, 1996]. More specifically, the Robust RL paradigm aims to control the dynamics in the worst-case scenario, *i.e.*, to optimise the minimal performance for a given objective function over a set of possible dynamics through a min-max problem formulation. This set is often called *ambiguity set* in the literature. It is defined as a region in the space of dynamics close enough w.r.t. to some divergence measure, such as the relative entropy [Nilim and Ghaoui, 2003]. Closer to this work, the recent paper from Eysenbach and Levine [2022] shows theoretically how Maximum-Entropy RL policies are inherently robust to a certain class of dynamics of fully-observed MDP. The finding of their article might still hold in the partially observable setting as any PO-MDP can be cast as fully-observed MDP with a larger state-space of probability measures [Hernández-Lerma and Lasserre, 1996], providing the ambiguity set is adapted to a more complicated space.

3 Problem Setup and Background

3.1 Partially Observable Markov Decision Process with Gaussian noise

First, the control problem when noisy observations are available to the agent is formulated. The study focuses on *Partially Observable Markov Decision Processes (PO-MDP)* with Gaussian noise of the form [Deisenroth and Peters, 2012]:

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

with $X_h \in \mathcal{X}$, $U_h \in \mathcal{U}$ and $Y_h \in \mathcal{Y}$ for any $h \in \mathbb{N}$, where $\mathcal{X}$, $\mathcal{U}$ and $\mathcal{Y}$ are respectively the corresponding state, action and observation spaces. The initial state starts from a reference state x_e^* on which centred Gaussian noise with diagonal covariance $\sigma_e^2 I_d$ is additively applied, $X_0 \sim \mathcal{N}(x_e^*,\, \sigma_e^2 I_d)$. Associated with the dynamics, an instantaneous cost function $c : \mathcal{X} \times \mathcal{U} \to \mathbb{R}_+$ is also given to define the control model.

In this context, a *policy* π is a transition kernel on $\mathcal{A}$ given $\mathcal{Y}$, *i.e.*, a distribution on actions conditioned on observations. This kind of policies are commonly used in the literature but can be very poor in the partially observable setting where information is missing. Together, a control model, a policy π and an initial distribution P_{X_0} on $\mathcal{X}$ define a stochastic process with distribution $P^{\pi,\epsilon}$ where the superscript ϵ highlights the dependency on the observation noise ϵ. Similarly, one denotes by P^{π} the distribution of the process when the noise is zero almost-surely, *i.e.*, $P^{\pi} = P^{\pi,0}$. More details about the PO-MDP control problem can be found in Hernández-Lerma and Lasserre [1996], Cassandra [1998].

Here, the maximum-entropy control problem is to find a policy π^* which minimises the following performance criterion

$$J_m^{\pi,\epsilon} = \mathbb{E}^{\pi,\epsilon}\left[\sum_{h=0}^{H} \gamma^h c\left(X_h, U_h\right)\right] + \alpha_m \mathbb{E}^{\pi,\epsilon}\left[\sum_{h=0}^{H} \gamma^h \mathcal{H}(\pi(\,\cdot \mid X_h))\right], \tag{2}$$

where $H \in \mathbb{N}$ is a given time horizon, $\mathbb{E}^{\pi,\epsilon}$ denotes the expectation under the probability measure $P^{\pi,\epsilon}$, $\mathcal{H}$ denotes the differential entropy [Cover and Thomas, 2006] and α_m is a time-dependent weighting parameter that evolves over training time $m \leq m_{\mathcal{D}} = |\mathcal{D}|$ with $|\mathcal{D}|$ being the total number of times the agent interacts with the system such that all observations used by the learning algorithm form the dataset $\mathcal{D}$ at the end of the training procedure (when $m_{\mathcal{D}}$ environment interactions are done).

$J_m^{\pi,\epsilon}$ is denoted $J^{\pi,\epsilon}$. The quantity $J^{\pi,\epsilon}$ is called the value function or, more generally, *loss*. Moreover, the performance gap for dynamics with noisy and noiseless observables will be considered in the sequel. In this context, the *(rate of) excess risk under noise* is defined as the difference between the loss under noisy dynamics and the loss under noiseless dynamics:

Definition 1 (Excess Risk Under Noise) *The excess risk under noise of a policy π for a PO-MDP with dynamics* (1) *is defined as:*

$$\mathcal{R}^{\pi} = \mathbb{E}^{\pi,\epsilon}\left[\sum_{h=0}^{H} \gamma^h c\left(X_h, U_h\right)\right] - \mathbb{E}^{\pi}\left[\sum_{h=0}^{H} \gamma^h c\left(X_h, U_h\right)\right] = J^{\pi,\epsilon} - J^{\pi} \tag{3}$$

Similarly, the rate of excess risk under noise is defined as:

$$\mathring{\mathcal{R}}^{\pi} = \frac{J^{\pi,\epsilon} - J^{\pi}}{J^{\pi}} = \frac{\mathcal{R}^{\pi}}{J^{\pi}} \tag{4}$$

Note that in the above definition,

expectations are taken with respect to the probability measure $P^{\pi,\epsilon}$ and P^{π} respectively. The *rate of excess risk under noise* represents the performance degradation after noise introduction in value function units. In the rest of the paper, arguments to derive complexity measures will be developed, allowing to predict the excess risk under noise and provide numerical evidence showing maximum-entropy policies are more robust regarding this metric. Hence, maximum-entropy policies implicitly learn a robust control policy in the sense of Definition 1.

In the next section, some concepts of statistical learning theory are introduced. Then, complexity measures will be defined to quantify the regularisation power of the maximum-entropy objective of (2).

4 Complexity Measures and Robustness

4.1 Complexity Measures

The principal objective of *statistical learning* is to provide bounds on the generalisation error, so-called *generalisation bounds*. In the following, it is assumed that an algorithm $\mathcal{A}$ returns a hypothesis $\pi \in \mathcal{F}$ from a dataset $\mathcal{D}$. Note that the dataset $\mathcal{D}$ is random and the algorithm $\mathcal{A}$ is a randomised algorithm.

As the hypothesis set $\mathcal{F}$ typically used in machine learning is infinite, a practical way to quantify the generalisation ability of such a set must be found. This quantification is done by introducing *complexity measures*, enabling the derivation of generalisation bounds.

Definition 2 (Complexity measure) *A complexity measure is a mapping $\mathcal{M} : \mathcal{F} \to \mathbb{R}_+$ that maps a hypothesis to a positive real number.*

According to Neyshabur et al. [2017] from which this formalism is inspired, an appropriate complexity measure satisfies several properties. In the case of parametric models $\pi_\theta \in \mathcal{F}(\Theta)$ with $\theta \in \Theta \subset \mathbb{R}^b$, it should increase with the dimension b of the parameter space Θ as well as being able to identify when the dataset $\mathcal{D}$ contains totally random, spurious or adversarial data. As a result, finding good complexity measures $\mathcal{M}$ allows the quantification of the generalisation ability of a hypothesis set $\mathcal{F}$ or a model π and an algorithm $\mathcal{A}$.

4.2 Complexity measures for PO-MDP with Gaussian Noise

This paper studies heuristics about generalisation bounds on the optimal *excess risk under noise* from Definition 1 when the optimal policy π_{θ^*} is learnt with an algorithm $\mathcal{A}$ on the non-noisy objective J^{π}, where $\alpha_m = 0$ for any m.

Definition 3 ((Rate of) Excess Risk Under Noise Bound) *Given an optimal policy π^* learnt with an algorithm $\mathcal{A}$ on the non-noisy objective J^{π}, the optimal excess risk under noise bound is a real-valued mapping φ such that*

$$\mathcal{R}^{\pi^*} \leq \varphi(\mathcal{M}(\pi^*, \mathcal{D}),\, m_{\mathcal{D}},\, \eta,\, \delta) \tag{5}$$

and φ is increasing with the complexity measure $\mathcal{M}$ and the sample complexity $m_{\mathcal{D}}$. The definition is similar for the rate of excess risk under noise bound where $\mathring{\mathcal{R}}^{\pi^*}$ is used instead of $\mathcal{R}^{\pi^*}$.

Hence, considering a learning algorithm $\mathcal{A}$ with a parameterised family $\mathcal{F}(\Theta) = (\pi_\theta)_{\theta \in \Theta}$, $\Theta \subset \mathbb{R}^b$, such that $\theta = (\theta_\mu,\, \theta_{\sigma_\pi})$ with $\pi_\theta(\cdot \mid x) \sim \mathcal{N}(\mu_{\theta_\mu}(x), \operatorname{diag}(\theta_{\sigma_\pi}))$, $x \in \mathcal{X}$, - where μ_{θ_μ} is a shallow multi-layer feed-forward neural network (with depth-size $l = 2$, width $w = 64$ neurons, weights matrix $(\theta_\mu^i)_{1 \leq i \leq l}$) and $\operatorname{diag}(\theta_{\sigma_\pi})$ is a diagonal matrix of dimension $q = \dim(\mathcal{U})$ parameterising the variance[2] - to learn the optimal policy π_{θ^*}, multiple complexity measures $\mathcal{M}$ are defined and details on their underlying rationale are given below.

4.2.1 Norm based complexity measures

First, the so-called norm-based complexity measures are functions of the norm of some subset of the parameters of the model. For instance, a common norm-based measure calculates the product of the operator norms of the neural network linear layers. The measures are commonly used in the statistical learning theory literature to derive bounds on the generalisation gap, especially in the context of neural networks [Neyshabur et al., 2015, Golowich et al., 2018, Miyato et al., 2018].
In fact, the product of the norm of the linear layers of a standard class of multi-layer neural networks (including Convolutional Neural Networks) serves as an upper bound on the often intractable Lipschitz constant of the network [Miyato et al., 2018]. Thus, controlling the magnitude of the weights of the linear layers increases the regularity of the model.
Consequently, the following complexity measures are defined:

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

In this context $\|\cdot\|_p$ with $p = 1, 2, \infty$ denotes the p-operator norm while $p = F$ denotes the Frobenius norm, which is discarded for the first case of the full parameters vector θ_μ (since Frobenius norm is defined for matrix).

4.2.2 Flatness based complexity measures

On the other hand, another measure of complexity is given by the flatness of the optimisation local minimum (see Section 2 for a brief overview). As McAllester [2003], Neyshabur et al. [2017] have pointed out, the generalisation ability of a parametric solution is controlled by two key components in the context of supervised learning: the norm of the parameter vector and its flatness w.r.t. to the objective function.

One might wonder if a similar robustness property still holds in the setting of Reinforcement Learning. In this manner, complexity measures quantifying the flatness of the solution are needed. Concretely, the interest lies in the flatness of the local minima of the objective function J^π. As stated earlier, there are several ways to quantify the flatness of a solution with metrics derived from the gradient or curvature of the loss function at the local optimum, such as the Hessian's largest eigenvalue - otherwise spectral norm [Keskar et al., 2017] or the trace of Hessian [Dinh et al., 2017].

Moreover, as discussed in Section 2, Ahmed et al. [2019] observed that *maximum entropy solutions are characterised by flat minima* while entropy regularisation has a smoothing effect on the loss landscape. *Hence, a central objective of this present study is to investigate this latter property further and relate it to the robustness aspect of the resulting policies.*

However, instead of dealing directly with the Hessian of the objective J^π this work proposes a measure based on the conditional Fisher Information $\mathcal{I}$ of the policy due to its link with a notion of model regularity in the parameter space.

Definition 4 (Conditional Fisher Information Matrix) *Let $x \in \mathcal{X}$ and π_θ a policy identified by its conditional density for a parameter $\theta \in \Theta \subset \mathbb{R}^b$ and suppose ρ is a distribution over $\mathcal{X}$. The conditional Fisher Information Matrix of the vector θ is defined under some regularity conditions as*

$$\mathcal{I}(\theta) = -\,\mathbb{E}^{X\sim\rho, U\sim\pi_\theta(\cdot|X)}\left[\nabla_\theta^2 \log \pi_\theta(U \mid X)\right], \tag{6}$$

where ∇_θ^2 denotes the Hessian matrix evaluated at θ.

Note that the distribution over states ρ is arbitrary and can be chosen as the discounted state visitation measure ρ^π induced by the policy π [Agarwal et al., 2019] or the stationary distribution of the induced Markov process if the policy is Markovian and the MDP ergodic[3] as it is done in Kakade [2001].

[2] Note this choice of state-independent policy variance is inspired by Ahmed et al. [2019] and simplifies the problem.

[3] With these choices, the following holds: $\mathbb{E}^{\rho^\pi(ds)\pi(da|s)} = \mathbb{E}^\pi$ up to taking the expectation *w.r.t.* the state-action space (no subscript under X and U) or the trajectory space (with subscripts such as X_h and U_h as trajectory coordinate) Agarwal et al. [2019].

As a matter of fact, it has already been mentioned in the early works of policy optimisation [Kakade, 2001] that this quantity $\mathcal{I}$ might be related to the Hessian of the objective function. Indeed, the Hessian matrix of the standard objective function reads (see Shen et al. [2019] for a proof):

$$\nabla_\theta^2 J^{\pi_\theta} = \mathbb{E}^{\pi_\theta}\left[\sum_{h,i,j=0}^{H} c\left(X_h, U_h\right)\Big(\nabla_\theta \log \pi_\theta\left(U_i \mid X_i\right) \nabla_\theta \log \pi_\theta\left(U_j \mid X_j\right)^T + \nabla_\theta^2\left[\log \pi_\theta\left(U_i \mid X_i\right)\right]\Big)\right]. \tag{7}$$

As suggested by the author mentioned above (S. Kakade), (7) might be related to $\mathcal{I}$ although being weighted by the cost c. Indeed, the Hessian of the state-conditional log-likelihoods ($\nabla_\theta^2 \log \pi_\theta$ on the rightmost part of the expectation of (7)) belongs to the objective-function Hessian $\nabla_\theta^2 J^{\pi_\theta}$ while the Fisher Information $\mathcal{I}(\theta)$ is an average of the Hessian of the policy log-likelihood.

In any case, the conditional FIM measures the regularity of a critical component of the objective to be minimised. Thus, the trace of the conditional FIM of the mean actor network parameter θ_μ is suggested as a complexity measure

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

Moreover, in the context of classification, a link between the degree of stochasticity of optimisation gradients (leading to flatter minima [Mulayoff and Michaeli, 2020, Xie et al., 2021]) and the FIM trace during training has recently been revealed in Jastrzebski et al. [2021]. Magnitudes of the FIM eigenvalues may be related to loss flatness and norm-based capacity measures to generalisation ability [Karakida et al., 2019] in deep learning.

5 Experiments

5.1 Robustness under noise of Maximum Entropy Policies

The first hypothesis is that maximum entropy policies are more robust to noise than those trained without entropy regularisation (which play the role of control experiments). Consequently, the robustness of the controlled policy π_{θ^*} is compared with the robustness of the maximum entropy policy $\pi_{\theta^*}^{\alpha}$ for different temperature evolutions $\alpha = (\alpha_m)_{0\leq m\leq m_\mathcal{D}}$. In this view, and since inter-algorithm comparisons are characterised by high uncertainty [Henderson et al., 2018, Colas et al., 2018, Agarwal et al., 2021], only one algorithm $\mathcal{A}$ (*Proximal Policy Optimisation* (PPO) Schulman et al. [2017]) is retained while results on multiple entropy constraint levels $\alpha = (\alpha_m)_{0\leq m\leq m_\mathcal{D}}$ are examined.

In this regard, ten independent PPO models are trained for each of the five arbitrarily chosen entropy temperatures $\alpha^i = (\alpha_m^i)_{0\leq m\leq m_\mathcal{D}}$ where $i \in \{1, \ldots, 5\}$, on dynamics without observation noise, *i.e.*, where $\sigma_Y^2 = 0$. The entropy coefficients linearly decay during training, and all vanish ($\alpha_m = 0$) when m reaches one-fourth of the training time $m_{1/4} = \lfloor \frac{m_\mathcal{D}}{4} \rfloor$ in order to replicate a sort of exploration-exploitation procedure, ensuring that all objectives J_m^π are the same whenever $m \geq m_{1/4}$, *i.e.*, $J_m^\pi = J^\pi$. This choice is different but inspired by Ahmed et al. [2019] as they optimise using only the *policy gradient* and manipulate the standard deviation of Gaussian policies directly, whereas, in the present approach, it is done implicitly with an adaptive entropy coefficient. An algorithm that learns a model with a given entropy coefficient $\alpha = (\alpha_m)_{0\leq m\leq m_\mathcal{D}}$ is denoted as $\mathcal{A}_\alpha$.

The chosen chaotic systems are the *Lorenz* [Vincent and Yu, 1991] (with $m_\mathcal{D} = 10^6$) and *Kuramoto-Sivashinsky (KS)* [Bucci et al., 2019] (with $m_\mathcal{D} = 2 \cdot 10^6$) controlled differential equations. The defaults training hyper-parameters from *Stable-Baselines3* [Raffin et al., 2021] are used.

5.2 Robustness against Complexity Measures

So far, three separate analyses on the 5 × 10 models obtained have been performed on the *Lorenz* and *Kuramoto-Sivashinsky (KS)* controlled differential equations.
First, as mentioned before, the robustness of the models for each of the chosen entropy temperatures α^i is tested against the same dynamics but now with a noisy observable, *i.e.*, $\sigma_Y > 0$. Second, norm-based complexity measures introduced in Section 4.2 are evaluated and compared to the generalisation performances of the distinct algorithms $\mathcal{A}_\alpha$. Third, numerical computation of the

conditional distribution of the trace of the Fisher Information Matrix given by (6) is performed to test the hypothesis that this regularity measure is an indicator of robust solutions. The state distribution ρ^{π_θ} is naturally chosen as the state visitation distribution induced by the policy π_θ. The following section discusses the results of those experiments.

6 Results

This section provides numerical evidence of maximum entropy's effect on the robustness, as defined by the Excess Risk Under Noise defined by (3). Then, after quantifying robustness, the relation between the complexity measures defined in Section 4.2 and robustness is studied.

6.1 Entropy Regularisation induces noise robustness

In the first place, a distributional representation[4] of the rate of excess risk under noise defined in (3) is computed for each of the 5×10 models obtained with the PPO algorithm $\mathcal{A}_{\alpha^i}$, $i \in \{1, \dots, 5\}$ and different levels of observation noise $\sigma_Y > 0$.

First and foremost, the results shown in Figure 1 indicate that the noise introduction to the system observable Y of KS and Lorenz leads to a global decrease in performance, as expected.

The robustness to noise contamination of the two systems is improved by initialising the policy optimisation procedure up to a certain intermediate threshold of the entropy coefficient $\alpha^i > 0$. Once this value is reached, two respective behaviours are observed depending on the system. In the case of the Lorenz dynamics, the robustness continues to improve after this entropy threshold, whereas the opposite trend is observed for KS (particularly with the maximal entropy coefficient chosen).

Hence, the sole introduction of entropy-regularisation in the objective function impacts the robustness. This behaviour difference between Lorenz and KS might be explained by the variability of the optimisation landscapes that can be observed with respect to the chosen underlying dynamics as underlined in Ahmed et al. [2019].

6.2 Maximum entropy as a norm-based regularisation on the policy

Norm-based complexity measures introduced in Section 4.2 are now evaluated. For a complexity measure $\mathcal{M}$ to be considered significant, it should be correlated with the robustness of the model.

Accordingly, the different norm-based measures presented in Section 4.2 are estimated. Figure 2 shows the layer-wise product norm of the policy actor network parameters ($\mathcal{M}(\pi_\theta, \mathcal{D}) = \Pi_{i=1}^{l} \|\theta_\mu^i\|_p$) w.r.t. to their associated entropy coefficient α^i for all the 50 independently trained models.

Again, policies obtained with initial $\alpha^i > 0$ exhibit a trend toward decreasing complexity measure values as α increases up to a certain threshold of the entropy coefficient. Similarly to Section 6.1, the complexity measure continues to decrease after surpassing this threshold for the Lorenz system. On the other hand, in the KS case, $\mathcal{M}(\pi_\theta, \mathcal{D})$ increases again once its entropy threshold is reached, notably for the larger entropy coefficient.

Moreover, the measures tend to be much more concentrated when $\alpha^i > 0$, especially in the case of KS (except for the higher α^i).

This may indicate that the entropy regularisation acts on the uncertainty of the policy parameters. Likewise, similar observations can be made for the total norm of the parameters but are not introduced here for the sake of brevity.

Consequently, this experiment highlights an existing correlation between maximum entropy regularisation and norm-based complexity measures. As this complexity measure is linked to the Lipschitz continuity of the policy, one might wonder if the regularity of the policy is more directly impacted. This is the purpose of the next subsection.

[4] By replacing the expectation operator $\mathbb{E}$ with the conditional expectation $\mathbb{E}[\,\cdot \mid X_0]$ in the definition of $\mathcal{R}^\pi$ in (3), the quantity becomes a random variable for which the distribution can be estimated by sampling the initial state distribution $X_0 \sim \mathcal{N}(x_e^*, \sigma_e^2 I_d)$. In fact, taking the conditional expectation gives the difference of the standard *value functions* under P^π and $P^{\pi,\epsilon}$.

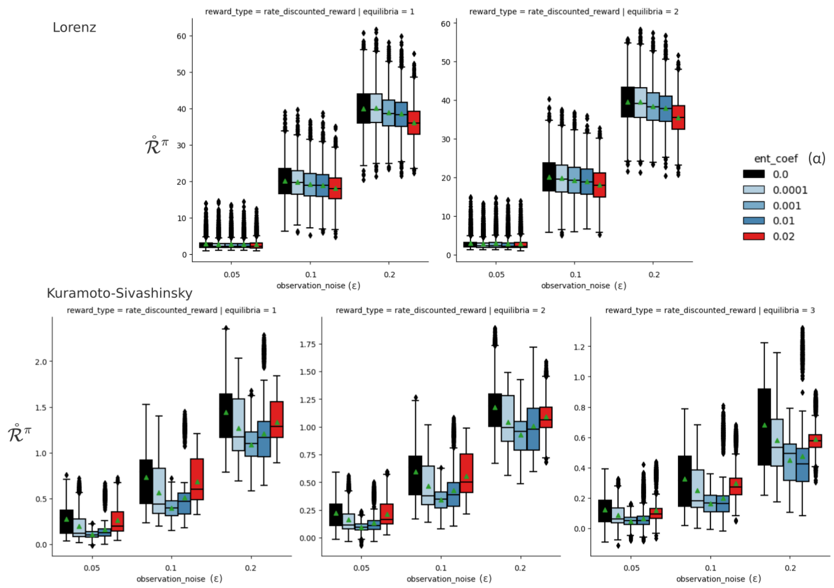


Figure 1: Distributional representation of the rate of excess risk under noise $\mathring{\mathcal{R}}^\pi$ conditioned on the α^i used during optimisation for different initial state distribution $X_0 \sim \mathcal{N}(x_e^*,\, \sigma_e^2 I_d)$. Each of the rows corresponds to one of the dynamical systems of interest. Each of the columns corresponds to one of the initial state distributions of interest. There are two non-zero fixed points (equilibria) x_e^* for Lorenz and three for KS. From top to bottom: KS; Lorenz.
For each box plot, three intensities σ_Y for the observation noise ϵ are evaluated. As expected, when the uncertainty regarding the observable Y increases through the variance σ_Y of the observation signal noise ϵ, the policy performance decreases globally ($\mathring{\mathcal{R}}^\pi$ increases). Moreover, the rate of excess risk under noise tends to decrease when α^i increases in the Lorenz case, whereas it decreases up to a certain entropy threshold for KS before increasing again.

6.3 Maximum entropy reduces the average Fisher-Information

Another regularity measure is considered: the average trace of the Fisher information ($\mathcal{M}(\pi_\theta, \mathcal{D}) = Tr(\mathcal{I}\left(\theta_\mu\right)) = Tr(-\,\mathbb{E}^{X\sim\rho, U\sim\pi_\theta(\cdot|X)}\left[\nabla^2_{\theta_\mu} \log \pi_\theta(U \mid X)\right]))$. As discussed in 4.2.2, this quantity reflects the regularity of the policy and might be related to the flatness of the local minima of the objective function.

Figure 3 shows the distribution under π_θ of the trace of the state conditional Fisher Information of the numerical optimal solution θ^*_{μ,α^i} for the policy w.r.t. the α^i used during optimisation. In other words, a kernel density estimator of the distribution of $Tr(\mathcal{I}(\pi_{\theta^*_{\mu,\alpha^i}}(\,\cdot \mid X)))$ when $X \sim \rho^{\pi_{\theta^*}}$ is represented. The results of this experiment suggest first, this distribution is skewed negatively and has a fat right tail. This means some regions of the support of $\rho^{\pi_{\theta^*}}$ provide FIM trace with extreme positive values, meaning the regularity of the policy may be poor in these regions of the state space. A comparison of the distribution w.r.t. the different α^i sheds further light on the relation between robustness and regularity. In fact, there appears to be a correspondence between the robustness, as indicated by the rate of excess risk under noise $\mathring{\mathcal{R}}^\pi$ shown in Figure 1 and the concentration of the trace distribution toward larger values (*i.e.* more irregular policies) when the model is less robust.

Meanwhile, under the considerations of 4.2.2 and since it is known that entropy regularisation favours flat minima in RL [Ahmed et al., 2019], these experimental results support the hypothesis of an existing relationship between robustness, objective function flatness around the solution θ^* and conditional Fisher information of θ^*.

For a complementary point of view, a supplementary experiment regarding the sensitivity of the policy updates during training w.r.t. to different level of entropy is also presented in Appendix A.

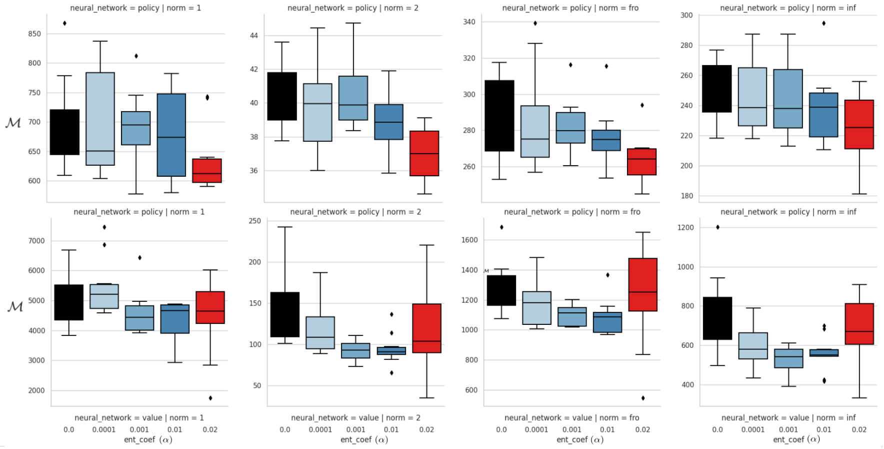


Figure 2: Measures of complexity $\mathcal{M}(\pi_\theta, \mathcal{D}) = \Pi_{i=1}^{l} \|\theta_\mu^i\|_p$ with $p = 1,\ 2,\ \infty,\ F$ conditioned on the α^i used during optimisation. Each row corresponds to one of the dynamical systems of interest while column represents a different norm order p. From top to bottom: Lorenz and KS.
For the Lorenz case, the barycenters of the measures tend to decrease when α^i increases. Regarding KS, passing a threshold, the complexity increases again with the entropy. In addition, the measures are much more concentrated when $\alpha^i > 0$. For $p = 2,\ F$, the separation of the measures w.r.t. the different α^i is more pronounced.

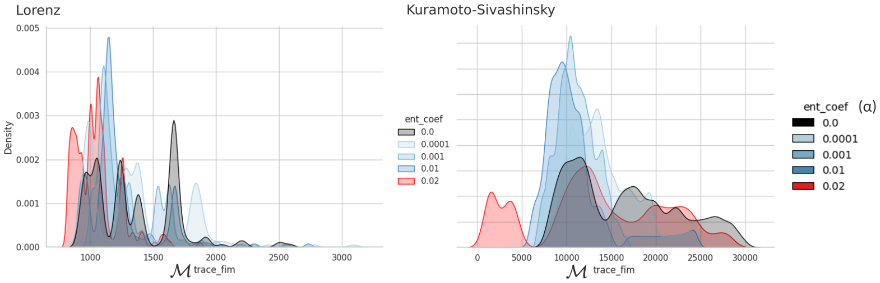


Figure 3: Distribution of the trace of the (conditional) Fisher information of the numerical optimal solution θ^*_{μ,α^i} for the policy w.r.t. the α^i used during optimisation. From left to right: Lorenz and KS environments. Colours: control experiment $\alpha^i = 0$ (black); intermediate entropy level α^i (blue); largest α^i (red).
A skewed distribution towards (relatively) larger values is observed for all controlled dynamical systems. Moreover, those right tails exhibit high kurtosis, especially for the control experiment (black) and the model with the larger entropy coefficient (red) for the KS system. Finally, solutions with intermediate entropy levels (blue) are much more concentrated - have lower variance than the others. About Lorenz, the barycenter of the more robust model (red) is shifted towards lower values than the others.

7 Discussion

In this paper, the question of the robustness of maximum entropy policies under noise is studied. After introducing the notion of complexity measures from the statistical learning theory literature, numerical evidence supports the hypothesis that maximum entropy regularisation induces robustness under noise. Moreover, norm-based complexity measures are shown to be correlated with the robustness of the model. Then, the average trace of the Fisher Information is shown to be a relevant indicator of the regularity of the policy. This suggests the existence of a link between robustness, regularity and entropy regularisation. Finally, this work contributes to bringing statistical learning concepts such as flatness into the field of Reinforcement Learning. New algorithms or metrics, such as in the work of Lecarpentier et al. [2021], may be built upon notions of regularity, *e.g.*, Lipschitz continuity, flatness or Fisher Information of the parameter in order to achieve robustness.

Acknowledgements

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR), under grant ANR-REASON (ANR-21-CE46-0008). This work was performed using HPC resources from GENCI-IDRIS (Grant 2023-[AD011014278]).

A Weights sensitivity during training

This section is intended to provide complementary insights on the optimisation landscape induced by the entropy coefficient α during training from the *conservative* or *trust region* policy iteration point of view Kakade and Langford [2002], Schulman et al. [2015].

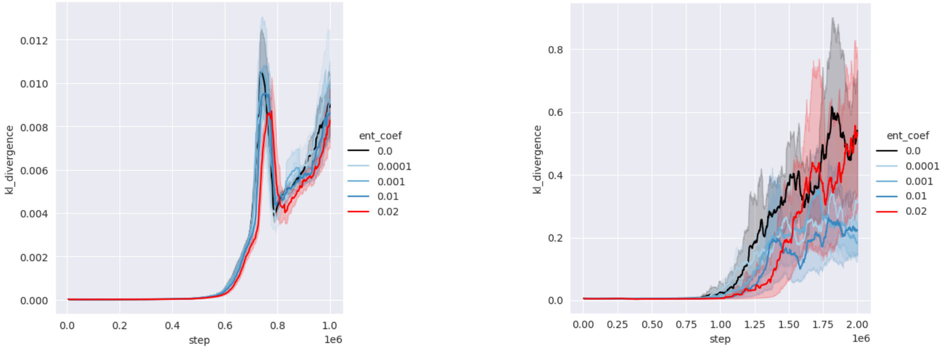
Let $(\theta_m^\alpha)_{m=1}^{m_\mathcal{D}}$ be the sequence of weights of the policy during the training of the model for some initial entropy coefficient α. The conditional Kullback-Leibler divergence between the policy identified by the parameters θ_m^α and the subsequent policy defined by the parameters θ_{m+1}^α is given by

$$\overline{D}_{KL}\left(\theta_m^\alpha,\, \theta_{m+1}^\alpha\right) = \mathbb{E}^{X\sim\rho}\left[\int_{\mathcal{U}} \log\left(\frac{\pi_{\theta_m^\alpha}(du|X)}{\pi_{\theta_{m+1}^\alpha}(du|X)}\right)\pi_{\theta_{m+1}^\alpha}(du\mid X)\right].$$

The above quantity is a measure of the divergence from the policy at time m to the policy at time $m+1$. Thus it may provide information on the local stiffness of the optimisation landscape during training.

Figure 4 shows the evolution of the Kullback-Leibler divergence between two subsequent policies during training for the Lorenz and KS controlled differential equations. Regarding the Lorenz system, the maximal divergence is reached for the optimisation performed with the two lowest α^i while increasing entropy seems to slightly reduce the divergence. On the other hand, the highest divergence values observed for the KS system are reached for $\alpha^i = 0$ and the maximal entropy coefficient. This observation is coherent with the results of the previous sections and suggests that the entropy coefficient α impacts the optimisation landscape during training.

Interesting questions regarding the optimisation landscape and its link with the Fisher Information (through the point of view of Information Geometry [Amari, 1998]) are raised by the results of this section but are left for future work.



(a) Lorenz

(b) Kuramoto-Sivashinsky

Figure 4: Evolution of $\overline{D}_{KL}\left(\theta_m^\alpha,\, \theta_{m+1}^\alpha\right)$ during training for the Lorenz and KS controlled differential equations. For Lorenz, the maximal divergence is reached for the optimisation performed with $\alpha^i = 0$ and the second lowest α^i. Regarding KS, the highest divergence values are observed for $\alpha^i = 0$ and the maximal entropy coefficient.

References

Ronald J. Williams, Jing Peng, and Hong Li. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268, 1991.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th*

International Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, pages 1352–1361. PMLR, 06–11 Aug 2017.

Olivier Sigaud and Olivier Buffet. *Markov Decision Processes in Artificial Intelligence*. Wiley, 2010. ISBN 978-1-848-21167-4.

Benjamin Eysenbach and Sergey Levine. Maximum entropy RL (provably) solves some robust RL problems. In *International Conference on Learning Representations*, 2022.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870. PMLR, 10–15 Jul 2018.

Behnam Neyshabur, Srinadh Bhojanapalli, David Mcallester, and Nati Srebro. Exploring generalization in deep learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Sepp Hochreiter and Jürgen Schmidhuber. Flat Minima. *Neural Computation*, 9(1):1–42, 01 1997. ISSN 0899-7667.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1019–1028. PMLR, 06–11 Aug 2017.

Nitish Shirish Keskar, Jorge Nocedal, Ping Tak Peter Tang, Dheevatsa Mudigere, and Mikhail Smelyanskiy. On large-batch training for deep learning: Generalization gap and sharp minima. 2017. 5th International Conference on Learning Representations, ICLR 2017 ; Conference date: 24-04-2017 Through 26-04-2017.

Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 151–160. PMLR, 09–15 Jun 2019.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized markov decision processes. *CoRR*, abs/1705.07798, 2017. URL `http://arxiv.org/abs/1705.07798`.

Esther Derman, Matthieu Geist, and Shie Mannor. Twice regularized mdps and the equivalence between robustness and regularization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 22274–22287. Curran Associates, Inc., 2021.

David Haussler and Manfred Opper. Mutual information, metric entropy and cumulative relative entropy risk. *The Annals of Statistics*, 25(6):2451–2492, 1997. ISSN 00905364.

Yang Zhao, Hao Zhang, and Xiuyuan Hu. Penalizing gradient norm for efficiently improving generalization in deep learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26982–26992. PMLR, 17–23 Jul 2022.

Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the hessian in deep learning: Singularity and beyond. 2017.

Yuichi Yoshida and Takeru Miyato. Spectral norm regularization for improving the generalizability of deep learning, 2017.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: biasing gradient descent into wide valleys*. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, dec 2019.

Florin Gogianu, Tudor Berariu, Mihaela C Rosca, Claudia Clopath, Lucian Busoniu, and Razvan Pascanu. Spectral normalisation for deep reinforcement learning: An optimisation perspective. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3734–3744. PMLR, 18–24 Jul 2021.

Jay K. Satia and Roy E. Lave. Markovian decision processes with uncertain transition probabilities. *Operations Research*, 21(3):728–740, 1973. ISSN 0030364X, 15265463. URL `http://www.jstor.org/stable/169381`.

K. Zhou, J.C. Doyle, and K. Glover. *Robust and Optimal Control*. Feher/Prentice Hall Digital and. Prentice Hall, 1996. ISBN 9780134565675.

J. Doyle. Robust and optimal control. In *Proceedings of 35th IEEE Conference on Decision and Control*, volume 2, pages 1595–1598 vol.2, 1996.

Arnab Nilim and Laurent Ghaoui. Robustness in markov decision problems with uncertain transition matrices. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003.

Onésimo Hernández-Lerma and Jean B. Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer New York, 1 edition, 1996.

Marc Peter Deisenroth and Jan Peters. Solving nonlinear continuous state-action-observation pomdps for mechanical systems with gaussian noise. In *European Workshop on Reinforcement Learning*, 2012.

Anthony R. Cassandra. *Exact and Approximate Algorithms for Partially Observable Markov Decision Processes*. PhD thesis, Brown University, 1998.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, July 2006. ISBN 0471241954.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1376–1401, Paris, France, 03–06 Jul 2015. PMLR.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 297–299. PMLR, 06–09 Jul 2018.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. 2018.

David McAllester. Simplified pac-bayesian margin bounds. In Bernhard Schölkopf and Manfred K. Warmuth, editors, *Learning Theory and Kernel Machines*, pages 203–215, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg. ISBN 978-3-540-45167-9.

Alekh Agarwal, Nan Jiang, and Sham M. Kakade. Reinforcement learning: Theory and algorithms. 2019.

Sham M Kakade. A natural policy gradient. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.

Zebang Shen, Alejandro Ribeiro, Hamed Hassani, Hui Qian, and Chao Mi. Hessian aided policy gradient. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5729–5738. PMLR, 6 2019.

Rotem Mulayoff and Tomer Michaeli. Unique properties of flat minima in deep networks. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. 2021.

Stanislaw Jastrzebski, Devansh Arpit, Oliver Astrand, Giancarlo B Kerg, Huan Wang, Caiming Xiong, Richard Socher, Kyunghyun Cho, and Krzysztof J Geras. Catastrophic fisher explosion: Early phase fisher matrix impacts generalization. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4772–4784. PMLR, 18–24 Jul 2021.

Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal statistics of fisher information in deep neural networks: Mean field approach. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1032–1041. PMLR, 16–18 Apr 2019.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. How many random seeds? statistical power analysis in deep reinforcement learning experiments, 2018.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in Neural Information Processing Systems*, 34, 2021.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, 2017.

Thomas L. Vincent and Jianzu Yu. Control of a chaotic system. *Dynamics and Control*, 1(1):35–52, Mar 1991. ISSN 1573-8450.

M. A. Bucci, O. Semeraro, A. Allauzen, G. Wisniewski, L. Cordier, and L. Mathelin. Control of chaotic systems by deep reinforcement learning. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 475(2231):20190351, 2019. URL `https://royalsocietypublishing.org/doi/abs/10.1098/rspa.2019.0351`.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021.

Erwan Lecarpentier, David Abel, Kavosh Asadi, Yuu Jinnai, Emmanuel Rachelson, and Michael L. Littman. Lipschitz lifelong reinforcement learning, 2021.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, page 267–274, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. ISBN 1558608737.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1889–1897, Lille, France, 07–09 Jul 2015. PMLR.

Shun-ichi Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10(2): 251–276, 02 1998. ISSN 0899-7667. doi: 10.1162/089976698300017746. URL `https://doi.org/10.1162/089976698300017746`.