
The Definitive Guide to Policy Gradients in Deep Reinforcement Learning: Theory, Algorithms and Implementations

Matthias Lehmann
University of Cologne

ABSTRACT

In recent years, various powerful policy gradient algorithms have been proposed in deep reinforcement learning. While all these algorithms build on the Policy Gradient Theorem, the specific design choices differ significantly across algorithms. We provide a holistic overview of on-policy policy gradient algorithms to facilitate the understanding of both their theoretical foundations and their practical implementations. In this overview, we include a detailed proof of the continuous version of the Policy Gradient Theorem, convergence results and a comprehensive discussion of practical algorithms. We compare the most prominent algorithms on continuous control environments and provide insights on the benefits of regularization. All code is available at <https://github.com/Matt0n/PolicyGradientsJax>.

Contents

1	Introduction	1
2	Preliminaries	2
2.1	Notation	2
2.2	Reinforcement Learning	2
2.2.1	Problem Setting	2
2.2.2	Value Functions	3
2.2.3	On-Policy Policy Gradient Methods	4
2.3	Deep Learning	5
3	Theoretical Foundations of Policy Gradients	8
3.1	Policy Gradient Theorem	8
3.2	Value Function Estimation with Baselines	12
3.3	Importance Sampling	14
4	Policy Gradient Algorithms	14
4.1	REINFORCE	15
4.2	A3C	15
4.3	TRPO	16
4.4	PPO	19
4.5	V-MPO	21
4.6	Comparing Design Choices in Policy Gradient Algorithms	23
5	Convergence Results	25
5.1	Literature Overview	25
5.2	Mirror Learning	25
5.2.1	Fundamentals of Mirror Learning	25
5.2.2	Policy Gradient Algorithms as Instances of Mirror Learning	26

5.2.3 Convergence Proof	28
6 Numerical Experiments	33
7 Conclusion	34
Appendices	41
A Hyperparameters	41
B Extended Experiments	41
B.1 Comparison to RL frameworks	41
B.2 Entropy Bonus in A2C	42
B.3 A2C and REINFORCE with Multiple Update Epochs	42
C V-MPO: Derivation Details	42
D Auxiliary Theory	46

1 Introduction

Reinforcement Learning (RL) is a powerful set of methods for an agent to learn how to act optimally in a given environment to maximize some reward signal. In contrast to other methods such as dynamic programming, RL achieves this task of learning an optimal policy, which dictates the optimal behavior, via a trial-and-error process of interacting with the environment [75]. Most early successful applications of RL use value-based methods (e.g., [84, 79, 55]), which estimate the expected future rewards to inform the agent’s decisions. However, these methods only indirectly optimize the true objective of learning an optimal policy [82] and are non-trivial to apply in settings with continuous action spaces [75].

In this work, we discuss policy gradient algorithms [75] as an alternative approach, which aims to directly learn an optimal policy. Policy gradient algorithms are by no means new [7, 78, 87, 88], but this subfield only gained traction in recent years following the emergence of deep RL [55] with the development of various powerful algorithms (e.g., [54, 71, 73]). Deep RL is a subfield of RL, which uses neural networks and other deep learning methods. The increased interest in policy gradient algorithms is due to several appealing properties of this class of algorithms. They can be used natively in continuous action spaces without compromising the applicability to discrete spaces [77]. In contrast to value-based methods, policy gradient algorithms inherently learn stochastic policies, which results in smoother search spaces and partly remedies the exploration problem of having to acquire knowledge about the environment in order to optimize the policy [75, 77]. In some settings, the optimal policy may also be stochastic itself [75]. Lastly, policy gradient methods enable smoother changes in the policy during the learning process, which may result in better convergence properties [77].

Our goal is to present a holistic overview of policy gradient algorithms. In doing so, we limit the scope to on-policy algorithms, which we will define in Section 2. Thus, we exclude some popular algorithms including DDPG [50], TD3 [24] and SAC [28]. See Figure 1 for an overview of RL and the subfields we cover. Our contributions are as follows:

- We give a comprehensive introduction to the theoretical foundations of policy gradient algorithms including a detailed proof of the continuous version of the Policy Gradient Theorem.
- We derive and compare the most prominent policy gradient algorithms and provide high quality pseudocode to facilitate understanding.
- We release competitive implementations of these algorithms, including the, to the best of our knowledge, first publicly available V-MPO implementation displaying performance on par with the results in the original paper.

The remainder of this paper is organized as follows. Section 2 introduces fundamental definitions in RL as well as an overview of deep learning. Section 3 derives the theoretical foundations of policy gradient algorithms with a special focus on proving the Policy Gradient Theorem, based on which we will construct several existing practical algorithms in Section 4. In Section 5, we discuss convergence results from literature. Section 6, presents the results of our numerical experiments comparing the discussed algorithms. Section 7 concludes.

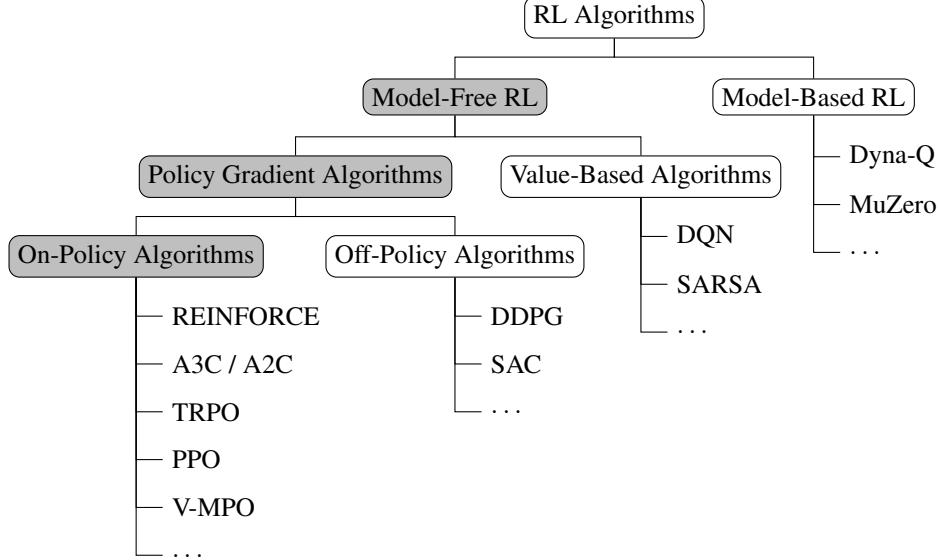


Figure 1: Simplified taxonomy of RL algorithms. Subfields of RL we focus on are highlighted in gray.

2 Preliminaries

In this section, we present prerequisites for subsequent chapters. Specifically, we introduce our notation in Section 2.1 and present overviews of RL in Section 2.2 and of deep learning in Section 2.3. Furthermore, we list several well-known definitions and results from probability theory, measure theory and analysis, which we use in our paper, in Appendix D.

2.1 Notation

We denote the set of natural numbers by \mathbb{N} , natural numbers including zero by \mathbb{N}_0 , real numbers by \mathbb{R} and positive real numbers by \mathbb{R}_+ . We denote d -dimensional real-numbered vector spaces as \mathbb{R}^d . By $\mathcal{P}(\mathcal{A})$, we denote the power set of a set \mathcal{A} . Where possible, we denote random variables with capital letters and their realizations with the corresponding lower case letters. For any probability measure \mathbb{P} , we denote the probability of an event $X = x$ as $\mathbb{P}(X = x)$. Similarly, we write $\mathbb{P}(X = x \mid Y = y)$ for conditional probabilities. When it is clear, which random variable is referred to, we regularly omit it to shorten notation, i.e. $\mathbb{P}(X = x) = \mathbb{P}(x)$. We identify measurable spaces (\mathcal{A}, Σ) just by the set \mathcal{A} as we always use the respective power set $\mathcal{P}(\mathcal{A})$ for discrete sets and the Borel algebra for intervals in \mathbb{R}^d as the respective σ -algebra Σ . We express most Lebesgue integrals w.r.t. the Lebesgue measure λ using Theorem D.9. To simplify notation, we write integrals for measurable functions f on \mathcal{A} as $\int_{a \in \mathcal{A}} f(a) da := \int_{a \in \mathcal{A}} f(a) d\lambda(a)$. We denote that a random variable X follows a probability distribution p by $X \sim p$. For any random variable $X \sim p$, we denote by $\mathbb{E}_{X \sim p}[X]$ and $\text{Var}_{X \sim p}[X]$ its expectation and variance. We denote the set of probability distributions over some measurable space \mathcal{A} as $\Delta(\mathcal{A})$. We write $|\mathcal{A}|$ for the cardinality of a finite set \mathcal{A} or area of a region $\int_{a \in \mathcal{A}} da$. For any variable or function x , we commonly denote approximations to it by \hat{x} .

2.2 Reinforcement Learning

In the following, we formally describe the general problem setting encountered in RL, define fundamental functions and introduce the subfields of RL our work is further concerned with. Sections 2.2.1 and 2.2.2 are based on [75], Chapter 3.

2.2.1 Problem Setting

Each problem instance in RL consists of an agent and an environment with which he interacts to achieve some specific goal. The environment comprises everything external to the agent and can be formalized as a Markov Decision Process (MDP). Let an action space \mathcal{A} be the set of all actions the agent can take and let a state space \mathcal{S} be the set of all possible states, i.e. snapshots of the environment at any given point in time. State and action spaces can be discrete or continuous¹ and we assume both to be compact and measurable. We write an MDP as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \gamma, p_0)$, where $P: \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S} \times \mathbb{R})$ is the environment's transition function, which defines the probability² $P(s', r \mid s, a)$

¹Here, we call a state/action space continuous if it is an interval in \mathbb{R}^d for $d \in \mathbb{N}$.

²Technically, this is the value of the probability density function for continuous distributions. However, we unify terminology by referring to the values of probability density functions as probabilities here and in the following.

of transitioning to a new environment state s' and receiving reward $r \in \mathbb{R}$ when the agent uses action a in state s , $\gamma \in [0, 1]$ is a discount rate and $p_0 \in \Delta(\mathcal{S})$ is a probability distribution over potential starting states. We assume rewards r to be bounded. In the following, our notation assumes state and action spaces to be continuous.

We call sequences of states, actions and rewards $(s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1}, r_{t+2}, \dots, s_{t+k-1}, a_{t+k-1}, r_{t+k}, s_{t+k})$ trajectories. A one-step trajectory, i.e. a tuple $(s_t, a_t, r_{t+1}, s_{t+1})$ is called a transition. In this work, we limit ourselves to episodic settings, where the agent only interacts with the environment for a finite number of at most T steps after which the environment is reset to a starting state. An episode may however be shorter than T if a terminal state is reached. Therefore each episode consists of a trajectory $(s_0, a_0, r_1, s_1, a_1, r_2, \dots, s_{\tilde{T}-1}, a_{\tilde{T}-1}, r_{\tilde{T}}, s_{\tilde{T}})$, with $\tilde{T} \leq T$. Rewards are occasionally omitted from the trajectory notation since they do not influence future states. Correspondingly, we also can compute the alternative transition probabilities $P(s' | s, a) = \int_{r \in \mathbb{R}} P(s', r | s, a) dr$.

The main goal in reinforcement learning is to solve the control problem of learning a policy $\pi: \mathcal{S} \rightarrow \Delta(\mathcal{A})$ to maximize the expected return. The return $G_t := \sum_{k=0}^T \gamma^k r_{t+k+1}$ is the discounted sum of rewards from timestep t onwards. Note that G_t is bounded since rewards are bounded. We denote the probability of taking action a in state s under policy π with $\pi(a | s)$. For a policy π , its stationary state distribution d^π determines the probability of being in a specific state $s \in \mathcal{S}$ at any point in time when following π .

Let Π be the set of all possible policies. RL algorithms $\mathfrak{A}: \Pi \rightarrow \Pi$ for the control problem now iteratively learn policies by interacting with the environment using the current policy to sample transitions, which are then used to update the policy. We will discuss how these updates can look like in Section 2.2.3. A key characteristic of many RL problems is a necessary trade-off between exploration and exploitation in this learning process. The agent has no prior knowledge of the environment and thus needs to explore different transitions in order to learn which states and actions are desirable. As state and action spaces are typically large however, exploiting the already acquired knowledge about the environment is also crucial to guide the search process for an optimal policy to subspaces that hold most promise. A common approach to this exploration problem is to add noise to the policy.

2.2.2 Value Functions

Based on the return, we define the value and action-value functions, which are fundamental in RL. The value function

$$V_\pi(s) := \mathbb{E}_\pi[G_t | S_t = s] \quad (1)$$

gives the expected return from state s onwards when following policy π , which selects all subsequent actions. Thus, the value function states how good it is to be in a specific state s given a policy π . Note that here we follow the general convention to write this just as an expectation over π . However, it should be noted that this expectation integrates over all subsequent states and actions that are obtained by following policy π , i.e. Equation (1) computes the expected return given that all subsequent actions are sampled from π and all rewards and next states are sampled from P . This is implicit in our notation here as well as in further expectations.

Next, we define the action-value function

$$Q_\pi(s, a) := \mathbb{E}_\pi[G_t | S_t = s, A_t = a],$$

which differs from the value function in that the very first action a is provided as an input to the function and not determined by the policy. We observe the following relation between V_π and Q_π :

$$V_\pi(s) = \int_{a \in \mathcal{A}} \pi(a | s) Q_\pi(s, a) da.$$

Further, we call

$$A_\pi(s, a) := Q_\pi(s, a) - V_\pi(s)$$

the advantage function, which determines how good an action a is in state s in relation to other possible actions.

From the definitions of V_π and Q_π we can derive the so-called Bellman equations [9]. Starting from Equation (1), we use the definition of the return G_t , explicitly write out the expectation for the first transition and then apply the definition of V_π again:

$$\begin{aligned} V_\pi(s) &= \mathbb{E}_\pi[G_t | S_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s] \\ &= \int_{a \in \mathcal{A}} \pi(a | s) \int_{s' \in \mathcal{S}} \int_{r \in \mathbb{R}} P(s', r | s, a) \left(r + \gamma \mathbb{E}_\pi[G_{t+1} | S_{t+1} = s'] \right) dr ds' da \\ &= \int_{a \in \mathcal{A}} \pi(a | s) \int_{s' \in \mathcal{S}} \int_{r \in \mathbb{R}} P(s', r | s, a) (r + \gamma V_\pi(s')) dr ds' da \end{aligned}$$

Thus, we find a formulation of the value function, which depends on the value of subsequent states. Collapsing the expectation again yields the form known as the Bellman equation of the value function:

$$V_\pi(s) = \mathbb{E}_\pi [R_{t+1} + \gamma V_\pi(S_{t+1})].$$

Similarly, we can find the Bellman equation for the action-value function:

$$Q_\pi(s, a) = \mathbb{E}_\pi [R_{t+1} + \gamma Q_\pi(S_{t+1}, A_{t+1})]. \quad (2)$$

Now, we can formally define what optimality means in RL. An optimal policy π^* is defined by $V_{\pi^*}(s) \geq V_\pi(s)$ for all states s and policies π , i.e. any optimal policy maximizes the expected return. It can be shown that in every finite MDP, a deterministic optimal policy exists [56]. All optimal policies share the same optimal value function $V^*(s) := \max_{\pi \in \Pi} V_\pi(s)$ and optimal action-value function $Q^*(s, a) := \max_{\pi \in \Pi} Q_\pi(s, a)$ and select actions $a \in \arg \max_{a'} Q^*(s, a')$ for every state. Applying this to Equation (2) yields the Bellman optimality equation

$$\begin{aligned} Q^*(s, a) &= \mathbb{E}_{\pi^*} [R_{t+1} + \gamma Q^*(S_{t+1}, A_{t+1})] \\ &= \mathbb{E}[R_{t+1} + \gamma \max_{a' \in \mathcal{A}} Q^*(S_{t+1}, a')] \end{aligned}$$

We cite the following result without proof from [75] on how to obtain an optimal policy, which we will revisit in Section 5.

Theorem 2.1. (*Generalized Policy Iteration*) Let π_{old} be the current policy. Then, Generalized Policy Iteration updates its policy by

$$\pi_{new} \in \arg \max_{\pi \in \Pi} \mathbb{E}_{A \sim \pi} [Q_{\pi_{old}}(s, A)]$$

for all $s \in \mathcal{S}$. Let $(\pi_n)_{n=0}^\infty$ be a sequence of policies obtained through Generalized Policy Iteration. Then, this sequence converges to an optimal policy, i.e.

$$\lim_{n \rightarrow \infty} \pi_n = \pi^*$$

and

$$\lim_{n \rightarrow \infty} Q_{\pi_n} = Q^*.$$

2.2.3 On-Policy Policy Gradient Methods

Finally, we will delineate the subfields of RL on which our work focuses. In this context, we will successively introduce function approximation, policy gradient methods and the on-policy paradigm.

RL algorithms are mostly concerned with learning functions such as π , V_π or Q_π . Early reinforcement methods learn exact representations of these by maintaining lookup tables with entries for each possible function input [75]. While this approach yields theoretical convergence guarantees [56], it is practically very limited. Similar states are treated independently such that learnings do not generalize from one state to others while specific states are only rarely visited in large state spaces [75]. Moreover, this approach is not applicable to continuous spaces. Function approximation remedies these shortcomings by parameterizing the function to be learned. Let $f_\theta(x)$ be this learnable function, where θ are the function's parameters, which are adjusted over the course of learning, and x are the functions inputs such as states and actions or representations thereof. By choosing f_θ to be continuous in its inputs, we can ensure that f_θ generalizes across its inputs when we fit it to sampled transitions [75]. f_θ can be as simple as a linear mapping, i.e. $f_\theta(x) = \theta^T x$, however recent works mostly use neural networks as function approximators (e.g., [55, 72]). The field using neural networks as function approximators is coined deep RL [55]. For the remainder of this paper, you can consider any learned function to be a neural network unless explicitly stated otherwise, although all our statements apply to any differentiable function approximators. We will introduce deep learning and neural networks in detail in Section 2.3.

Policy gradient methods pose an alternative to value-based methods in RL. Most early successes in RL use value-based methods such as Q-Learning [84] or SARSA [67], that aim at learning a sequence of value functions converging to the optimal value function, from which an optimal policy can then be inferred. In contrast, policy-based RL, which we focus on in this work, directly learns a parameterized policy π_θ . The main idea in this learning process is to increase the probability of those actions that lead to higher returns until we reach an (approximately) optimal policy [75]. While this optimization problem can be approached in several ways, gradient-based methods are most commonly used [82]. Following [77], we define policy gradient methods as follows.

Definition 2.2. (*Policy Gradient Algorithm*) Let $\pi_\theta: \mathcal{S} \rightarrow \Delta(\mathcal{A})$ be a fully differentiable function with learnable parameters $\theta \in \mathbb{R}^d$ mapping states to a probability distribution over actions. Let $J: \mathbb{R}^d \rightarrow \mathbb{R}$ be some performance

measure of the parameters. We call any learning algorithm a policy gradient algorithm if it learns its policy π_θ by updating θ via gradient ascent (or descent) on J , i.e. its updates have the general form

$$\theta_{\text{new}} \leftarrow \theta + \alpha \nabla_\theta J(\theta), \quad (3)$$

where $\alpha \in \mathbb{R}$ is a step size parameter of the algorithm.

In policy-based RL, two distinct ways exist to have the policy output a probability distribution over actions, from which actions can be sampled [75]. For discrete action spaces, we construct a discrete distribution over the action space by normalizing the policies' raw outputs via a softmax function [26]. In this case, we have

$$\pi(a | s) = \frac{\exp(\pi_\theta(a | s))}{\sum_{a' \in \mathcal{A}} \exp(\pi_\theta(a' | s))}.$$

For continuous action spaces, we let π_θ output the mean μ and standard deviation σ of a Gaussian distribution, i.e. $\pi_\theta(s) = (\mu_\theta(s), \sigma_\theta(s))$ such that

$$\pi(a | s) = \frac{1}{\sigma_\theta(s)\sqrt{2\pi}} \exp\left(-\frac{(a - \mu_\theta(s))^2}{2\sigma_\theta(s)^2}\right).$$

This parameterization of a Gaussian distribution for the policy was first introduced by [87, 88]. As action spaces are commonly bounded, the actions sampled from such a Gaussian are typically transformed to be within these bounds either by clipping or by applying a squashing distribution [5]. Further, we highlight that the policies learned by policy gradient methods in both the discrete and the continuous case are generally stochastic. This stands in contrast to value-based methods which generally learn deterministic policies [77]. Policy gradient methods are the core focus of this paper and will be discussed in-depth in subsequent sections.

Lastly, we delineate on-policy from off-policy algorithms. In RL, we distinguish between behavior and target policies [75]. A behavior policy is a policy which generates the data in form of trajectories from which we want to learn, i.e. this is the policy from which we sample actions when interacting with the environment. Conversely, the target policy is the policy which we want to learn about to evaluate how good it is in the given environment and improve it. Algorithms where behavior and target policy are not identical, e.g. Q-Learning [84] or DQN [55], are referred to as off-policy algorithms. In this work, we only discuss on-policy algorithms, where behavior and target policy are identical. Hence, when speaking of policy gradient algorithms in the following, we always implicitly mean on-policy policy gradient algorithms if not mentioned otherwise.

2.3 Deep Learning

In this section, we introduce deep learning as a subfield of machine learning since its methods are commonly used in policy gradient algorithms. In recent years, deep learning has emerged as the premier machine learning method in various fields, enabling state-of-the-art performance in domains such as computer vision (e.g., [44, 29, 22]) and natural language processing (e.g., [83, 14]). Following [47] and [26], we define deep learning as a set of techniques to solve prediction tasks by learning multiple levels of representations from raw data using a composition of simple non-linear functions. This composition of functions, that we will describe in detail later, is referred to as (deep) neural network. Deep learning stands in contrast to conventional machine learning techniques like logistic regressions, which typically require hand-engineered representations as inputs to be effective [47]. In the following, we introduce the general problem setting of deep learning using the notation of [10], formalize neural networks and describe how they are trained.

Consider measurable spaces \mathcal{X} and \mathcal{Y} . $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ is the data space with each element $z = (x, y) \in \mathcal{Z}$ being a tuple of input features $x \in \mathcal{X}$ and a label $y \in \mathcal{Y}$. Let $\mathcal{M}(\mathcal{X}, \mathcal{Y})$ be the set of measurable functions from \mathcal{X} to \mathcal{Y} . The problems we encounter in deep learning are prediction tasks. Thus, the goal is to learn a mapping $f \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$ from inputs to labels by minimizing some loss function \mathcal{L} over training data $S = \{z^{(1)}, \dots, z^{(m)}\}$ such that it generalizes to unseen data $z \in \mathcal{Z}$. To learn the function f , we first select a hypothesis set $\mathcal{F} \subset \mathcal{M}(\mathcal{X}, \mathcal{Y})$. Deep learning then provides learning algorithms $\mathfrak{A}: \mathcal{Z} \rightarrow \mathcal{F}$ that use training data S to learn the desired function $f = \mathfrak{A}(S)$. Before we discuss this learning process, we will first further characterize the mapping to be learned.

In deep learning, functions in the hypothesis set \mathcal{F} represent instances of neural networks. Note that here we limit ourselves to feedforward networks, also called multilayer perceptrons (MLP), and will not discuss transformers [83], recurrent (RNN) [32] or convolutional neural networks (CNN)[44].

Definition 2.3. (*Feedforward Neural Network*) A feedforward neural network $f: \mathcal{X} \rightarrow \mathcal{Y}$ is a composition of functions

$$f = f^{(n+1)} \circ \dots \circ f^{(1)}$$

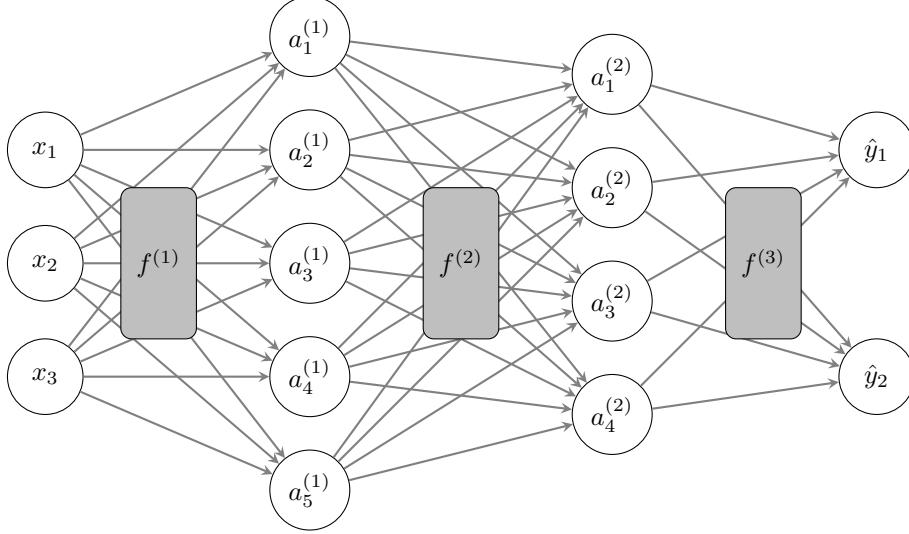


Figure 2: A neural network with hidden layers of sizes 5 and 4 as a directed graph.

that is differentiable almost everywhere. We refer to $f^{(i)}$, $i = 1, \dots, n$ as *hidden layers*, whereas $f^{(n+1)}$ is the non-hidden output layer. Consequently, n denotes the number of hidden layers in the network. Each hidden layer is characterized by a layer width N_i . Let N_0 and N_{n+1} further be the size of the input and output vectors respectively. Then, we can write each layer as

$$f^{(i)}(x) = g\left(W^{(i)}x + b^{(i)}\right),$$

where x is the output of the previous layer or the network's inputs for $i = 1$, $W^{(i)} \in \mathbb{R}^{N_i \times N_{i-1}}$ and $b^{(i)} \in \mathbb{R}^{N_i}$ are the layer's weight matrix and bias vector respectively and $g: \mathbb{R} \rightarrow \mathbb{R}$ is a differentiable activation function introducing non-linearity. g is applied element-wise.

An MLP can be characterized by its architecture $a = ((N_i)_{i=0}^{n+1}, g)$, consisting of layer sizes and the activation function to be used. The number of layers $n + 1$ is also referred to as the depth of the network. The Universal Approximation Theorem [17, 35] underpins the expressivity of neural networks: a two-layer network can already approximate any measurable function arbitrarily well under weak conditions on the activation function. Technically, each layer can feature a different activation function albeit this is uncommon. The activation function of the output layer is not defined by the architecture a but is derived from the prediction task. The standard choice for activation functions in the hidden layers is a rectified linear unit (ReLU)³ [57], due to typically fast learning [25]. See [48] for an overview of other commonly used activation functions. The output layer typically uses no activation function for regression tasks and sigmoid or softmax functions for classification tasks. Each element of a layer $f^{(i)}$ is called a neuron. The outputs $a^{(i)} = (f^{(i)} \circ \dots \circ f^{(1)})(x)$ of any layer are the learned representations of the inputs x . We denote the outputs $f(x)$, i.e. the predictions, of the neural network with \hat{y} . Figure 2 depicts a neural network as an acyclic directed graph.

Selecting a hypothesis set \mathcal{F} is done implicitly by specifying an architecture a . Hence, we denote the hypothesis set for architecture a , whose elements are all MLPs with that architecture, by \mathcal{F}_a . The MLPs in \mathcal{F}_a therefore differ only in their weights and biases. We call these the (learnable) parameters of the network and typically collect them in a flattened parameter vector $\theta \in \mathbb{R}^d$. We denote an MLP with parameters θ as f_θ .

Given a hypothesis set \mathcal{F}_a , we now aim to learn a neural network $f_\theta \in \mathcal{F}_a$, i.e. to learn parameters θ , such that we reduce the expected loss or risk,

$$\mathcal{R}(f) := \mathbb{E}_{Z \sim \mathbb{P}_Z} [\mathcal{L}(f, Z)] = \int_{z \in \mathcal{Z}} \mathcal{L}(f, z) d\mathbb{P}_Z(z),$$

over the data distribution \mathbb{P}_Z for some appropriately chosen differentiable loss function $\mathcal{L}: \mathcal{F} \times \mathcal{Z} \rightarrow \mathbb{R}$ [10, 26]. Here \mathbb{P}_Z is the image measure of Z on \mathcal{Z} , from which the training data $S = \{z^{(1)}, \dots, z^{(m)}\}$ and unknown out-of-sample data z is drawn. We generally assume that training $z^{(1)}, \dots, z^{(m)}$ and out-of-sample data z are realizations of i.i.d. random variables $Z^{(1)}, \dots, Z^{(m)}, Z \sim \mathbb{P}_Z$ [10]. For a given MLP $f_\theta = \mathcal{A}(S)$ trained on S , the risk becomes

³Note that the ReLU function is not differentiable at 0. In practice, this is circumvented by using its sub-derivatives.

$\mathcal{R}(f_\theta) = \mathbb{E}_{Z \sim \mathbb{P}_Z} [\mathcal{L}(f_\theta, Z) | S]$. In practice, noisy data results in a positive lower bound on risk, i.e. an irreducible error [10]. Common loss functions are binary cross-entropy loss,

$$\mathcal{L}(f, (x, y)) = -(y \cdot \ln(f(x)) + (1 - y) \cdot \ln(1 - f(x))),$$

for (binary) classification and mean squared error (MSE),

$$\mathcal{L}(f, (x, y)) = (y - f(x))^2,$$

for regression tasks. Sometimes, loss functions are augmented by regularization terms $\Omega(\theta)$ such as an L2-penalty of the parameters, i.e. $\beta \|\theta\|_2^2$ with $\beta \in \mathbb{R}$ [26].

The data distribution \mathbb{P}_Z is generally unknown. Hence, we replace it by an empirical distribution based on the sampled training data S and use empirical risk minimization (ERM) as the learning algorithm to minimize it [26, 10].

Definition 2.4. (*Empirical Risk*). Given training data $S = \{z^{(1)}, \dots, z^{(m)}\}$ and a function $f_\theta \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$, the empirical risk is defined by

$$\hat{\mathcal{R}}_S(f_\theta) := \frac{1}{m} \sum_{i=1}^m \mathcal{L}(f_\theta, z^{(i)}). \quad (4)$$

Definition 2.5. (*ERM learning algorithm*). Given hypothesis set \mathcal{F}_a and training data S , an empirical risk minimization algorithm $\mathfrak{A}^{\text{erm}}$ terminates with an (approximate⁴) minimizer $\hat{f}_S \in \mathcal{F}_a$ of empirical risk:

$$\mathfrak{A}^{\text{erm}}(S) = \hat{f}_S \in \arg \min_{f \in \mathcal{F}_a} \hat{\mathcal{R}}_S(f).$$

We approximately minimize empirical risk typically via gradient-based methods due to efficient computation of point-wise derivatives via the backpropagation algorithm [66, 40]. Backpropagation means the practical application of the chain rule to neural networks. The gradient of the objective function \mathcal{L} with respects to the i -th layer's inputs $a^{(i-1)}$ can be computed by working backwards from the gradient with respects to the layer's outputs $a^{(i)}$ as $\nabla_{a^{(i-1)}} \mathcal{L} = \sum_j (\nabla_{a^{(i-1)}} a_j^{(i)}) \frac{\partial \mathcal{L}}{\partial a_j^{(i)}}$. From these gradients, the gradients with respects to the weights and biases in each layer can be calculated similarly. Due to this flow of information from the objective function to each of the layers, the optimization of MLPs is also referred to as backwards pass, in contrast to the forward pass of calculating $\hat{y} = f(x)$. The full backpropagation algorithm for MLPs is formulated in Algorithm 1.

Algorithm 1 Backpropagation, pseudocode taken from [26]

```

Require: labels  $y$ , regularizer  $\Omega(\theta)$ , network outputs  $\hat{y}$ , activated and unactivated layer outputs  $a^{(k)}$  and  $h^{(k)}$  for  $k = 1, \dots, n$ , activation function  $g$ , loss  $\mathcal{L}$ 
 $\delta \leftarrow \nabla_{\hat{y}} \mathcal{L}$ 
for  $k = n, \dots, 1$  do
     $\delta \leftarrow \nabla_{h^{(k)}} \mathcal{L} = \delta \odot g'(h^{(k)})$  ▷ hadamard product if  $g$  is element-wise
     $\nabla_{b^{(k)}} \mathcal{L} \leftarrow \delta + \nabla_{b^{(k)}} \Omega(\theta)$ 
     $\nabla_{W^{(k)}} \mathcal{L} \leftarrow \delta h^{(k-1)\top} + \nabla_{W^{(k)}} \Omega(\theta)$ 
     $\delta \leftarrow \nabla_{a^{(k-1)}} \mathcal{L} = W^{(k)\top} \delta$ 
end for

```

The gradients computed via backpropagation are used to update the parameters in each layer using gradient descent. However, due to the prohibitive computational costs of evaluating the expectation in Equation (4), computing gradients only on a subset of the training data is generally preferred and typically also results in faster convergence [26]. At each iteration, a batch S' of data with size $m' \leq m$ (typically $m' \ll m$) is randomly sampled from the training data to conduct the update [10]

$$\Theta^{(k)} := \Theta^{(k-1)} - \alpha_k \frac{1}{m'} \sum_{z \in S'} \nabla_{\theta} \mathcal{L}\left(f_{\Theta^{(k-1)}}, z\right). \quad (5)$$

Here, Θ is a random variable whose realizations are neural network parameters θ . α_k is the step size or learning rate on the k -th optimization step. The learning rate is commonly decayed over the training process to help convergence [26].

⁴In practice, the empirical risk is generally highly non-convex prohibiting guaranteed convergence to a global minimum [10].

The procedure using updates as in Equation (5) is known as stochastic (minibatch) gradient descent (SGD)⁵ [26] and dates back to [63, 41]. Using SGD has the additional benefit of introducing random fluctuations which enable escaping saddle points [10]. SGD in its general form is depicted in Algorithm 2, where in our context $r(\theta) = \hat{\mathcal{R}}_S(f_\theta)$. The neural network parameters θ are set to be the realization of the final $\Theta^{(K)}$ or a convex combination of $(\Theta^{(k)})_{k=1}^K$ [10].

Algorithm 2 Stochastic Gradient Descent, pseudocode from [10]

Require: Differentiable function $r: \mathbb{R}^d \rightarrow \mathbb{R}$, step sizes $\alpha_k \in (0, \infty)$, $k = 1, \dots, K$, \mathbb{R}^d -valued random variable $\Theta^{(0)}$

for $k = 1, \dots, K$ **do**

Let $D^{(k)}$ be a random variable such that $\mathbb{E}[D^{(k)} | \Theta^{(k-1)}] = \nabla r(\Theta^{(k-1)})$

$\Theta^{(k)} \leftarrow \Theta^{(k-1)} - \alpha_k D^{(k)}$

end for

Despite the stochasticity of SGD and highly non-convex loss landscapes, SGD's convergence can be guaranteed in some regimes [10], and it exhibits strong performance in practice [26]. Hence, SGD and its variants are the default choice to optimize neural networks. The most prominently used variant is Adam [42], which uses momentum [59] and an adaptive scaling of gradients to stabilize learning. Nonetheless, the initialization of θ is also important for convergence. Biases are commonly initialized to 0 whereas weights are randomly initialized close to 0 using various strategies [26]. Finally, note that regardless of the non-convexity of the loss landscapes, local minima are not considered problematic if the neural networks are large enough [18, 15].

In practice, the training of neural networks is an iterative process. We alternate between choosing the network architecture a as well as further hyperparameters of the learning algorithm such as the learning rates α , and approximately minimizing the empirical risk for this set of hyperparameters. This is generally a trial-and-error process to find a suitable set of hyperparameters to maximize generalization performance, i.e. to minimize risk. To approximate risk, the trained models are typically evaluated by the empirical risk on a held-out test data set, which was not seen during training [26]. The achieved empirical risk can be decomposed into a generalization error, an optimization error, an approximation error and the irreducible error [10]. The generalization error is the difference between empirical and actual risk stemming from the random sampling of training data, which may not be representative of the actual data distribution \mathbb{P}_Z . The optimization error is the result of potentially not finding a global minimum during the learning process. The approximation error is the difference between the minimum achievable risk over functions in \mathcal{F}_a and over all $f \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$.

3 Theoretical Foundations of Policy Gradients

Having introduced the fundamentals of deep RL, we can now discuss policy gradient algorithms in detail. In this section, we derive their theoretical foundations. Our main focus is going to be the Policy Gradient Theorem, on which all policy gradient algorithms build. This theorem will be discussed in Section 3.1. Furthermore, Sections 3.2 and 3.3 introduce the theoretical justifications for additional methods that are frequently used in policy gradient algorithms.

3.1 Policy Gradient Theorem

Given an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \gamma, p_0)$, consider a parameterized policy π_θ , which is differentiable almost everywhere, and the following objective function J for maximizing the expected episodic return:

$$\begin{aligned} J(\theta) &= \mathbb{E}_{S_0 \sim p_0, \pi_\theta} [G_0] \\ &= \mathbb{E}_{S_0 \sim p_0} \left[\mathbb{E}_{\pi_\theta} [G_t | S_t = S_0] \right] \\ &= \mathbb{E}_{S_0 \sim p_0} [V_{\pi_\theta}(S_0)] \end{aligned}$$

The idea of policy gradient algorithms is to maximize $J(\theta)$ over the parameters θ by performing gradient ascent [75]. Hence, we require the gradients $\nabla_\theta J(\theta)$, however it is a priori not obvious how the right-hand side $\mathbb{E}_{S_0 \sim p_0, \pi_\theta} [G_0]$ depends on θ as changes in the policy π also affect the state distribution d^π . The Policy Gradient Theorem [76, 52] yields an analytic form of $\nabla_\theta J(\theta)$ from which we can sample gradients that does not involve the derivative of d^π . Here, we focus on the undiscounted case, i.e. $\gamma = 1$. Note that any discounted problem instance can be reduced to the undiscounted case by letting the reward function absorb the discount factor [70].

⁵Sometimes SGD refers to updates which only involve a single data point. We however follow the nowadays common terminology of calling any sample-based gradient descent stochastic.

Theorem 3.1. (*Policy Gradient Theorem*) For a given MDP, let π_θ be differentiable w.r.t. θ and $\nabla_\theta \pi_\theta$ be bounded, let Q_{π_θ} be differentiable w.r.t. θ and $\nabla_\theta Q_{\pi_\theta}$ be bounded for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Then, there exists a constant η such that

$$\nabla_\theta J(\theta) = \eta \mathbb{E}_{S \sim d^{\pi_\theta}, A \sim \pi_\theta} [Q_{\pi_\theta}(S, A) \nabla_\theta \ln \pi_\theta(A | S)]. \quad (6)$$

Proof. We largely follow the proof by [75] albeit in a more detailed form and extended to continuous state and action spaces. To enhance readability, we omit subscripts θ for the policy π and all gradients ∇ but both always depend on the parameters θ .

Starting from the definition of the objective function, we explicitly write out the expectation over starting states, use the relationship between value and action-value function, $V_\pi(s) = \int_{a \in \mathcal{A}} \pi(a | s) Q_\pi(s, a) da$, and differentiate by parts.

$$\begin{aligned} \nabla J(\theta) &= \nabla \mathbb{E}_{S \sim p_0} [V_\pi(S)] \\ &= \nabla \int_{s \in \mathcal{S}} p_0(s) V_\pi(s) ds \\ &= \nabla \int_{s \in \mathcal{S}} p_0(s) \int_{a \in \mathcal{A}} \pi(a | s) Q_\pi(s, a) da ds \\ &= \int_{s \in \mathcal{S}} p_0(s) \left(\int_{a \in \mathcal{A}} (\nabla \pi(a | s)) Q_\pi(s, a) da + \int_{a \in \mathcal{A}} \pi(a | s) \nabla Q_\pi(s, a) da \right) ds. \end{aligned} \quad (7)$$

Note that in the last step via used the Leibniz integral rule (Theorem D.10) to swap the order of integration and differentiation prior to applying the product rule. The conditions for Leibniz are satisfied since $\pi(\cdot | s) Q_\pi(s, \cdot)$ is integrable for any $s \in \mathcal{S}$ and its partial derivatives exist and are bounded for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$ since π and Q_π are bounded and ∇Q_π and $\nabla \pi$ exist and are bounded by assumption.

Now, consider the recursive formulation of the action-value function

$$Q_\pi(s, a) = \int_{s' \in \mathcal{S}} \int_{r \in \mathbb{R}} P(s', r | s, a) (r + V_\pi(s')) dr ds'.$$

Due to the identity $\int_{r \in \mathbb{R}} P(s', r | s, a) dr = P(s' | s, a)$ and since realized rewards r and environment transitions for a given action no longer depend on the policy, we can reformulate the gradients of Q_π w.r.t. θ , again using the Leibniz integral rule.

$$\begin{aligned} \nabla Q_\pi(s, a) &= \nabla \int_{s' \in \mathcal{S}} \int_{r \in \mathbb{R}} P(s', r | s, a) (r + V_\pi(s')) dr ds' \\ &= \int_{s' \in \mathcal{S}} \int_{r \in \mathbb{R}} P(s', r | s, a) \nabla(r + V_\pi(s')) dr ds' \\ &= \int_{s' \in \mathcal{S}} \int_{r \in \mathbb{R}} P(s', r | s, a) \nabla V_\pi(s') dr ds' \\ &= \int_{s' \in \mathcal{S}} \left(\int_{r \in \mathbb{R}} P(s', r | s, a) dr \right) \nabla V_\pi(s') ds' \\ &= \int_{s' \in \mathcal{S}} P(s' | s, a) \nabla V_\pi(s') ds'. \end{aligned} \quad (8)$$

Further, note that for all $s \in \mathcal{S}$

$$\begin{aligned} \nabla V_\pi(s) &= \nabla \int_{a \in \mathcal{A}} \pi(a | s) Q_\pi(s, a) da \\ &= \int_{a \in \mathcal{A}} (\nabla \pi(a | s)) Q_\pi(s, a) da + \int_{a \in \mathcal{A}} \pi(a | s) \nabla Q_\pi(s, a) da, \end{aligned} \quad (9)$$

which is equivalent to the inner expression in Equation (7). By using (8) and (9), we can transform (7) into a recursive form, which we are then going to unroll subsequently to yield an explicit form. In the following, we simply notation by defining

$$\phi(s) := \int_{a \in \mathcal{A}} (\nabla \pi(a | s)) Q_\pi(s, a) da. \quad (10)$$

Applying (10) and (8) to (7) in order and rearranging the integrals gives

$$\begin{aligned} \nabla J(\theta) &= \int_{s \in \mathcal{S}} p_0(s) \left(\int_{a \in \mathcal{A}} (\nabla \pi(a | s)) Q_\pi(s, a) da + \int_{a \in \mathcal{A}} \pi(a | s) \nabla Q_\pi(s, a) da \right) ds \\ &= \int_{s \in \mathcal{S}} p_0(s) \left(\phi(s) + \int_{a \in \mathcal{A}} \pi(a | s) \nabla Q_\pi(s, a) da \right) ds \\ &= \int_{s \in \mathcal{S}} p_0(s) \left(\phi(s) + \int_{a \in \mathcal{A}} \pi(a | s) \int_{s' \in \mathcal{S}} P(s' | s, a) \nabla V_\pi(s') ds' da \right) ds \\ &= \int_{s \in \mathcal{S}} p_0(s) \left(\phi(s) + \int_{s' \in \mathcal{S}} \int_{a \in \mathcal{A}} \pi(a | s) P(s' | s, a) da \nabla V_\pi(s') ds' \right) ds \end{aligned} \quad (11)$$

In the final step, we switched the order of integration using Fubini's Theorem (Theorem D.11), which is applicable since ∇V_π is bounded and $\pi(\cdot | s)P(\cdot | s, \cdot)$ is a probability measure on $\mathcal{S} \times \mathcal{A}$ such that $|\pi(\cdot | s)P(\cdot | s, \cdot)|\nabla V_\pi|$ is integrable over the product space $\mathcal{S} \times \mathcal{A}$. To unroll Equation (11) across time, we introduce notation for multi-step transition probabilities. Let $\rho_\pi(s \rightarrow s', k)$ be the probability of transitioning from state s to s' after k steps under policy π . We have that

$$\rho_\pi(s \rightarrow s', 0) := \begin{cases} 1 & \text{if } s = s', \\ 0 & \text{else} \end{cases}$$

and $\rho_\pi(s \rightarrow s', 1) := \int_{a \in \mathcal{A}} \pi(a | s) P(s' | s, a) da$. Now, we can recursively write

$$\rho_\pi(s \rightarrow s'', k+1) = \int_{s' \in \mathcal{S}} \rho_\pi(s \rightarrow s', k) \rho_\pi(s' \rightarrow s'', 1) ds'.$$

Using this notation, iteratively substituting in (8) and (9) and applying Fubini, we can unroll (11):

$$\begin{aligned}
 \nabla J(\theta) &= \int_{s \in \mathcal{S}} p_0(s) \left(\phi(s) + \int_{s' \in \mathcal{S}} \int_{a \in \mathcal{A}} \pi(a | s) P(s' | s, a) da \nabla V_\pi(s') ds' \right) ds \\
 &= \int_{s \in \mathcal{S}} p_0(s) \left(\phi(s) + \int_{s' \in \mathcal{S}} \rho_\pi(s \rightarrow s', 1) \nabla V_\pi(s') ds' \right) ds \\
 &= \int_{s \in \mathcal{S}} p_0(s) \left(\phi(s) + \int_{s' \in \mathcal{S}} \rho_\pi(s \rightarrow s', 1) \left(\phi(s') + \int_{a \in \mathcal{A}} \pi(a | s') \nabla Q_\pi(s', a) da \right) ds' \right) ds \\
 &= \int_{s \in \mathcal{S}} p_0(s) \left(\phi(s) + \int_{s' \in \mathcal{S}} \rho_\pi(s \rightarrow s', 1) \left(\phi(s') + \int_{s'' \in \mathcal{S}} \rho_\pi(s' \rightarrow s'', 1) \nabla V_\pi(s'') ds'' \right) ds' \right) ds \\
 &= \int_{s \in \mathcal{S}} p_0(s) \left(\phi(s) + \int_{s' \in \mathcal{S}} \rho_\pi(s \rightarrow s', 1) \phi(s') ds' \right. \\
 &\quad \left. + \int_{s'' \in \mathcal{S}} \left(\int_{s' \in \mathcal{S}} \rho_\pi(s \rightarrow s', 1) \rho_\pi(s' \rightarrow s'', 1) ds' \right) \nabla V_\pi(s'') ds'' \right) ds \\
 &= \int_{s \in \mathcal{S}} p_0(s) \left(\phi(s) + \int_{s' \in \mathcal{S}} \rho_\pi(s \rightarrow s', 1) \phi(s') ds' + \int_{s'' \in \mathcal{S}} \rho_\pi(s \rightarrow s'', 2) \nabla V_\pi(s'') ds'' \right) ds \\
 &= \int_{s \in \mathcal{S}} p_0(s) \left(\rho_\pi(s \rightarrow s, 0) \phi(s) + \int_{s' \in \mathcal{S}} \rho_\pi(s \rightarrow s', 1) \phi(s') ds' \right. \\
 &\quad \left. + \int_{s'' \in \mathcal{S}} \rho_\pi(s \rightarrow s'', 2) \phi(s'') ds'' + \int_{s''' \in \mathcal{S}} \rho_\pi(s \rightarrow s''', 3) \nabla V_\pi(s''') ds''' \right) ds \\
 &\quad \vdots \\
 &= \int_{s \in \mathcal{S}} p_0(s) \int_{s' \in \mathcal{S}} \sum_{t=0}^T \rho_\pi(s \rightarrow s', t) \phi(s') ds' ds
 \end{aligned}$$

We set $\eta_s(s') := \sum_{t=0}^T \rho_\pi(s \rightarrow s', t)$, rearrange the integrals and multiply by 1 to obtain

$$\begin{aligned}
 \nabla_\theta J(\theta) &= \int_{s \in \mathcal{S}} p_0(s) \int_{s' \in \mathcal{S}} \sum_{t=0}^T \rho_\pi(s \rightarrow s', t) \phi(s') ds' ds \\
 &= \int_{s' \in \mathcal{S}} \int_{s \in \mathcal{S}} p_0(s) \eta_s(s') \phi(s') ds ds' \\
 &= \frac{\int_{s'' \in \mathcal{S}} \int_{s \in \mathcal{S}} p_0(s) \eta_s(s'') ds ds''}{\int_{s'' \in \mathcal{S}} \int_{s \in \mathcal{S}} p_0(s) \eta_s(s'') ds ds''} \int_{s' \in \mathcal{S}} \int_{s \in \mathcal{S}} p_0(s) \eta_s(s') ds \phi(s') ds' \\
 &= \int_{s'' \in \mathcal{S}} \int_{s \in \mathcal{S}} p_0(s) \eta_s(s'') ds ds'' \int_{s' \in \mathcal{S}} \frac{\int_{s \in \mathcal{S}} p_0(s) \eta_s(s') ds}{\int_{s'' \in \mathcal{S}} \int_{s \in \mathcal{S}} p_0(s) \eta_s(s'') ds ds''} \phi(s') ds' \\
 &= \int_{s \in \mathcal{S}} p_0(s) \int_{s'' \in \mathcal{S}} \eta_s(s'') ds'' ds \int_{s' \in \mathcal{S}} d^\pi(s') \phi(s') ds'.
 \end{aligned} \tag{12}$$

In the final step, we used the identity

$$d^\pi(s') = \frac{\int_{s \in \mathcal{S}} p_0(s) \eta_s(s') ds}{\int_{s'' \in \mathcal{S}} \int_{s \in \mathcal{S}} p_0(s) \eta_s(s'') ds ds''},$$

which can be seen as $\eta_s(s')$ is the accumulate sum over probabilities of reaching s' after any number of steps for a given starting state. Integrating over the starting state distribution and normalizing hence yields the probability of visiting state s' and thereby the stationary distribution d^π over states under the current policy.

Finally, we can derive the canonical form of the Policy Gradient Theorem from (12) by using the definition of $\phi(s)$, setting

$$\eta := \int_{s \in \mathcal{S}} p_0(s) \int_{s'' \in \mathcal{S}} \eta_s(s'') ds'' ds$$

and multiplying with 1:

$$\begin{aligned} \nabla J(\theta) &= \int_{s \in \mathcal{S}} p_0(s) \int_{s'' \in \mathcal{S}} \eta_s(s'') ds'' ds \int_{s' \in \mathcal{S}} d^\pi(s') \phi(s') ds' \\ &= \eta \int_{s' \in \mathcal{S}} d^\pi(s') \int_{a \in \mathcal{A}} (\nabla \pi(a | s')) Q_\pi(s', a) da ds' \\ &= \eta \int_{s' \in \mathcal{S}} d^\pi(s') \int_{a \in \mathcal{A}} \pi(a | s') \frac{\nabla \pi(a | s')}{\pi(a | s')} Q_\pi(s', a) da ds' \\ &= \eta \int_{s' \in \mathcal{S}} d^\pi(s') \int_{a \in \mathcal{A}} \pi(a | s') (\nabla \ln \pi(a | s')) Q_\pi(s', a) da ds' \\ &= \eta \mathbb{E}_{S \sim d^\pi} \left[\mathbb{E}_{A \sim \pi} [Q_\pi(S, A) \nabla \ln \pi(A | S)] \right]. \end{aligned}$$

□

The Policy Gradient Theorem provides us with an explicit form of the policy gradients from which we can sample gradients. This allows the use of gradient-based optimization to directly optimize the policy using the methods presented in Section 2.3. Thus, the theorem serves as the foundation for the policy gradient algorithms which we will discuss in Section 4.

We conclude this section with some further remarks on Equation (6). First, we note that for any starting state $s \in \mathcal{S}$, we have that

$$\begin{aligned} \eta &= \int_{s \in \mathcal{S}} p_0(s) \int_{s' \in \mathcal{S}} \eta_s(s') ds' ds = \int_{s \in \mathcal{S}} p_0(s) \int_{s' \in \mathcal{S}} \sum_{t=0}^T \rho_\pi(s \rightarrow s', t) ds' ds \\ &= \mathbb{E}_{S \sim p_0} \left[\sum_{t=0}^T \int_{s' \in \mathcal{S}} \rho_\pi(S \rightarrow s', t) ds' \right], \end{aligned}$$

which is the average episode length⁶ under policy π [75]. Second, the use of gradient-based methods makes it sufficient to sample gradients which are only proportional to the actual gradients since any constant of proportionality can be absorbed by the learning rate parameter of the optimization algorithms. Hence, η is commonly omitted [75], i.e.

$$\nabla_\theta J(\theta) \propto \mathbb{E}_{S \sim d^{\pi_\theta}, A \sim \pi_\theta} [Q_{\pi_\theta}(S, A) \nabla_\theta \ln \pi_\theta(A | S)]. \quad (13)$$

We observe that all terms on the right hand side are known or can be estimated via sampling.

3.2 Value Function Estimation with Baselines

In practice, the resulting estimates of the policy gradients can become very noisy when sampling from Equation (13). Therefore, a main practical challenge of policy gradient algorithms is to introduce measures to reduce the variance of the gradients while keeping the bias low [77]. In this context, a well-known and widely used technique is to use a baseline [88] when sampling an estimate of the action-value function Q_π [27]. In this section, we show that using an appropriately chosen baseline does not bias the estimate but can greatly reduce the variance of the sampled gradients.

Let $\hat{Q}(s, a)$ be a sampled estimate of $Q_\pi(s, a)$, assuming $\mathbb{E}[\hat{Q}(s, a)] = Q_\pi(s, a)$. Then, we can construct a new estimator $\hat{Q}_b(s, a)$ by subtracting some baseline $b: \mathcal{S} \rightarrow \mathbb{R}$, i.e. $\hat{Q}_b(s, a) = \hat{Q}(s, a) - b(s)$. Our only condition towards

⁶Note that $\rho_\pi(s_t \rightarrow s_{t+1}, 1) = 0$ if the episode already terminated due to reaching a terminal state on any previous step.

b is that it does not depend on the action a , though it can depend on the state s and even be a random variable [75]. Our sampled estimate of the gradient $\nabla_\theta J(\theta)$ becomes

$$\hat{\nabla}_\theta J(\theta) = \nabla_\theta \ln \pi_\theta(a | s) (\hat{Q}(s, a) - b(s)).$$

In expectation over the policy π , this yields

$$\begin{aligned} \mathbb{E}_\pi[\hat{\nabla}_\theta J(\theta)] &= \mathbb{E}_\pi[\nabla_\theta \ln \pi_\theta(A | S) (\hat{Q}(S, A) - b(S))] \\ &= \mathbb{E}_\pi[\nabla_\theta \ln \pi_\theta(A | S) \hat{Q}(S, A)] - \mathbb{E}_\pi[\nabla_\theta \ln \pi_\theta(A | S) b(S)] \end{aligned}$$

using the linearity of the expectation. Now, we show that the second part is 0. Using the Leibniz integral rule, we have that

$$\begin{aligned} \mathbb{E}_{S \sim d^{\pi_\theta}, A \sim \pi_\theta}[\nabla_\theta \ln \pi_\theta(A | S) b(S)] &= \int_{s \in \mathcal{S}} d^\pi(s) \int_{a \in \mathcal{A}} \pi_\theta(a | s) \nabla_\theta \ln \pi_\theta(a | s) b(s) da ds \\ &= \int_{s \in \mathcal{S}} d^\pi(s) b(s) \int_{a \in \mathcal{A}} \pi_\theta(a | s) \nabla_\theta \ln \pi_\theta(a | s) da ds \\ &= \int_{s \in \mathcal{S}} d^\pi(s) b(s) \int_{a \in \mathcal{A}} \pi_\theta(a | s) \frac{\nabla_\theta \pi_\theta(a | s)}{\pi_\theta(a | s)} da ds \\ &= \int_{s \in \mathcal{S}} d^\pi(s) b(s) \nabla_\theta \int_{a \in \mathcal{A}} \pi_\theta(a | s) da ds \\ &= \int_{s \in \mathcal{S}} d^\pi(s) b(s) \nabla_\theta 1 ds \\ &= 0 \end{aligned}$$

since $\pi(\cdot | s)$ is a probability distribution over actions. Thus, subtracting an action-independent baseline b from an action-value function estimator \hat{Q} does indeed not add any bias to the gradient estimate. While here we have shown this for a baseline which only depends on the current state, this result can be extended to baselines which depend on the current and all subsequent states [70].

Next, we analyze the effect on the variance of the gradient estimates. Here, we only provide an approximate explanation, see [27] for a more thorough analysis which derives bounds of the true variance. We can compute the variance using $\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. Due to the above, $\mathbb{E}[X]^2$ is independent of the baseline in our case. This yields

$$\begin{aligned} \arg \min_b \text{Var}_\pi[\nabla_\theta \ln \pi_\theta(A | S) (\hat{Q}(S, A) - b(S))] &= \arg \min_b \mathbb{E}_\pi \left[(\nabla_\theta \ln \pi_\theta(A | S) (\hat{Q}(S, A) - b(S)))^2 \right] \\ &\approx \arg \min_b \left(\mathbb{E}_\pi[(\nabla_\theta \ln \pi_\theta(A | S))^2] \cdot \mathbb{E}_\pi[(\hat{Q}(S, A) - b(S))^2] \right), \end{aligned}$$

where we approximated the variance by assuming independence of the two terms in the second step. Under this approximation, the variance of sampled gradients can be minimized by minimizing $\mathbb{E}_\pi[(\hat{Q}(S, A) - b(S))^2]$. This is a common least squares problem resulting in the optimal choice of $b(s) = \mathbb{E}_\pi[\hat{Q}(s, A)]$ (see Theorem D.8). This result indicates that an appropriately chosen baseline can potentially significantly reduce variance of the gradients. Using this choice for the baseline, we would like to compute gradients for sampled states and actions as

$$\begin{aligned} \nabla_\theta \ln \pi_\theta(a | s) (Q_\pi(s, a) - \mathbb{E}_{A \sim \pi_\theta}[Q_\pi(s, A)]) &= \nabla_\theta \ln \pi_\theta(a | s) (Q_\pi(s, a) - V_\pi(s)) \\ &= \nabla_\theta \ln \pi_\theta(a | s) A_\pi(s, a). \end{aligned}$$

Here, we used the relation of the value function V_π to Q_π and the definition of the advantage function A_π . Despite our approximations, this choice of a baseline turns out to yield almost the lowest possible variance of the gradients [70]. However, note that in practice the advantage function must also be estimated. Learning this estimate typically introduces bias [43, 76].

3.3 Importance Sampling

Importance sampling is a technique to calculate expectations under one distribution given samples from another [65, 31, 75]. Traditionally, this is only needed in off-policy RL, where we sample transitions using a behavior policy β but want to calculate expectations over the target policy π . However, in some implementations of on-policy algorithms the policy may be updated before all data generated by it is processed. This makes these implementations slightly off-policy and thus importance sampling becomes relevant even for theoretically on-policy algorithms [85]. We build our presentation of importance sampling on [75], Section 5.5.

Given a behavior policy β , we want to estimate the value function V_π of our target policy π . Generally, we have

$$V_\beta(s) = \mathbb{E}_\beta[G_t | S_t = s] \neq V_\pi(s).$$

We can calculate the probability of a trajectory $(a_t, s_{t+1}, a_{t+1}, \dots, a_{T-1}, s_T)$ under any policy π as

$$\prod_{k=t}^{T-1} \pi(a_k | s_k) P(s_{k+1} | s_k, a_k).$$

Now, we can define the importance sampling ratio.

Definition 3.2. (*Importance Sampling Ratio*) Given a target policy π , a behavior policy β and a trajectory $\tau = (a_t, s_{t+1}, a_{t+1}, \dots, s_T)$ generated by β , the importance sampling ratio is defined as

$$\rho_{t:T-1} := \frac{\prod_{k=t}^{T-1} \pi(a_k | s_k) P(s_{k+1} | s_k, a_k)}{\prod_{k=t}^{T-1} \beta(a_k | s_k) P(s_{k+1} | s_k, a_k)} = \frac{\prod_{k=t}^{T-1} \pi(a_k | s_k)}{\prod_{k=t}^{T-1} \beta(a_k | s_k)}.$$

Let \mathcal{T} be the set of possible trajectories. By multiplying returns of trajectories $\tau \in \mathcal{T}$ generated by the behavior policy β with the importance sampling ratio ρ we get

$$\begin{aligned} \mathbb{E}_\beta[\rho_{t:T-1} G_t | S_t = s] &= \mathbb{E}_\beta[\rho_{t:T-1} G(\tau) | S_t = s] \\ &= \sum_{\tau \in \mathcal{T}} \rho_{t:T-1} G(\tau) \prod_{k=t}^{T-1} \beta(a_k | s_k) P(s_{k+1} | s_k, a_k) \\ &= \sum_{\tau \in \mathcal{T}} \frac{\prod_{k=t}^{T-1} \pi(a_k | s_k)}{\prod_{k=t}^{T-1} \beta(a_k | s_k)} G(\tau) \prod_{k=t}^{T-1} \beta(a_k | s_k) P(s_{k+1} | s_k, a_k) \\ &= \sum_{\tau \in \mathcal{T}} G(\tau) \prod_{k=t}^{T-1} \frac{\pi(a_k | s_k)}{\beta(a_k | s_k)} \beta(a_k | s_k) P(s_{k+1} | s_k, a_k) \\ &= \sum_{\tau \in \mathcal{T}} G(\tau) \prod_{k=t}^{T-1} \pi(a_k | s_k) P(s_{k+1} | s_k, a_k) \\ &= \mathbb{E}_\pi[G_t | S_t = s] \\ &= V_\pi(s). \end{aligned}$$

The intuition behind this importance sampling correction is that, to evaluate π , we want to weigh returns more heavily that are more likely under π than under β and vice versa. As an extension of the derivation above, we also get the per-decision importance sampling ratio $\rho := \frac{\pi(a|s)}{\beta(a|s)}$ [75].

Using importance sampling, we can derive the following approximate policy gradients of the target policy π_θ in an off-policy setting with behavior policy β :

$$\nabla_\theta J(\theta) \approx \eta \mathbb{E}_{S \sim d^\beta, A \sim \beta} \left[\frac{\pi_\theta(A | S)}{\beta(A | S)} Q_{\pi_\theta}(S, A) \nabla_\theta \ln \pi_\theta(A | S) \right].$$

See [19] for a proof. Note that η now is the average episode length under β .

4 Policy Gradient Algorithms

Building on Theorem 3.1, several policy gradient algorithms have been proposed, which compute sample-based estimates $\hat{\nabla}_\theta J(\theta)$ of the actual policy gradients $\nabla_\theta J(\theta)$. This is done by constructing surrogate objectives J_* such that

$\hat{\nabla}_\theta J(\theta) = \nabla_\theta J_*(\theta)$. Additionally, most algorithms focus on stabilizing learning by regularizing the policy [5] and reducing the variance of $\hat{\nabla}_\theta J(\theta)$ [77]. In this section, we derive the most prominent⁷ algorithms before comparing them in the final subsection.

4.1 REINFORCE

REINFORCE (REward Increment = Non-negative Factor \times Offset Reinforcement \times Characteristic Eligibility) [88] is the earliest policy gradient algorithm. While this algorithm precedes the formulation of the Policy Gradient Theorem, REINFORCE can be seen as a straightforward application of it. By using Monte Carlo methods [75] to estimate Q_π in Equation (13), i.e. by sampling entire episodes to compute the sample returns $G_t = \sum_{k=0}^T \gamma^k r_{t+k+1}$, REINFORCE samples policy gradients

$$\hat{\nabla}_\theta J(\theta) = G_t \nabla_\theta \ln \pi_\theta(a_t | s_t).$$

Using the generic policy gradient update from Equation (3) results in the gradient ascend updates

$$\theta_{\text{new}} = \theta + \alpha G_t \nabla_\theta \ln \pi_\theta(a_t | s_t)$$

where $\alpha \in (0, 1]$ is the learning rate determining the step size of the gradient steps and is set as a hyperparameter. At times, REINFORCE is extended by subtracting some baseline value from G_t to reduce variance [88]. The pseudocode for REINFORCE is presented in Algorithm 3.

Algorithm 3 REINFORCE

Require: $\alpha \in (0, 1], \gamma \in [0, 1]$

Initialize θ at random

for all episodes **do**

 Generate trajectory $s_0, a_0, r_1, s_1 \dots, s_T$ under policy π_θ

for $t = 1, \dots, T$ **do**

$G_t \leftarrow \sum_{k=t}^T \gamma^{k-t} r_k$ ▷ estimate expected return Q_π
 $\theta \leftarrow \theta + \alpha G_t \nabla_\theta \ln \pi_\theta(a_t | s_t)$ ▷ update policy parameters

end for

end for

4.2 A3C

Instead of estimating Q_π directly via sampling as in REINFORCE, we can alternatively learn such an estimate via function approximation. Algorithms that use this approach to learn a parameterized action-value function \hat{Q}_ϕ or value function \hat{V}_ϕ (called critic) with parameters ϕ in addition to learning the parameterized policy π_θ (called actor) are referred to as actor-critic algorithms [75]. Note that in practice the actor and the critic may also share parameters.

The most archetypical representative of this class of algorithms is Asynchronous Advantage Actor-Critic (A3C) [54]. A3C builds on two main ideas from which the algorithm's name originates. First, as suggested by the results from Section 3.2, A3C learns an estimate \hat{A}_ϕ of the advantage function indirectly by learning an estimate \hat{V}_ϕ of the value function. Second, A3C introduces the concept of using multiple parallel actors to interact with the environment to stabilize training. We will discuss both ideas in detail below. The algorithm samples policy gradients

$$\hat{\nabla}_\theta J(\theta) = \frac{1}{|\mathcal{D}|} \sum_{s,a \in \mathcal{D}} \hat{A}_\phi(s, a) \nabla_\theta \ln \pi_\theta(a | s),$$

where \mathcal{D} is a batch of transitions collected by the actors. The pseudocode for A3C is presented in Algorithm 4.

In the original work [55], the advantage function is estimated via

$$\hat{A}_\phi(s_t, a_t) = \left(\sum_{i=0}^{k-1} \gamma^i r_{t+i} + \gamma^k \hat{V}_\phi(s_{t+k}) \right) - \hat{V}_\phi(s_t). \quad (14)$$

⁷As determined by their impact on subsequent research and the adoption rate by users.

To understand this estimate, observe that

$$\begin{aligned}
 A_\pi(s_t, a_t) &= Q_\pi(s_t, a_t) - V_\pi(s_t) \\
 &= \mathbb{E}_\pi \left[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s_t, A_t = a_t \right] - V_\pi(s_t) \\
 &= \mathbb{E}_\pi \left[R_{t+1} + \gamma R_{t+2} + \gamma^2 V_\pi(S_{t+2}) \mid S_t = s_t, A_t = a_t \right] - V_\pi(s_t) \\
 &\quad \vdots \\
 &= \mathbb{E}_\pi \left[\sum_{i=0}^{k-1} \gamma^i R_{t+i} + \gamma^k V_\pi(S_{t+k}) \mid S_t = s_t, A_t = a_t \right] - V_\pi(s_t),
 \end{aligned}$$

for any $k \in \mathbb{N}$, which follows from the definition of the value and action-value functions as well as their relationship. Sampling this n-step temporal difference [75] expression and replacing V_π with our learned \hat{V}_ϕ yields Equation (14). Simultaneously to updating π_θ , we learn \hat{V}_ϕ by minimizing the mean squared error loss

$$\frac{1}{|\mathcal{D}|} \sum_{\mathcal{D}} \left(\left(\sum_{i=0}^{k-1} \gamma^i r_{t+i} + \gamma^k \hat{V}_\phi(s_{t+k}) \right) - \hat{V}_\phi(s_t) \right)^2$$

over ϕ via SGD. Note that the inner expression is identical to the right hand side in Equation (14). In Equation (14), we compute the difference between the estimated return when choosing action a_t in state s_t and the estimated return when in state s_t , under policy π respectively. However, a_t is sampled from π such that in expectation this difference should be 0 for the true value function V_π . Hence, we minimize this squared difference to optimize ϕ by treating the first term, $\sum_{i=0}^{k-1} \gamma^i r_{t+i} + \gamma^k \hat{V}_\phi(s_{t+k})$, as independent of ϕ .

The use of multiple parallel actors is justified as follows. Deep RL is notoriously unstable, which was first resolved by off-policy algorithms using replay buffers that store and reuse sampled transitions for multiple updates [55]. As an alternative, [54] propose using several actors $\pi_\theta^{(1)}, \dots, \pi_\theta^{(k)}$ to decrease noise by accumulating the gradients over multiple trajectories. These accumulated gradients are applied to a centrally maintained copy of θ , which is then redistributed to each actor. By doing this asynchronously, each actor has a potentially unique set of parameters at any point in time compared to the other actors. This decreases the correlation of the sampled trajectories across actors, which can further stabilize learning.

As a final implementation detail, the policy loss function of A3C, from which the policy gradients are obtained, is typically augmented with an entropy bonus for the policy. Thus, the policy gradients become

$$\hat{\nabla}_\theta J(\theta) = \frac{1}{|\mathcal{D}|} \sum_{\mathcal{D}} \left(\left(\sum_{i=0}^{k-1} \gamma^i r_{t+i} + \gamma^k \hat{V}_\phi(s_{t+k}) - \hat{V}_\phi(s_t) \right) \nabla_\theta \ln \pi_\theta(a_t \mid s_t) + \beta \nabla_\theta H(\pi_\theta(\cdot \mid s_t)) \right),$$

where H is the entropy (see Definition D.4) and the entropy coefficient β is a hyperparameter. This entropy bonus, first proposed by [89], regularizes the policy such that it does not prematurely converges to a suboptimal policy. By rewarding entropy, the policy is encouraged to spread probability mass over actions which improves exploration [54].

4.3 TRPO

Excessively large changes in the policy can result in instabilities during the training of RL algorithms. Even small changes in policy parameters θ can lead to significant changes in the resulting policy and its performance. Hence, small step sizes during gradient ascent cannot fully remedy this problem and would impair the sample efficiency of the algorithm [3]. Trust Region Policy Optimization (TRPO) [69] mitigates these issues by imposing a trust region constraint on the Kullback-Leibler (KL) divergence (see Definition D.5) between consecutive policies. In addition, TRPO uses an off-policy correction through importance sampling as discussed in Section 3.3 to account for the interleaved optimization and collection of transitions.

TRPO samples gradients

$$\hat{\nabla}_\theta J(\theta) = \frac{1}{|\mathcal{D}|} \sum_{s,a \in \mathcal{D}} \hat{A}_\phi(s, a) \nabla_\theta \frac{\pi_\theta(a \mid s)}{\pi_{\text{old}}(a \mid s)}$$

and postprocesses them as detailed below to solve the approximate trust region optimization problem

$$\begin{aligned}
 \max_{\theta} \quad & \left(J_{\text{TRPO}}(\theta) = \mathbb{E}_{S \sim d^{\pi_{\text{old}}}, A \sim \pi_{\text{old}}} \left[\hat{A}_\phi(S, A) \frac{\pi_\theta(A \mid S)}{\pi_{\text{old}}(A \mid S)} \right] \right) \\
 \text{subject to} \quad & \mathbb{E}_{S \sim d^{\pi_{\text{old}}}} [D_{KL}(\pi_{\text{old}}(\cdot \mid S) \parallel \pi_\theta(\cdot \mid S))] \leq \delta
 \end{aligned}$$

Algorithm 4 A3C

Require: $n \in \mathbb{N}$, $\alpha \in (0, 1]$, $\gamma \in [0, 1]$, $t_{\text{MAX}} \in \mathbb{N}$, $T_{\text{MAX}} \in \mathbb{N}$

Initialize θ and ϕ at random

for $i = 1, \dots, n$ **do** ▷ in parallel

- while** $T \leq T_{\text{MAX}}$ **do**
 - $d\theta \leftarrow 0$, $d\phi \leftarrow 0$ ▷ reset gradients
 - $\theta^{(i)} \leftarrow \theta$, $\phi^{(i)} \leftarrow \phi$ ▷ synchronize parameters on actors
 - $s_t \sim p_0$ ▷ sample start state
 - $t_{\text{start}} \leftarrow t$
 - while** s_t not terminal and $t - t_{\text{start}} \leq t_{\text{MAX}}$ **do**
 - $a_t \sim \pi_{\theta^{(i)}}$ ▷ sample action
 - $s_{t+1}, r_{t+1} \sim P(s_t, a_t)$ ▷ sample next state and reward
 - $t \leftarrow t + 1$, $T \leftarrow T + 1$
 - end while**
 - $R \leftarrow \begin{cases} 0 & \text{if } s_t \text{ is terminal} \\ V_{\phi^{(i)}}(s_t) & \text{else} \end{cases}$ ▷ bootstrap if necessary
 - for** $j = t - 1, \dots, t_{\text{start}}$ **do**
 - $R \leftarrow r_j + \gamma R$
 - $A = R - V_{\phi^{(i)}}(s_j)$
 - $d\theta \leftarrow d\theta + \nabla_{\theta^{(i)}} \ln \pi_{\theta^{(i)}}(a_j | s_j) A$ ▷ accumulate gradients
 - $d\phi \leftarrow d\phi + \nabla_{\phi^{(i)}} (R - V_{\phi^{(i)}}(s_j))^2$ ▷ accumulate gradients
 - end for**
 - update θ and ϕ using $d\theta$ and $d\phi$ via gradient ascent / descent
- end while**

end for

where $\pi_{\text{old}} = \pi_{\theta_{\text{old}}}$ is the previous policy and θ_{old} the corresponding parameters. This optimization problem is an approximation to an objective with convergence guarantees, which we will show in the following. We start by presenting [69]'s main theoretical result. Consider the objective of maximizing the expected return $\mathbb{E}_{S_0 \sim p_0, \pi}[G_0]$ under policy π , which we denote as $\eta(\pi)$ here. Let L_π be the following local approximation of η :

$$L_\pi(\tilde{\pi}) = \eta(\pi) + \int_{s \in \mathcal{S}} d\pi(s) \int_{a \in \mathcal{A}} \tilde{\pi}(a | s) A_\pi(s, a) da ds,$$

with $L_{\pi_\theta}(\pi_\theta) = \eta(\pi_\theta)$ and $\nabla_\theta L_{\pi_{\theta_0}}(\pi_\theta)|_{\theta=\theta_0} = \nabla_\theta \eta(\pi_\theta)|_{\theta=\theta_0}$ [38]. Based on the total variation divergence D_{TV} (see Definition D.6), we define

$$D_{TV}^{\max}(\pi, \tilde{\pi}) := \max_{s \in \mathcal{S}} D_{TV}(\pi(\cdot | s) \| \tilde{\pi}(\cdot | s)).$$

Then, we have [69]:

Theorem 4.1. *Let $\alpha = D_{TV}^{\max}(\pi_{\text{old}}, \pi_{\text{new}})$, then*

$$\eta(\pi_{\text{new}}) \geq L_{\pi_{\text{old}}}(\pi_{\text{new}}) - \frac{4\varepsilon\gamma}{(1-\gamma)^2} \alpha^2, \quad (15)$$

where $\varepsilon = \max_{s \in \mathcal{S}, a \in \mathcal{A}} |A_\pi(s, a)|$.

See the appendix in [69] for a proof. By using the relationship between total variation divergence and KL divergence $D_{TV}(\pi \| \tilde{\pi})^2 \leq D_{KL}(\pi \| \tilde{\pi})$ [58] and setting $D_{KL}^{\max}(\pi, \tilde{\pi}) := \max_{s \in \mathcal{S}} D_{KL}(\pi(\cdot | s) \| \tilde{\pi}(\cdot | s))$ and $C = \frac{4\varepsilon\gamma}{(1-\gamma)^2}$, we derive the following lower bound for the objective from Equation (15):

$$\eta(\pi_{\text{new}}) \geq L_{\pi_{\text{old}}}(\pi_{\text{new}}) - CD_{KL}^{\max}(\pi_{\text{old}}, \pi_{\text{new}}) \quad (16)$$

Iteratively maximizing the right-hand side yields a sequence of policies $\pi_i, \pi_{i+1}, \pi_{i+2}, \dots$ with the monotonic improvement guarantee $\eta(\pi_i) \leq \eta(\pi_{i+1}) \leq \eta(\pi_{i+2}) \leq \dots$. This is because we have equality in (16) for $\pi_{\text{new}} = \pi_{\text{old}}$ and hence

$$\eta(\pi_{i+1}) - \eta(\pi_i) \geq \left(L_{\pi_i}(\pi_{i+1}) - CD_{KL}^{\max}(\pi_i, \pi_{i+1}) \right) - \left(L_{\pi_i}(\pi_i) - CD_{KL}^{\max}(\pi_i, \pi_i) \right),$$

which is non-negative as we maximize over π each iteration. Thus, we could construct a Minorization-Maximization-type algorithm [37] which maximizes the right-hand side of Inequality (16) at each iteration and is thereby guaranteed to converge to an optimum as the objective is bounded.

Such an algorithm would be impractical as it requires evaluating the advantage function at every point in the state-action product space $\mathcal{S} \times \mathcal{A}$ and the KL penalty at every point in the state space \mathcal{S} . Hence, [69] apply several approximations to the objective stemming from Inequality (16). We replace the KL penalty, which would yield restrictively small step sizes given by C , by a trust region constraint:

$$\begin{aligned} & \max_{\theta} L_{\pi_{\text{old}}}(\pi_{\theta}) \\ \text{subject to } & D_{KL}^{\max}(\pi_{\text{old}}, \pi_{\theta}) \leq \delta. \end{aligned}$$

To avoid computing D_{KL}^{\max} , we use the average KL divergence

$$\bar{D}_{KL}^{\pi_{\text{old}}}(\pi \| \tilde{\pi}) := \mathbb{E}_{S \sim d^{\pi_{\text{old}}}} [D_{KL}(\pi(\cdot | S) \| \tilde{\pi}(\cdot | S))]$$

between policies as heuristic constraint, which we can sample. Further, we rewrite the surrogate objective $\max_{\theta} L_{\pi_{\text{old}}}(\pi_{\theta})$ as an expectation over the old policy π_{old} via importance sampling. Note that $\eta(\pi_{\text{old}})$ is a constant w.r.t θ :

$$\begin{aligned} \arg \max_{\theta} L_{\pi_{\text{old}}}(\pi_{\theta}) &= \arg \max_{\theta} \left(\eta(\pi_{\text{old}}) + \int_{s \in \mathcal{S}} d^{\pi_{\text{old}}}(s) \int_{a \in \mathcal{A}} \pi_{\theta}(a | s) A_{\pi_{\text{old}}}(s, a) da ds \right) \\ &= \arg \max_{\theta} \int_{s \in \mathcal{S}} d^{\pi_{\text{old}}}(s) \int_{a \in \mathcal{A}} \pi_{\theta}(a | s) A_{\pi_{\text{old}}}(s, a) da ds \\ &= \arg \max_{\theta} \int_{s \in \mathcal{S}} d^{\pi_{\text{old}}}(s) \int_{a \in \mathcal{A}} \frac{\pi_{\text{old}}(a | s)}{\pi_{\text{old}}(a | s)} \pi_{\theta}(a | s) A_{\pi_{\text{old}}}(s, a) da ds \\ &= \arg \max_{\theta} \mathbb{E}_{S \sim d^{\pi_{\text{old}}}, A \sim \pi_{\text{old}}} \left[\frac{\pi_{\theta}(A | S)}{\pi_{\text{old}}(A | S)} A_{\pi_{\text{old}}}(S, A) \right]. \end{aligned}$$

Using these modifications, we are now left with solving the trust region problem

$$\begin{aligned} & \max_{\theta} \mathbb{E}_{S \sim d^{\pi_{\text{old}}}, A \sim \pi_{\text{old}}} \left[\frac{\pi_{\theta}(A | S)}{\pi_{\text{old}}(A | S)} A_{\pi_{\text{old}}}(S, A) \right] \\ \text{subject to } & \mathbb{E}_{S \sim d^{\pi_{\text{old}}}} [D_{KL}(\pi_{\text{old}}(\cdot | S) \| \pi_{\theta}(\cdot | S))] \leq \delta. \end{aligned} \tag{17}$$

To approximately solve this constrained problem, [69] use backtracking line search, where the search direction is computed by Taylor-expanding (see Theorem D.12) the objective function and the constraint. Let $g = \nabla_{\theta} \mathbb{E}_{S \sim d^{\pi_{\text{old}}}, A \sim \pi_{\text{old}}} \left[\frac{\pi_{\theta}(A | S)}{\pi_{\text{old}}(A | S)} A_{\pi_{\text{old}}}(S, A) \right]$. Approximating $L_{\pi_{\text{old}}}(\pi_{\theta})$ to first order around θ_{old} yields

$$L_{\pi_{\text{old}}}(\pi_{\theta}) \approx g^T (\theta - \theta_{\text{old}}),$$

where we again ignored the constant $\eta(\pi_{\text{old}})$. The quadratic approximation of the constraint at θ_{old} is

$$\bar{D}_{KL}^{\pi_{\text{old}}}(\pi \| \tilde{\pi}) \approx \frac{1}{2} (\theta - \theta_{\text{old}})^T H (\theta - \theta_{\text{old}}),$$

where H is the Fisher information matrix, which is estimated via

$$\hat{H}_{i,j} = \frac{1}{|\mathcal{D}|} \sum_{s \in \mathcal{D}} \frac{\partial^2}{\partial \theta_i \partial \theta_j} D_{KL}(\pi_{\text{old}}(\cdot | s) \| \pi(\cdot | s)),$$

albeit the full matrix is not required. We solve the resulting approximate optimization problem analytically using Lagrangian duality methods [12] leading to

$$\theta_{\text{new}} = \theta_{\text{old}} + \sqrt{\frac{2\delta}{g^T \hat{H}^{-1} g}} \hat{H}^{-1} g.$$

However, due to the Taylor approximations, this solution may not satisfy the original trust region constraint or may not improve the surrogate objective of Problem (17). Therefore, TRPO employs backtracking line search along the search direction $H^{-1}g$ with search parameter $\beta \in (0, 1)$:

$$\theta_{\text{new}} = \theta_{\text{old}} + \beta^m \sqrt{\frac{2\delta}{g^T H^{-1} g}} H^{-1} g.$$

Algorithm 5 TRPO

Require: $\delta \in \mathbb{R}$, $b \in (0, 1)$, $K \in \mathbb{N}$, $\alpha \in (0, 1]$, $U \in \mathbb{N}$, $T \in \mathbb{N}$

Initialize θ and ϕ at random
 $t \leftarrow 0$
while $t \leq T$ **do**

for $i = 1, \dots, U$ **do**

$a \sim \pi_\theta$ ▷ sample action
 $\beta(a | s) \leftarrow \pi_\theta(a | s)$
 $s, r \sim P(s, a)$ ▷ sample next state and reward
 $t \leftarrow t + 1$
Store $(a, s, r, \beta(a | s))$ in \mathcal{D}

end for

for all epochs **do**

Compute returns R and advantages A
 $g \leftarrow \frac{1}{|\mathcal{D}|} \sum_{\mathcal{D}} \nabla_\theta \frac{\pi_\theta(a|s)}{\beta(a|s)} A$
Compute \hat{H} as the Hessian of the sample average KL-divergence
Compute $d \approx \hat{H}^{-1} g$ via conjugate gradient algorithm
 $m \leftarrow 0$
repeat
 $\theta \leftarrow \theta_{\text{old}} + b^m \sqrt{\frac{2\delta}{d^\top \hat{H} d}} d$
 $m \leftarrow m + 1$

until (sample loss improves and KL constraint satisfied) or $m > K$
 $d\phi \leftarrow \frac{1}{|\mathcal{D}|} \sum_{\mathcal{D}} \nabla_\phi (R - V_\phi(s))^2$
Update ϕ using $d\phi$ via gradient descent

end for

end while

We choose the exponent m as the smallest non-negative integer such that the trust region constraint is satisfied and the surrogate objective improves. We circumvent the computationally expensive matrix inversion of H for the search direction $d \approx \hat{H}^{-1} g$ by computing d via the conjugate gradient algorithm [30]. To further reduce the computational costs, the Fisher-vector products in this process can also be only calculated on a subset of the dataset \mathcal{D} of sampled transitions.

[69] do not specify an advantage estimator to be used in TRPO. The algorithm is commonly used with either the estimator used by A3C or the one which we present in the next subsection. TRPO is typically used with multiple parallel actors as A3C. The pseudocode for TRPO is presented in Algorithm 5. We remark that while being a policy-based algorithm, TRPO does not strictly adhere to Definition 2.2 as it solves a constrained optimization problem via line search. Yet, it does compute gradients of its objective function w.r.t. the policy parameters and therefore we treat it as a policy gradient algorithm.

4.4 PPO

Given the complexity of TRPO, Proximal Policy Optimization (PPO) [71] is designed to enforce comparable constraints on the divergence between consecutive policies during the learning process while simplifying the algorithm to not require second-order methods. This is achieved by heuristically flattening the gradients outside of an approximate trust region around the old policy. In addition, PPO uses a novel method to learn an estimate of the advantage function.

Let $r_\theta(a | s) = \frac{\pi_\theta(a|s)}{\pi_{\text{old}}(a|s)}$. Then, PPO uses the following estimate of the policy gradients:

$$\hat{\nabla}_\theta J(\theta) = \frac{1}{|\mathcal{D}|} \sum_{s,a \in \mathcal{D}} \hat{A}_\phi(s, a) \nabla_\theta \min \left\{ r_\theta(a | s), \text{clip} \left(r_\theta(a | s), 1 - \varepsilon, 1 + \varepsilon \right) \right\}. \quad (18)$$

Here, the clip-function $\text{clip}: \mathbb{R} \times \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ is defined by

$$\text{clip}(x, a, b) = \begin{cases} a & \text{if } x < a, \\ x & \text{if } a \leq x \leq b, \\ b & \text{if } b < x. \end{cases}$$

and is applied element-wise to r_θ . ε is a hyperparameter.

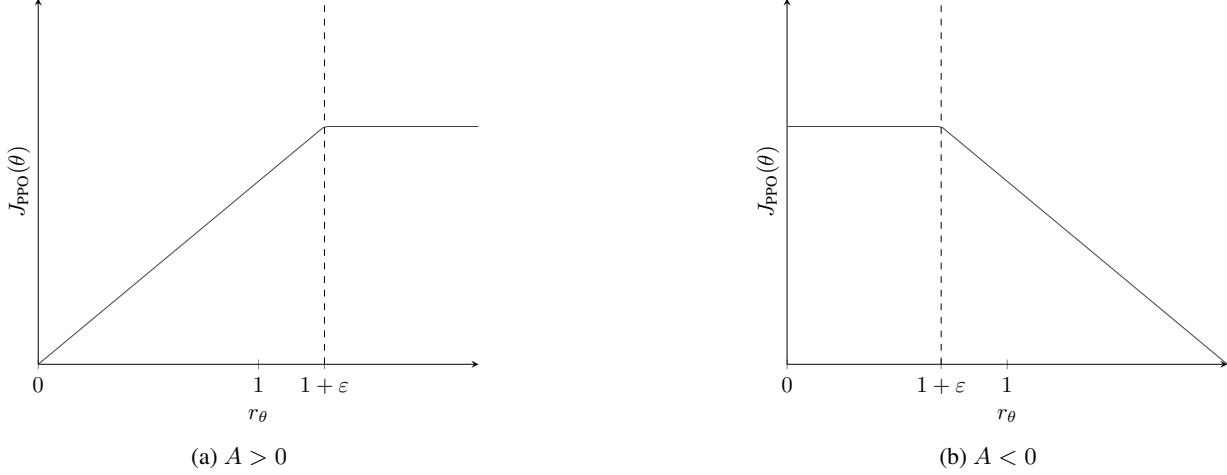


Figure 3: Illustration of the conservative clipping of PPO’s objective function, which is shown as a function of the ratio r_θ for a single transition depending on whether the advantages are positive (a) or negative (b). Replicated from [71].

This clipped objective conservatively removes the incentive for moving the new policy to far away from the old one. Intuitively, this can be seen as follows. We distinguish two cases: positive and negative estimated advantages $\hat{A}(s, a)$, i.e. whether action a is good or bad. If $\hat{A}(s, a) > 0$, the surrogate objective $J_{\text{PPO}}(\theta)$ increases when a becomes more likely. Similarly, if $\hat{A}(s, a) < 0$, $J_{\text{PPO}}(\theta)$ increases when a becomes less likely. Hence, we want to adjust the policy parameters θ accordingly. However, by clipping the policy ratio r_θ , this positive effect on the objective function disappears once we move outside the clip range. This clipping process is conservative as we only clip if the objective function would improve. If the policy is changed in the opposite direction such that $J_{\text{PPO}}(\theta)$ decreases, r_θ is not clipped due to taking the minimum in Equation (18). Figure 3 illustrates this explanation. The pseudocode for PPO is presented in Algorithm 6.

Algorithm 6 PPO

Require: $\varepsilon \in \mathbb{R}$, $\alpha \in (0, 1]$, $\gamma \in [0, 1]$, $\lambda \in [0, 1]$, $U \in \mathbb{N}$, $T \in \mathbb{N}$
 Initialize θ and ϕ at random
 $t \leftarrow 0$
while $t \leq T$ **do**
for $i = 1, \dots, U$ **do**
 $a \sim \pi_\theta$ ▷ sample action
 $\beta(a | s) \leftarrow \pi_\theta(a | s)$
 $s, r \sim P(s, a)$ ▷ sample next state and reward
 $t \leftarrow t + 1$
 Store $(a, s, r, \beta(a | s))$ in \mathcal{D}
end for
for all epochs do
 $R, A \leftarrow \text{computeGAE}(v, r, \lambda, \gamma)$ ▷ Compute returns and advantages
 $d\theta \leftarrow \nabla_\theta \frac{1}{|\mathcal{D}|} \sum_{\mathcal{D}} \min\left(\frac{\pi(a|s)}{\beta(a|s)}, \text{clip}\left(\frac{\pi(a|s)}{\beta(a|s)}, 1 - \varepsilon, 1 + \varepsilon\right)\right) A$
 $d\phi \leftarrow \nabla_\phi \frac{1}{|\mathcal{D}|} \sum_{\mathcal{D}} (R - V_\phi(s))^2$
 update θ and ϕ using $d\theta$ and $d\phi$ via gradient ascent / descent
end for
end while

To compute the estimate \hat{A}_ϕ of the advantage function, PPO uses generalized advantage estimation (GAE) [70] to further reduce the variance of gradients. GAE computes the estimated advantage as

$$\hat{A}_\phi(s_t, a_t) = \sum_{i=t}^{T-1} (\gamma \lambda)^{i-t} \delta_i, \quad (19)$$

where $\delta_i = r_i + \gamma \hat{V}_\phi(s_{i+1}) - \hat{V}_\phi(s_i)$. The value function estimate \hat{V}_ϕ is learned by minimizing

$$\frac{1}{|\mathcal{D}|} \sum_{\mathcal{D}} \left((\hat{A}_\phi(s, a) + \hat{V}_\phi(s)) - \hat{V}_\phi(s) \right)^2,$$

where the first term is treated as independent of ϕ . GAE relates to the idea of eligibility traces [74] to use both the sampled rewards and the current value function estimate on every time step. By computing such an exponentially weighted estimator, GAE reduces the variance of the policy gradients at the cost of introducing a slight bias to the value function estimate [70]. The hyperparameters γ and λ both adjust this bias-variance tradeoff. γ does so by scaling the value function estimate \hat{V} whereas λ controls the dependence on delayed rewards. Note that GAE is a strict generalization of A3C's advantage estimate as Equation (19) reduces to Equation (14) when $\lambda = 1$. The pseudocode for GAE is presented in Algorithm 7.

Algorithm 7 GAE

```

Require:  $\gamma \in [0, 1]$ ,  $\lambda \in [0, 1]$ 
Require: rewards  $(r_k)_{k=t}^{t+n}$ , values  $(v_k)_{t=k}^{t+n+1}$ 
 $A_t, \dots, A_{t+n} \leftarrow 0$ 
 $x \leftarrow 0$ 
for  $i = t + n, \dots, t$  do
    if transition was terminal then
         $\omega \leftarrow 1$ 
    else
         $\omega \leftarrow 0$ 
    end if
     $\delta \leftarrow r_i + \gamma \cdot v_{i+1} \cdot (1 - \omega) - v_i$ 
     $x \leftarrow \delta + \gamma \cdot \lambda \cdot (1 - \omega) \cdot x$ 
     $A_i \leftarrow x$ 
end for
for  $i = t, \dots, t + n$  do
     $R_i \leftarrow A_i + v_i$ 
end for
```

Beyond these main innovations, PPO uses several implementational details to improve learning. PPO conducts multiple update epochs for each batch of data such that several gradient descent steps are based on the same transitions to increase sample efficiency and speed up learning. Moreover, PPO commonly augments its surrogate objective with an entropy bonus $H(\pi_\theta(\cdot | s))$ and uses multiple actors similarly to A3C. Lastly, we note that further algorithms have been proposed as modifications of PPO, e.g. Phasic Policy Gradients [16] and Robust Policy Optimization [61], which we will not discuss further as they only modify minor details.

4.5 V-MPO

In the previous algorithms, we learn a policy from the control perspective by selecting actions to maximize expected rewards. In this subsection, we consider an alternative formulation of RL problems, which casts them as probabilistic inference problems of estimating posterior policies that are consistent with a desired outcome [1]. This problem is then solved via Expectation Maximization (EM) [20]. This procedure was first proposed in the off-policy algorithm Maximum a-posteriori Policy Optimization (MPO) [2, 1]. Here, we discuss its on-policy variant V-MPO [73], where the "V" in the name refers to learning the value function V_π instead of Q_π as in MPO.

The main idea of V-MPO is to find a maximum a posteriori estimate of the policy by sequentially finding a tight lower bound on the posterior and then maximizing this lower bound. This problem can be transformed into the objective function

$$J_{\text{V-MPO}}(\theta, \eta, \nu) = \mathcal{L}_\pi(\theta) + \mathcal{L}_\eta(\eta) + \mathcal{L}_\nu(\theta, \nu),$$

where \mathcal{L}_π is the policy loss

$$\mathcal{L}_\pi(\theta) = - \sum_{a, s \in \mathcal{D}} \frac{\exp\left(\frac{\hat{A}_\phi(s, a)}{\eta}\right)}{\sum_{a', s' \in \mathcal{D}} \exp\left(\frac{\hat{A}_\phi(s', a')}{\eta}\right)} \ln \pi_\theta(a | s), \quad (20)$$

\mathcal{L}_η is the temperature loss

$$\mathcal{L}_\eta(\eta) = \eta \varepsilon_\eta + \eta \ln \left[\frac{1}{|\mathcal{D}|} \sum_{a, s \in \mathcal{D}} \exp\left(\frac{\hat{A}_\phi(s, a)}{\eta}\right) \right] \quad (21)$$

and \mathcal{L}_ν is the trust-region loss

$$\mathcal{L}_\nu(\theta, \nu) = \frac{1}{|\mathcal{D}|} \sum_{s \in \mathcal{D}} \left(\nu \left(\varepsilon_\nu - \text{sg} \left[\left[D_{KL}(\pi_{\text{old}}(\cdot | s) \| \pi_\theta(\cdot | s)) \right] \right] \right) + \text{sg}[\nu] D_{KL}(\pi_{\text{old}}(\cdot | s) \| \pi_\theta(\cdot | s)) \right). \quad (22)$$

Here, $\text{sg}[\cdot]$ is a stop-gradient operator, meaning its arguments are treated as constants when computing gradients, η is a learnable temperature parameter, ν is a learnable KL-penalty parameter, ε_ν and ε_η are hyperparameters, \mathcal{D} is a batch of transitions and $\tilde{\mathcal{D}} \subset \mathcal{D}$ is the half of these transitions with the largest advantages. We will provide a sketch of how to derive this objective function in the following. We refer the interested reader to Appendix C for a more detailed derivation.

Let $p_\theta(s, a) = \pi_\theta(a | s) d^{\pi_\theta}(s)$ denote the joint state-action distribution under policy π_θ conditional on the parameters θ . Let \mathcal{I} be a binary random variable whether the updated policy π_θ is an improvement over the old policy π_{old} , i.e. $\mathcal{I} = 1$ if it is an improvement. We assume the conditional probability of π_θ being an improvement given a state s and an action a is proportional to the following expression

$$p_\theta(\mathcal{I} = 1 | s, a) \propto \exp \left(\frac{A_{\pi_{\text{old}}}(s, a)}{\eta} \right). \quad (23)$$

Given the desired outcome $\mathcal{I} = 1$, we seek the posterior distribution conditioned on this event. Specifically, we seek the maximum a posteriori estimate

$$\begin{aligned} \theta^* &= \arg \max_{\theta} [p_\theta(\mathcal{I} = 1) \rho(\theta)] \\ &= \arg \max_{\theta} [\ln p_\theta(\mathcal{I} = 1) + \ln \rho(\theta)], \end{aligned} \quad (24)$$

where ρ is some prior distribution. $\ln p_\theta(\mathcal{I} = 1)$ can be rewritten as

$$\ln p_\theta(\mathcal{I} = 1) = \mathbb{E}_{S, A \sim \psi} \left[\ln \frac{p_\theta(\mathcal{I} = 1, S, A)}{\psi(S, A)} \right] + D_{KL}(\psi \| p_\theta(\cdot, \cdot | \mathcal{I} = 1)) \quad (25)$$

for some distribution ψ over $\mathcal{S} \times \mathcal{A}$. Observe that, since the KL divergence is non-negative, the first term is a lower bound for $\ln p_\theta(\mathcal{I} = 1)$. Akin to EM algorithms, V-MPO now iterates by choosing the variational distribution ψ in the expectation (E) step to minimize the KL divergence in Equation (25) to make the lower bound as tight as possible. In the maximization (M) step, we maximize this lower bound and the prior $\ln \rho(\theta)$ to obtain a new estimate of θ^* via Equation (24).

First, we consider the E-step. Under the proportionality assumption (23), we turn the problem of finding a variational distribution ψ to minimize $D_{KL}(\psi \| p_{\theta_{\text{old}}}(\cdot, \cdot | \mathcal{I} = 1))$ into an optimization problem over the temperature η . This is formulated as a constrained problem subject to a bound on the KL divergence between ψ and the previous state-action distribution $p_{\theta_{\text{old}}}$ while ensuring that ψ is a state-action distribution. To enable optimizing η via gradient descent, we transform this constrained problem into an unconstrained problem via Lagrangian relaxation, which emits both the form of the variational distribution

$$\psi(s, a) = \frac{p_{\theta_{\text{old}}}(s, a) p_{\theta_{\text{old}}}(\mathcal{I} = 1 | s, a)}{\int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} p_{\theta_{\text{old}}}(s, a) p_{\theta_{\text{old}}}(\mathcal{I} = 1 | s, a) da ds}$$

and the temperature loss (21)

$$\mathcal{L}_\eta(\eta) = \eta \varepsilon_\eta + \eta \ln \left(\int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \exp \left(\frac{A_{\pi_{\text{old}}}(s, a)}{\eta} \right) da ds \right).$$

[73] find that using only the highest 50 % of advantages per batch when sampling these expressions, i.e. replacing \mathcal{D} with $\tilde{\mathcal{D}}$, substantially improves the algorithm. The advantage function A_π is estimated by \hat{A}_ϕ , which is learned as in A3C.

Then, in the M-Step we solve the maximum a posterior estimation problem (24) over the policy parameters θ for the constructed variational distribution $\psi(s, a)$ and the thereby implied lower bound. This lower bound, i.e. the first term in Equation (25), becomes the weighted maximum likelihood policy loss (20)

$$\mathcal{L}_\pi(\theta) = - \int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \psi(s, a) \ln \pi_\theta(a | s) da ds$$

after dropping terms independent of θ . This loss is computed over the same reduced batch $\tilde{\mathcal{D}}$ as the temperature loss, effectively assigning out-of-sample transitions a weight of zero. Simultaneously, we want to maximize the prior $\rho(\theta)$ according to the maximization problem (24). V-MPO follows TRPO and PPO to choose a prior such that the new policy is kept close to the previous one, i.e.

$$\rho(\theta) = -\nu \mathbb{E}_{S \sim d^{\pi_{\text{old}}}} [D_{KL}(\pi_{\text{old}}(\cdot | S) \| \pi_{\theta}(\cdot | S))],$$

with learnable parameter ν . However, optimizing the resulting sample-based maximum likelihood objective directly tends to result in overfitting. Hence, a sequence of transformations is applied. First, the prior is transformed into a hard constraint on the KL divergence when optimizing the policy loss. To employ gradient-based optimization, we use Lagrangian relaxation to transform this constrained optimization problem back into an unconstrained problem and use a coordinate-descent strategy to simultaneously optimize for θ and ν . This can equivalently be written via the stop-gradient operator yielding the trust-region loss (22).

Algorithm 8 V-MPO

```

Require:  $\eta \in \mathbb{R}, \nu \in \mathbb{R}, \varepsilon_\eta \in \mathbb{R}, \varepsilon_\nu \in \mathbb{R}, U \in \mathbb{N}, T \in \mathbb{N}$ 
    Initialize  $\theta$  and  $\phi$  at random
     $t \leftarrow 0$ 
    while  $t \leq T$  do
        for  $i = 1, \dots, U$  do
             $a \sim \pi_{\theta}$  ▷ sample action
             $\beta(a | s) \leftarrow \pi_{\theta}(a | s)$ 
             $s, r \sim P(s, a)$  ▷ sample next state and reward
             $t \leftarrow t + 1$ 
            Store  $(a, s, r, \beta(a | s))$  in  $\mathcal{D}$ 
        end for
        for all epochs do
            Compute returns  $R$  and advantages  $A$ 
            Compute  $\tilde{\mathcal{D}}$ 
             $L_\nu \leftarrow \frac{1}{|\tilde{\mathcal{D}}|} \sum_{\tilde{\mathcal{D}}} \nu (\varepsilon_\nu - \text{sg}(D_{KL}(\pi_{\text{old}} \| \pi_{\theta}))) + \text{sg}(\nu) D_{KL}(\pi_{\text{old}} \| \pi_{\theta})$  ▷ KL loss
             $L_\pi \leftarrow -\frac{1}{|\tilde{\mathcal{D}}|} \sum_{\tilde{\mathcal{D}}} \ln \pi_{\theta}(a | s) \psi(s, a)$  ▷ Policy loss
             $L_\eta \leftarrow \eta \varepsilon_\eta + \eta \ln \left( \frac{1}{|\tilde{\mathcal{D}}|} \sum_{\tilde{\mathcal{D}}} \exp \frac{A}{\eta} \right)$  ▷ Temperature loss
             $d\theta \leftarrow \nabla_{\theta}(L_\pi + L_\nu), d\eta \leftarrow \frac{\partial}{\partial \eta} L_\eta, d\nu \leftarrow \frac{\partial}{\partial \nu} L_\nu$  ▷ Compute gradients
             $d\phi \leftarrow \frac{1}{|\tilde{\mathcal{D}}|} \sum_{\tilde{\mathcal{D}}} \nabla_{\phi}(R - V_{\phi}(s))^2$ 
            update  $\theta, \eta, \nu$  and  $\phi$  using  $d\theta, d\eta, d\nu$  and  $d\phi$  via gradient ascent / descent
        end for
    end while

```

The learnable parameters η and ν are Lagrangian multipliers and hence must be positive. We enforce this by projecting the computed values to small positive values η_{\min} and ν_{\min} respectively if necessary. The pseudocode for V-MPO is depicted in Algorithm 8. As implementational details, V-MPO typically uses decoupled KL constraints for the mean and covariance of the policy in continuous action spaces following [2]. This enables better exploration without moving the policy mean as well as fast learning by rapidly changing the mean without resulting in a collapse of the policy due to vanishing standard deviations. In addition, V-MPO can be used with an off-policy correction via an importance sampling ratio similarly to TRPO and PPO and uses multiple actors following A3C.

4.6 Comparing Design Choices in Policy Gradient Algorithms

Having outlined the main on-policy policy gradient algorithms, we want to shortly compare them to characterize the main design choices in constructing such algorithms.

The predominant differences across policy gradient algorithms lie in the estimators $\hat{\nabla}_{\theta} J(\theta)$ of the policy gradients. We summarize these estimates in Table 1⁸. The algorithms can be distinguished along several dimensions with respects to the gradients. First, they use different variance reduction techniques, which are especially reflected in how Q_{π} in the policy gradient formula (13) is estimated. Second, various policy regularization strategies are used. Third, the algorithms employ further lower-level details to stabilize learning. We will discuss each of these dimensions in the following.

⁸For V-MPO, we focus on the policy loss, thus ignoring the gradient of the KL loss \mathcal{L}_ν w.r.t. the policy parameters θ here.

Algorithm	Gradient estimator
REINFORCE	$G \nabla_{\theta} \ln \pi_{\theta}(a s)$
A3C	$\frac{1}{ \mathcal{D} } \sum_{\mathcal{D}} \hat{A} \nabla_{\theta} \ln \pi_{\theta}(a s)$
TRPO	$\frac{1}{ \mathcal{D} } \sum_{\mathcal{D}} \hat{A} \nabla_{\theta} \frac{\pi_{\theta}(a s)}{\pi_{\text{old}}(a s)}$
PPO	$\frac{1}{ \mathcal{D} } \sum_{\mathcal{D}} \hat{A} \nabla_{\theta} \min \left(\frac{\pi_{\theta}(a s)}{\pi_{\text{old}}(a s)}, \text{clip} \left(\frac{\pi_{\theta}(a s)}{\pi_{\text{old}}(a s)}, 1 - \varepsilon, 1 + \varepsilon \right) \right)$
V-MPO	$\frac{1}{\sum_{\tilde{\mathcal{D}}} \exp \left(\frac{\hat{A}}{\eta} \right)} \sum_{\tilde{\mathcal{D}}} \exp \left(\frac{\hat{A}}{\eta} \right) \nabla_{\theta} \ln \pi_{\theta}(a s)$

Table 1: Policy gradient estimates used by various policy gradient algorithms.

Reducing variance is important to stabilize learning and speed up convergence [77]. However, while high variance means algorithms require more samples to converge, bias in the estimates is not resolvable even with infinite samples [70]. All contemporary policy gradient algorithms, i.e. all presented algorithms except REINFORCE, make use of baselines to reduce variance as discussed in Section 3.2 when approximating the unknown Q_{π} . While REINFORCE samples returns as an unbiased but high-variance estimate [75], the other algorithms learn a value function \hat{V} to estimate the advantage function \hat{A} . Notably, this reduces variance at the cost of introducing bias [43, 76, 70]. Further differences arise from how advantages are estimated, albeit these strategies can be easily transferred between algorithms. PPO uses GAE to estimate advantages, which generalizes the n-step temporal difference estimates used in A3C, TRPO and V-MPO. In addition, V-MPO scales the advantages via the learned temperature η and only uses the top 50 % of advantages per batch.

To stabilize learning beyond variance reduction, several regularization techniques are proposed by TRPO, PPO and V-MPO to limit the change in policies across iterations. TRPO imposes a constraint on the KL divergence between the newly learned and the previous policy. Thus, the policy gradients are not directly applied to update the policy parameters θ but instead they are postprocessed to yield an approximate solution to this constrained optimization problem. This comes at the cost of algorithmic complexity however. Whereas the other algorithms directly compute the estimated policy gradients via automatic differentiation [86, 53], TRPO requires the estimation of a hessian and the application of the conjugate gradient algorithm followed by a line search. PPO avoids such complexity by introducing a heuristic, which bounds the probability ratio $\frac{\pi_{\theta}(a|s)}{\pi_{\text{old}}(a|s)}$, into its objective function. By conservatively clipping this ratio, large policy changes induced by overfitting the advantage function are prevented. This also enables PPO to conduct multiple updates on the same data to accelerate learning [71]. V-MPO too limits the KL divergence across policies. The prior distribution is selected such that V-MPO arrives at a similar optimization problem with a penalty on the KL divergence as TRPO. Following TRPO, V-MPO transforms this into a constrained optimization problem, albeit now with the goal of automatically tuning the penalty parameter by applying coordinate-descent to the Lagrangian relaxation of this constrained optimization problem.

Lastly, we point out that different lower-level details are employed by the discussed algorithm. Except for REINFORCE, all algorithms use several actors, which are potentially updated asynchronously, and average gradients over batches of transitions to further reduce their variance. Here, V-MPO slightly diverges from the other algorithms as it computes a weighted average based on the advantages of the transitions, i.e. with weights

$$\frac{\exp \left(\frac{A_{\pi_{\theta_{\text{old}}}}(s,a)}{\eta} \right)}{\sum_{a,s \in \tilde{\mathcal{D}}} \exp \left(\frac{A_{\pi_{\theta_{\text{old}}}}(s,a)}{\eta} \right)}.$$

A3C and PPO commonly use an entropy bonus to prevent premature convergence to a suboptimal policy by incentivizing a higher standard deviation of the Gaussian output by the policy. We observe that V-MPO does not use an entropy bonus but achieves a comparable effect in continuous action spaces by constraining the policy mean and standard deviation separately. Lastly, TRPO and PPO include an importance sampling correction to compensate for the slight off-policy nature of the algorithms induced by using multiple asynchronous workers. This is also mentioned as an option for V-MPO [73].

5 Convergence Results

In this section, we discuss convergence results for policy gradient algorithms from literature. First, we present an overview of different convergence proofs in Section 5.1. Then, we thoroughly present one selected result in Section 5.2.

5.1 Literature Overview

Several convergence results have been proposed for policy gradient algorithms. They differ along various dimensions: the specific algorithms covered, the shown strength of convergence, the problem settings and the employed proving techniques. The following overview of convergence results is not intended to be complete but rather shall showcase these differences.

As previously discussed, REINFORCE uses an unbiased estimator of the policy gradients, which in expectation therefore point in the direction of the true gradients. Hence, under common stochastic approximation assumptions towards the step sizes, REINFORCE can be shown to converge to a locally optimal policy [75]. In the general form of policy gradient algorithms, agnostic to the specific estimator of Q_π , showing convergence is more complex as the estimated gradients are typically biased when using a learned value function [43, 76]. [4] and [11] consider the simplest case where state and action spaces \mathcal{S} and \mathcal{A} are finite and no function approximation is used, i.e. the policy uses a tabular parameterization with one parameter for each state-action combination. Using that the exact policy gradients can be calculated in this case, both studies show the global convergence to an optimal policy with a linear convergence rate.

[76] and [90] generalize these results to settings with function approximation, albeit under impractical conditions on the approximators, which are required to be linear in their inputs. The extension to function approximation comes at the cost of only being able to proof local convergence using stochastic approximation and the Supermartingale Convergence Theorem [64].

Finally, some convergence results exist for the specific algorithms such as TRPO and PPO. TRPO is based on Theorem 4.1, which comes with monotonic improvement guarantees. However, TRPO is only an approximation to the algorithm stemming from Theorem 4.1, so that no such guarantee holds for TRPO in practice. We further remark that PPO is similarly intended as an heuristic of this theoretical algorithm [71]. Nonetheless, efforts have been made to proof the convergence of these practical algorithms. [51] show that a slightly modified version of PPO converges to a globally optimal policy at a sublinear rate under specific assumptions. In particular, they require an overparameterized neural network as the function approximator such that they can use infinite-dimensional mirror-descent [8] to proof the convergence. [34] provide a proof using two time-scale stochastic approximation [39] that PPO converges to a locally optimal policy under more realistic assumptions akin to typical learning scenarios.

5.2 Mirror Learning

In this subsection, we focus on the convergence proof provided by [46]. While primarily of theoretical interest, we choose to discuss this particular result as it is agnostic to the selected algorithm and parameterization and can hence be applied to a range of policy gradient algorithms. [46] introduce a framework called *mirror learning*, which comes with global convergence guarantees for all policy gradient algorithms that adhere to a specific form. In the following, we follow [46] in deriving their results. We start by giving some definitions, based on which we then present the general form of mirror learning updates. We show that the discussed algorithms largely adhere to this form. Finally, we proof that this implies the convergence to an optimal policy.

5.2.1 Fundamentals of Mirror Learning

From here onwards, we do not explicitly write down the policy parameters, i.e. we omit the subscript θ when describing a policy π . [46] define the drift \mathfrak{D} and the neighborhood operator \mathcal{N} as follows.

Definition 5.1. (*Drift*) A drift functional

$$\mathfrak{D}: \Pi \times \mathcal{S} \rightarrow \{\mathfrak{D}_\pi(\cdot | s) : \Delta(\mathcal{A}) \rightarrow \mathbb{R}\}$$

is a map which satisfies the following conditions for all $s \in \mathcal{S}$ and $\pi, \bar{\pi} \in \Pi$:

1. $\mathfrak{D}_\pi(\bar{\pi} | s) \geq \mathfrak{D}_\pi(\pi | s) = 0$, (non-negativity)
2. $\mathfrak{D}_\pi(\bar{\pi} | s)$ has zero gradient with respects to $\bar{\pi}(\cdot | s)$ at $\bar{\pi}(\cdot | s) = \pi(\cdot | s)$, more precisely all its Gâteaux derivatives⁹ are zero, (zero gradient)

where we used $\mathfrak{D}_\pi(\bar{\pi}(\cdot | s) | s) := \mathfrak{D}_\pi(\bar{\pi} | s)$.

For any state distribution $\nu_{\bar{\pi}}^{\bar{\pi}} \in \Delta(\mathcal{S})$, that can depend on both $\bar{\pi}$ and π , the drift from $\bar{\pi}$ to π is given by

$$\mathfrak{D}_\pi^\nu(\bar{\pi}) := \mathbb{E}_{s \sim \nu_{\bar{\pi}}^{\bar{\pi}}} [\mathfrak{D}_\pi(\bar{\pi} | s)].$$

⁹See Definition D.16.

We require $\nu_{\pi}^{\bar{\pi}}$ to be such that this expectation is continuous in $\bar{\pi}$ and π . We call a drift trivial if $\mathcal{D}_{\pi}^{\nu}(\bar{\pi}) = 0$ for all $\pi, \bar{\pi} \in \Pi$.

Definition 5.2. (Neighborhood Operator) A mapping

$$\mathcal{N}: \Pi \rightarrow \mathcal{P}(\Pi)$$

is a neighborhood operator if it satisfies the following conditions:

1. \mathcal{N} is continuous, (continuity)
2. $\mathcal{N}(\pi)$ is compact for all $\pi \in \Pi$, (compactness)
3. There exists a metric $\chi: \Pi \times \Pi \rightarrow \mathbb{R}$ such that $\chi(\pi, \bar{\pi}) \leq \zeta$ implies $\bar{\pi} \in \mathcal{N}(\pi)$ for all $\pi, \bar{\pi} \in \Pi$ given some $\zeta \in \mathbb{R}_+$. (closed ball)

We call $\mathcal{N}(\cdot) = \Pi$ the trivial neighborhood operator.

With these definitions, we can define the mirror learning update rule.

Definition 5.3. (Mirror Learning Update) Let π_{old} be the previous policy and $d^{\pi_{old}}$ the state distribution under π_{old} . Further, let

$$[\mathcal{M}_{\mathfrak{D}}^{\bar{\pi}} V_{\pi}] (s) := \mathbb{E}_{A \sim \bar{\pi}} [Q_{\pi}(s, A)] - \frac{\nu_{\pi}^{\bar{\pi}}}{d^{\bar{\pi}}} \mathcal{D}_{\pi}(\bar{\pi} \mid s)$$

be the mirror learning operator. Then, the mirror learning update chooses the new policy π_{new} as

$$\pi_{new} \in \arg \max_{\bar{\pi} \in \mathcal{N}(\pi_{old})} \mathbb{E}_{S \sim d^{\pi_{old}}} \left[[\mathcal{M}_{\mathfrak{D}}^{\bar{\pi}} V_{\pi_{old}}] (S) \right]. \quad (26)$$

Under the light of this mirror learning update, the drift \mathfrak{D} from one policy to the next induces some penalty on the objective while the neighborhood operator puts a hard constraint on the divergence of subsequent policies.

5.2.2 Policy Gradient Algorithms as Instances of Mirror Learning

Before proving the convergence of mirror learning to an optimal policy, we first show that the discussed policy gradient algorithms can partly be seen as instances of mirror learning, i.e. use updates of the form

$$\pi_{new} \in \arg \max_{\pi \in \mathcal{N}(\pi_{old})} \mathbb{E}_{S \sim d^{\pi_{old}}} \left[\mathbb{E}_{A \sim \pi} [Q_{\pi_{old}}(S, A)] - \frac{\nu_{\pi_{old}}^{\pi}}{d^{\pi_{old}}} \mathcal{D}_{\pi_{old}}(\pi \mid S) \right].$$

A3C

A3C is a direct application of the Policy Gradient Theorem, albeit with a learned advantage function. Thus, at each iteration it approximately solves the optimization problem

$$\pi_{new} \in \arg \max_{\pi \in \Pi} \mathbb{E}_{S \sim d^{\pi_{old}}, A \sim \pi} \left[A_{\pi_{old}}(S, A) \right] = \arg \max_{\pi \in \Pi} \mathbb{E}_{S \sim d^{\pi_{old}}} \left[\mathbb{E}_{A \sim \pi} [Q_{\pi_{old}}(S, A)] \right].$$

This is the most trivial instantiation of mirror learning by using the trivial drift $\mathfrak{D}(\cdot \mid \cdot) = 0$ and the trivial neighborhood operator $\mathcal{N}(\cdot) = \Pi$. The same argumentation also applies to REINFORCE. Note that in practice however, we maximize the expectation over π_{old} rather than π . For this reason, these are not exact instances of mirror learning.

TRPO

TRPO's constrained optimization problems

$$\begin{aligned} \pi_{new} &\in \arg \max_{\pi \in \Pi} \mathbb{E}_{S \sim d^{\pi_{old}}, A \sim \pi} \left[\frac{\pi(A \mid S)}{\pi_{old}(A \mid S)} A_{\pi_{old}}(S, A) \right] \\ \text{subject to } & \mathbb{E}_{S \sim d^{\pi_{old}}} [D_{KL}(\pi_{old}(\cdot \mid S) \parallel \pi(\cdot \mid S))] \leq \delta \end{aligned}$$

can be rewritten as

$$\pi_{new} \in \arg \max_{\pi \in \mathcal{N}_{\text{TRPO}}(\pi_{old})} \mathbb{E}_{S \sim d^{\pi_{old}}} \left[\mathbb{E}_{A \sim \pi} [Q_{\pi_{old}}(S, A)] \right]$$

with the average-KL ball as the neighborhood operator, i.e.

$$\mathcal{N}_{\text{TRPO}}(\pi_{old}) = \{ \pi \mid \mathbb{E}_{S \sim d^{\pi_{old}}} [D_{KL}(\pi_{old}(\cdot \mid S) \parallel \pi(\cdot \mid S))] \leq \delta \}.$$

Here, we used that

$$\begin{aligned}\mathbb{E}_{A \sim \pi_{\text{old}}} \left[\frac{\pi(A | s)}{\pi_{\text{old}}(A | s)} A_{\pi_{\text{old}}}(s, A) \right] &= \int_{a \in \mathcal{A}} \pi_{\text{old}}(a | s) \frac{\pi(a | s)}{\pi_{\text{old}}(a | s)} A_{\pi_{\text{old}}}(s, a) da \\ &= \mathbb{E}_{A \sim \pi} [A_{\pi_{\text{old}}}(s, A)]\end{aligned}$$

and that maximizing over the action-value function is identical to maximizing over the advantage function following the discussion in Section 3.2. Thus, TRPO is a mirror learning instance with the trivial drift $\mathfrak{D}(\cdot | \cdot) = 0$.

PPO

Each iteration, PPO searches for

$$\pi_{\text{new}} \in \arg \max_{\pi \in \Pi} \mathbb{E}_{\pi_{\text{old}}} \left[\min \left\{ r_{\pi}(A | S), \text{clip}(r_{\pi}(A | S), 1 - \varepsilon, 1 + \varepsilon) \right\} A_{\pi_{\text{old}}}(S, A) \right],$$

where we write $r_{\pi}(a | s)$ for $\frac{\pi(a | s)}{\pi_{\text{old}}(a | s)}$. We can rewrite the expectation over actions by adding zero as

$$\begin{aligned}&\mathbb{E}_{A \sim \pi_{\text{old}}} \left[\min \left\{ r_{\pi}(A | s) A_{\pi_{\text{old}}}(s, A), \text{clip}(r_{\pi}(A | s), 1 - \varepsilon, 1 + \varepsilon) A_{\pi_{\text{old}}}(s, A) \right\} \right] \\ &= \mathbb{E}_{A \sim \pi_{\text{old}}} \left[r_{\pi}(A | s) A_{\pi_{\text{old}}}(s, A) \right] - \mathbb{E}_{A \sim \pi_{\text{old}}} \left[r_{\pi}(A | s) A_{\pi_{\text{old}}}(s, A) \right. \\ &\quad \left. - \min \left\{ r_{\pi}(A | s) A_{\pi_{\text{old}}}(s, A), \text{clip}(r_{\pi}(A | s), 1 - \varepsilon, 1 + \varepsilon) A_{\pi_{\text{old}}}(s, A) \right\} \right].\end{aligned}$$

Using the same technique as before, we can write the first expectation equivalently as $\mathbb{E}_{A \sim \pi} [A_{\pi_{\text{old}}}(s, A)]$. We now focus on the second expectation. We replace the min operator with a max and push the first term inside the max to obtain

$$\begin{aligned}&\mathbb{E}_{A \sim \pi_{\text{old}}} \left[r_{\pi}(A | s) A_{\pi_{\text{old}}}(s, A) - \min \left\{ r_{\pi}(A | s) A_{\pi_{\text{old}}}(s, A), \text{clip}(r_{\pi}(A | s), 1 - \varepsilon, 1 + \varepsilon) A_{\pi_{\text{old}}}(s, A) \right\} \right] \\ &= \mathbb{E}_{A \sim \pi_{\text{old}}} \left[r_{\pi}(A | s) A_{\pi_{\text{old}}}(s, A) + \max \left\{ -r_{\pi}(A | s) A_{\pi_{\text{old}}}(s, A), -\text{clip}(r_{\pi}(A | s), 1 - \varepsilon, 1 + \varepsilon) A_{\pi_{\text{old}}}(s, A) \right\} \right] \\ &= \mathbb{E}_{A \sim \pi_{\text{old}}} \left[\max \left\{ r_{\pi}(A | s) A_{\pi_{\text{old}}}(s, A) - r_{\pi}(A | s) A_{\pi_{\text{old}}}(s, A), \right. \right. \\ &\quad \left. \left. r_{\pi}(A | s) A_{\pi_{\text{old}}}(s, A) - \text{clip}(r_{\pi}(A | s), 1 - \varepsilon, 1 + \varepsilon) A_{\pi_{\text{old}}}(s, A) \right\} \right] \\ &= \mathbb{E}_{A \sim \pi_{\text{old}}} \left[\max \left\{ 0, (r_{\pi}(A | s) - \text{clip}(r_{\pi}(A | s), 1 - \varepsilon, 1 + \varepsilon)) A_{\pi_{\text{old}}}(s, A) \right\} \right].\end{aligned}$$

This final expression is non-negative. Moreover, it is zero for π sufficiently close to π_{old} , i.e. such that for all actions $a \in \mathcal{A}$ we have $r_{\pi}(a | s) = \frac{\pi(a | s)}{\pi_{\text{old}}(a | s)} \in [1 - \varepsilon, 1 + \varepsilon]$, because then the clip-function reduces to the identity function w.r.t. its first argument. Thus, the derivatives of this expression must also be zero at $\pi(\cdot | s) = \pi_{\text{old}}(\cdot | s)$. These properties are the exact conditions for a mapping to be considered a drift in the sense of Definition 5.1. With this preparation, we can now write the PPO update as

$$\pi_{\text{new}} \in \arg \max_{\pi \in \Pi} \mathbb{E}_{S \sim d^{\pi_{\text{old}}}} \left[\mathbb{E}_{A \sim \pi} [A_{\pi_{\text{old}}}(S, A)] - \mathfrak{D}_{\pi_{\text{old}}}(\pi | S) \right],$$

where $\mathfrak{D}_{\pi_{\text{old}}}$ is a drift given by

$$\mathfrak{D}_{\pi_{\text{old}}}(\pi | s) = \mathbb{E}_{A \sim \pi_{\text{old}}} \left[\max \left\{ 0, (r_{\pi}(A | s) - \text{clip}(r_{\pi}(A | s), 1 - \varepsilon, 1 + \varepsilon)) A_{\pi_{\text{old}}}(s, A) \right\} \right].$$

This is an instance of the mirror learning update with the trivial neighborhood operator $\mathcal{N}(\cdot) = \Pi$ and $\nu_{\pi_{\text{old}}}^{\pi} = d^{\pi_{\text{old}}}$.

5.2.3 Convergence Proof

Now, we present the main theoretical result of [46].

Theorem 5.4. *Let \mathfrak{D}^ν be a drift, \mathcal{N} a neighborhood operator and d^π the sampling distribution, all continuous in π . Let the objective, i.e. the expected returns under a policy π , be written as $J(\pi) = \mathbb{E}_{S_0 \sim p_0, \pi}[G_0]$. Let $\pi_0 \in \Pi$ be the initial policy and the sequence of policies $(\pi_n)_{n=0}^\infty$ be obtained through the mirror learning update rule (26) under \mathfrak{D}^ν , \mathcal{N} and d^π . Then,*

1. (Strict monotonic improvement)

$$J(\pi_{n+1}) \geq J(\pi_n) + \mathbb{E}_{S \sim p_0} \left[\frac{\nu_{\pi_n}^{\pi_{n+1}}(S)}{d^{\pi_n}(S)} \mathfrak{D}_{\pi_n}(\pi_{n+1} | S) \right] \quad \forall n \in \mathbb{N}_0.$$

2. (Value function optimality)

$$\lim_{n \rightarrow \infty} V_{\pi_n} = V^*.$$

3. (Maximum attainable return)

$$\lim_{n \rightarrow \infty} J(\pi_n) = \max_{\pi \in \Pi} J(\pi).$$

4. (Policy optimality)

$$\lim_{n \rightarrow \infty} \pi_n = \pi^*.$$

Proof. We structure the proof by [46] in five steps. In step 1, we start by showing that mirror learning updates lead to improvements under the mirror learning operator $\mathcal{M}_{\mathfrak{D}}^{\pi_n} V_{\pi_{n-1}}$, which implies improvements in the value function V_{π_n} . In step 2, we prove that the sequence of value functions $(V_{\pi_n})_{n=0}^\infty$ converges to some limit. In step 3, we show the existence of limit points of the sequence of policies $(\pi_n)_{n=0}^\infty$, which are fixed points of the mirror learning update (26). In step 4, we prove that these limit points are also fixed points of Generalized Policy Iteration (GPI) [75], from which we conclude that these limit points are optimal policies in step 5. For simplicity, we proof Theorem 5.4 for discrete state and actions spaces. However, the results are straightforward to extended to the continuous cases (see the appendix in [46] for details).

Step 1

We start by showing by contradiction that for all $n \in \mathbb{N}_0$ and for all $s \in \mathcal{S}$:

$$[\mathcal{M}_{\mathfrak{D}}^{\pi_{n+1}} V_{\pi_n}](s) \geq [\mathcal{M}_{\mathfrak{D}}^{\pi_n} V_{\pi_n}](s). \quad (27)$$

Suppose there exists $s_0 \in \mathcal{S}$, which violates (27). We define a policy $\hat{\pi}$ with

$$\hat{\pi}(\cdot | s) = \begin{cases} \pi_{n+1}(\cdot | s) & \text{if } s \neq s_0, \\ \pi_n(\cdot | s) & \text{if } s = s_0. \end{cases}$$

This way, we guarantee $\hat{\pi} \in \mathcal{N}(\pi_n)$ because $\pi_{n+1} \in \mathcal{N}(\pi_n)$ is forced by the mirror learning update (26) and the distance between $\hat{\pi}$ and π_n is similar to the distance between π_{n+1} and π_n at every $s \neq s_0$ but smaller at $s = s_0$.

By assumption, we have at s_0 that

$$\begin{aligned} [\mathcal{M}_{\mathfrak{D}}^{\hat{\pi}} V_{\pi_n}](s_0) &= \mathbb{E}_{A \sim \hat{\pi}} [Q_{\pi_n}(s_0, A)] - \frac{\nu_{\pi_n}^{\hat{\pi}}}{d^{\pi_n}} \mathfrak{D}_{\pi_n}(\hat{\pi} | s_0) \\ &= \mathbb{E}_{A \sim \pi_n} [Q_{\pi_n}(s_0, A)] - \frac{\nu_{\pi_n}^{\pi_n}}{d^{\pi_n}} \mathfrak{D}_{\pi_n}(\pi_n | s_0) \\ &= [\mathcal{M}_{\mathfrak{D}}^{\pi_n} V_{\pi_n}](s_0) \\ &> [\mathcal{M}_{\mathfrak{D}}^{\pi_{n+1}} V_{\pi_n}](s_0). \end{aligned}$$

It follows that

$$\mathbb{E}_{S \sim d^{\pi_n}} \left[[\mathcal{M}_{\mathfrak{D}}^{\hat{\pi}} V_{\pi_n}](S) \right] - \mathbb{E}_{S \sim d^{\pi_n}} \left[[\mathcal{M}_{\mathfrak{D}}^{\pi_{n+1}} V_{\pi_n}](S) \right] = d^{\pi_n}(s_0) \left([\mathcal{M}_{\mathfrak{D}}^{\hat{\pi}} V_{\pi_n}](s_0) - [\mathcal{M}_{\mathfrak{D}}^{\pi_{n+1}} V_{\pi_n}](s_0) \right) > 0,$$

where we used that $[\mathcal{M}_{\mathfrak{D}}^{\hat{\pi}} V_{\pi_n}](s) = [\mathcal{M}_{\mathfrak{D}}^{\pi_{n+1}} V_{\pi_n}](s)$ for $s \neq s_0$. Thus,

$$\mathbb{E}_{S \sim d^{\pi_n}} \left[[\mathcal{M}_{\mathfrak{D}}^{\hat{\pi}} V_{\pi_n}](S) \right] > \mathbb{E}_{S \sim d^{\pi_n}} \left[[\mathcal{M}_{\mathfrak{D}}^{\pi_{n+1}} V_{\pi_n}](S) \right],$$

which contradicts the mirror learning update rule, i.e. that

$$\mathbb{E}_{S \sim d^{\pi_n}} \left[[\mathcal{M}_{\mathfrak{D}}^{\pi_{n+1}} V_{\pi_n}](S) \right] = \max_{\bar{\pi} \in \mathcal{N}(\pi_n)} \mathbb{E}_{S \sim d^{\pi_n}} \left[[\mathcal{M}_{\mathfrak{D}}^{\bar{\pi}} V_{\pi_n}](S) \right].$$

Hence, we have shown that the sequence of policies $(\pi_n)_{n=0}^\infty$ created by the mirror learning updates monotonically increases the mirror learning operator at every state. Next, we show that this property, i.e. $[\mathcal{M}_{\mathfrak{D}}^{\pi_{n+1}} V_{\pi_n}](s) \geq [\mathcal{M}_{\mathfrak{D}}^{\pi_n} V_{\pi_n}](s)$, implies the monotonic improvement in the value function

$$V_{\pi_{n+1}}(s) \geq V_{\pi_n}(s) \quad (28)$$

for all $s \in \mathcal{S}$ and $n \in \mathbb{N}_0$.

By using the definitions of the value function V_π , the action-value function Q_π , the mirror learning operator $\mathcal{M}_{\mathfrak{D}}^{\bar{\pi}} V_\pi$ and the identity $\mathfrak{D}_\pi(\pi | s) = 0$, adding zeros and rearranging, we obtain

$$\begin{aligned} V_{\pi_{n+1}}(s) - V_{\pi_n}(s) &= \mathbb{E}_{\pi_{n+1}} \left[R + \gamma V_{\pi_{n+1}}(S') \right] - \mathbb{E}_{\pi_n} \left[R + \gamma V_{\pi_n}(S') \right] \\ &= \mathbb{E}_{\pi_{n+1}} \left[R + \gamma V_{\pi_{n+1}}(S') \right] - \mathbb{E}_{\pi_n} \left[R + \gamma V_{\pi_n}(S') \right] \\ &\quad + \frac{\nu_{\pi_n}^{\pi_{n+1}}(s)}{d^{\pi_n}(s)} \mathfrak{D}_{\pi_n}(\pi_{n+1} | s) - \frac{\nu_{\pi_n}^{\pi_{n+1}}(s)}{d^{\pi_n}(s)} \mathfrak{D}_{\pi_n}(\pi_{n+1} | s) \\ &= \mathbb{E}_{\pi_{n+1}} \left[R + \gamma V_{\pi_{n+1}}(S') + \gamma V_{\pi_n}(S') - \gamma V_{\pi_n}(S') \right] - \mathbb{E}_{\pi_n} \left[R + \gamma V_{\pi_n}(S') \right] \\ &\quad + \frac{\nu_{\pi_n}^{\pi_{n+1}}(s)}{d^{\pi_n}(s)} \mathfrak{D}_{\pi_n}(\pi_{n+1} | s) - \frac{\nu_{\pi_n}^{\pi_{n+1}}(s)}{d^{\pi_n}(s)} \mathfrak{D}_{\pi_n}(\pi_{n+1} | s) \\ &= \left(\mathbb{E}_{\pi_{n+1}} \left[R + \gamma V_{\pi_n}(S') \right] - \frac{\nu_{\pi_n}^{\pi_{n+1}}(s)}{d^{\pi_n}(s)} \mathfrak{D}_{\pi_n}(\pi_{n+1} | s) \right) \\ &\quad - \left(\mathbb{E}_{\pi_n} \left[R + \gamma V_{\pi_n}(S') \right] - \frac{\nu_{\pi_n}^{\pi_n}(s)}{d^{\pi_n}(s)} \mathfrak{D}_{\pi_n}(\pi_n | s) \right) \\ &\quad + \gamma \mathbb{E}_{\pi_{n+1}} \left[V_{\pi_{n+1}}(S') - V_{\pi_n}(S') \right] + \frac{\nu_{\pi_n}^{\pi_{n+1}}(s)}{d^{\pi_n}(s)} \mathfrak{D}_{\pi_n}(\pi_{n+1} | s) \\ &= \left(\mathbb{E}_{\pi_{n+1}} \left[Q_{\pi_n}(s, A) \right] - \frac{\nu_{\pi_n}^{\pi_{n+1}}(s)}{d^{\pi_n}(s)} \mathfrak{D}_{\pi_n}(\pi_{n+1} | s) \right) \\ &\quad - \left(\mathbb{E}_{\pi_n} \left[Q_{\pi_n}(s, A) \right] - \frac{\nu_{\pi_n}^{\pi_n}(s)}{d^{\pi_n}(s)} \mathfrak{D}_{\pi_n}(\pi_n | s) \right) \\ &\quad + \gamma \mathbb{E}_{\pi_{n+1}} \left[V_{\pi_{n+1}}(S') - V_{\pi_n}(S') \right] + \frac{\nu_{\pi_n}^{\pi_{n+1}}(s)}{d^{\pi_n}(s)} \mathfrak{D}_{\pi_n}(\pi_{n+1} | s) \\ &= [\mathcal{M}_{\mathfrak{D}}^{\pi_{n+1}} V_{\pi_n}] - [\mathcal{M}_{\mathfrak{D}}^{\pi_n} V_{\pi_n}] \\ &\quad + \gamma \mathbb{E}_{\pi_{n+1}} \left[V_{\pi_{n+1}}(S') - V_{\pi_n}(S') \right] + \frac{\nu_{\pi_n}^{\pi_{n+1}}(s)}{d^{\pi_n}(s)} \mathfrak{D}_{\pi_n}(\pi_{n+1} | s) \\ &\geq \gamma \mathbb{E}_{\pi_{n+1}} \left[V_{\pi_{n+1}}(S') - V_{\pi_n}(S') \right] + \frac{\nu_{\pi_n}^{\pi_{n+1}}(s)}{d^{\pi_n}(s)} \mathfrak{D}_{\pi_n}(\pi_{n+1} | s), \end{aligned} \quad (29)$$

where we used Inequality (27) in the final step. We take the infimum over states and replace the expectation with another infimum over states as a lower bound:

$$\begin{aligned} \inf_{s \in \mathcal{S}} \left[V_{\pi_{n+1}}(s) - V_{\pi_n}(s) \right] &\geq \inf_{s \in \mathcal{S}} \left[\gamma \mathbb{E}_{\pi_{n+1}} \left[V_{\pi_{n+1}}(S') - V_{\pi_n}(S') \right] + \frac{\nu_{\pi_n}^{\pi_{n+1}}(s)}{d^{\pi_n}(s)} \mathfrak{D}_{\pi_n}(\pi_{n+1} | s) \right] \\ &\geq \inf_{s \in \mathcal{S}} \left[\gamma \inf_{s' \in \mathcal{S}} \left[V_{\pi_{n+1}}(s') - V_{\pi_n}(s') \right] + \frac{\nu_{\pi_n}^{\pi_{n+1}}(s)}{d^{\pi_n}(s)} \mathfrak{D}_{\pi_n}(\pi_{n+1} | s) \right] \\ &= \gamma \inf_{s' \in \mathcal{S}} \left[V_{\pi_{n+1}}(s') - V_{\pi_n}(s') \right] + \inf_{s \in \mathcal{S}} \left[\frac{\nu_{\pi_n}^{\pi_{n+1}}(s)}{d^{\pi_n}(s)} \mathfrak{D}_{\pi_n}(\pi_{n+1} | s) \right]. \end{aligned}$$

From this expression, we obtain

$$\inf_{s \in \mathcal{S}} [V_{\pi_{n+1}}(s) - V_{\pi_n}(s)] \geq \frac{1}{1-\gamma} \inf_{s \in \mathcal{S}} \left[\frac{\nu_{\pi_n}^{\pi_{n+1}}(s)}{d^{\pi_n}(s)} \mathfrak{D}_{\pi_n}(\pi_{n+1} | s) \right] \geq 0,$$

since $\nu_{\pi_n}^{\pi_{n+1}}(s)$ and $d^{\pi_n}(s)$ are probabilities and the drift \mathfrak{D} is non-negative. Thus, we have proven the monotonic improvement of value functions $V_{\pi_{n+1}}(s) \geq V_{\pi_n}(s)$. We observe that this already implies the strict monotonic improvement property

$$J(\pi_{n+1}) \geq J(\pi_n) + \mathbb{E}_{S \sim p_0} \left[\frac{\nu_{\pi_n}^{\pi_{n+1}}(S)}{d^{\pi_n}(S)} \mathfrak{D}_{\pi_n}(\pi_{n+1} | S) \right]$$

for all $n \in \mathbb{N}_0$ since applying (28) and (27) sequentially to (29) yields for all $s \in \mathcal{S}$

$$\begin{aligned} V_{\pi_{n+1}}(s) - V_{\pi_n}(s) &= [\mathcal{M}_{\mathfrak{D}}^{\pi_{n+1}} V_{\pi_n}] - [\mathcal{M}_{\mathfrak{D}}^{\pi_n} V_{\pi_n}] \\ &\quad + \gamma \mathbb{E}_{\pi_{n+1}} [V_{\pi_{n+1}}(S') - V_{\pi_n}(S')] + \frac{\nu_{\pi_n}^{\pi_{n+1}}(s)}{d^{\pi_n}(s)} \mathfrak{D}_{\pi_n}(\pi_{n+1} | s) \\ &\geq [\mathcal{M}_{\mathfrak{D}}^{\pi_{n+1}} V_{\pi_n}] - [\mathcal{M}_{\mathfrak{D}}^{\pi_n} V_{\pi_n}] + \frac{\nu_{\pi_n}^{\pi_{n+1}}(s)}{d^{\pi_n}(s)} \mathfrak{D}_{\pi_n}(\pi_{n+1} | s) \\ &\geq \frac{\nu_{\pi_n}^{\pi_{n+1}}(s)}{d^{\pi_n}(s)} \mathfrak{D}_{\pi_n}(\pi_{n+1} | s). \end{aligned}$$

We obtain the desired inequality by taking the expectation over $S \sim p_0$.

Step 2

From step 1, we know that the value functions increase uniformly over the state space, i.e. $V_{\pi_{n+1}}(s) - V_{\pi_n}(s) \geq 0$, for all $s \in \mathcal{S}$, $n \in \mathbb{N}_0$. As the rewards r are bounded by assumption and we consider the episodic case where episode lengths are also bounded by T (albeit the same argument applies for infinite time horizons via discounting), the value functions $V_{\pi}(s) = \mathbb{E}_{\pi} [\sum_{k=0}^T \gamma^k R_{t+k+1} | S_t = s]$ are also uniformly bounded. Via the Monotone Convergence Theorem (Theorem D.13), the sequence of value functions $(V_{\pi_n})_{n=0}^{\infty}$ must therefore converge to some limit V .

Step 3

Now, we show the existence of limit points of the sequence of policies $(\pi_n)_{n=0}^{\infty}$ and prove by contradiction that these are fixed points of the mirror learning update (26).

The sequence $(\pi_n)_{n=0}^{\infty}$ is bounded, thus the Bolzano-Weierstrass Theorem (Theorem D.14) yields the existence of limits $\bar{\pi}$ to which some respective subsequence $(\pi_{n_i})_{i=0}^{\infty}$ converges. We denote this set of limit points as $L\Pi$. For each element of such a convergent subsequence $(\pi_{n_i})_{i=0}^{\infty}$, mirror learning solves the optimization problem

$$\max_{\pi \in \mathcal{N}(\pi_{n_i})} \mathbb{E}_{S \sim d^{\pi_{n_i}}} [\mathcal{M}_{\mathfrak{D}}^{\pi} V_{\pi_{n_i}}](S) \tag{30}$$

This expression is continuous in π_{n_i} due to the continuity of the value function [45], the drift and neighborhood operator (by definition) and the sampling distribution (by assumption). Let $\bar{\pi} = \lim_{i \rightarrow \infty} \pi_{n_i}$. Berge's Maximum Theorem (Theorem D.15) [6] now guarantees the convergence of the above expression, yielding

$$\lim_{i \rightarrow \infty} \max_{\pi \in \mathcal{N}(\pi_{n_i})} \mathbb{E}_{S \sim d^{\pi_{n_i}}} [\mathcal{M}_{\mathfrak{D}}^{\pi} V_{\pi_{n_i}}](S) = \max_{\pi \in \mathcal{N}(\bar{\pi})} \mathbb{E}_{S \sim d^{\bar{\pi}}} [\mathcal{M}_{\mathfrak{D}}^{\pi} V_{\bar{\pi}}](S). \tag{31}$$

For all $i \in \mathbb{N}_0$, we obtain the next policy π_{n_i+1} as the argmax of Expression (30). Since this expression converges to the limit in (31), there must exist some subsequence $(\pi_{n_{i_k}+1})_{k=0}^{\infty}$ of $(\pi_{n_i+1})_{i=0}^{\infty}$ which converges to some policy π' , which is the solution to the optimization problem (31). We now show by contradiction that $\pi' = \bar{\pi}$, which implies that $\bar{\pi}$ is a fixed point of the mirror learning update rule.

Suppose $\pi' \neq \bar{\pi}$. As π' is induced by the mirror learning update rule, the monotonic improvement results from step 1 yield

$$Q_{\pi'}(s, a) = \mathbb{E}_{R, S' \sim P} [R + \gamma V_{\pi'}(S')] \geq \mathbb{E}_{R, S' \sim P} [R + \gamma V_{\bar{\pi}}(S')] = Q_{\bar{\pi}}(s, a) \tag{32}$$

and

$$[\mathcal{M}_{\mathfrak{D}}^{\pi'} V_{\bar{\pi}}](s) \geq [\mathcal{M}_{\mathfrak{D}}^{\bar{\pi}} V_{\bar{\pi}}](s).$$

Suppose

$$\mathbb{E}_{S \sim d^{\bar{\pi}}} [\mathcal{M}_{\mathfrak{D}}^{\pi'} V_{\bar{\pi}}](S) > \mathbb{E}_{S \sim d^{\bar{\pi}}} [\mathcal{M}_{\mathfrak{D}}^{\bar{\pi}} V_{\bar{\pi}}](S),$$

then we have for some state s

$$\begin{aligned} [\mathcal{M}_{\mathfrak{D}}^{\pi'} V_{\bar{\pi}}](s) &= \mathbb{E}_{\pi'} [Q_{\bar{\pi}}(s, A)] - \frac{\nu_{\bar{\pi}}^{\pi'}(s)}{d^{\bar{\pi}}(s)} \mathfrak{D}_{\bar{\pi}}(\pi' | s) \\ &> [\mathcal{M}_{\mathfrak{D}}^{\bar{\pi}} V_{\bar{\pi}}](s) = \mathbb{E}_{\bar{\pi}} [Q_{\bar{\pi}}(s, A)] - \frac{\nu_{\bar{\pi}}^{\bar{\pi}}(s)}{d^{\bar{\pi}}(s)} \mathfrak{D}_{\bar{\pi}}(\bar{\pi} | s) \\ &= \mathbb{E}_{\bar{\pi}} [Q_{\bar{\pi}}(s, A)] = V_{\bar{\pi}}(s) = V(s). \end{aligned}$$

In the last equality, we used that the sequence of value functions converges to some unique limit V , which implies $V_{\bar{\pi}} = V$. We obtain the following via this result, Inequality (32), which must be strict for s , and the non-negativity of the drift \mathfrak{D} :

$$\begin{aligned} V_{\pi'}(s) &= \mathbb{E}_{\pi'} [Q_{\pi'}(s, A)] \\ &> \mathbb{E}_{\pi'} [Q_{\bar{\pi}}(s, A)] \\ &> \mathbb{E}_{\pi'} [Q_{\bar{\pi}}(s, A)] - \frac{\nu_{\bar{\pi}}^{\pi'}(s)}{d^{\bar{\pi}}(s)} \mathfrak{D}_{\bar{\pi}}(\pi' | s) \\ &> V(s). \end{aligned}$$

However due to $V_{\pi'}(s) = \lim_{k \rightarrow \infty} V_{\pi_{n_{i_k}+1}}$, this contradicts the uniqueness of the value limit, which gives $V_{\pi'} = V$. Therefore, we have shown by contradiction that

$$\bar{\pi} \in \arg \max_{\pi \in \mathcal{N}(\bar{\pi})} \mathbb{E}_{S \sim d^{\bar{\pi}}} [\mathcal{M}_{\mathfrak{D}}^{\pi} V_{\bar{\pi}}](S).$$

Step 4

Following step 3, let $\bar{\pi}$ be a limit point of $(\pi_n)_{n=0}^\infty$. We will show by contradiction that $\bar{\pi}$ is also a fixed point of GPI (see Theorem 2.1), i.e. that for all $s \in \mathcal{S}$

$$\bar{\pi} \in \arg \max_{\pi \in \Pi} \mathbb{E}_{A \sim \pi} [A_{\bar{\pi}}(s, A)] = \arg \max_{\pi \in \Pi} \mathbb{E}_{A \sim \pi} [Q_{\bar{\pi}}(s, A)]. \quad (33)$$

From step 3, we know that

$$\begin{aligned} \bar{\pi} &\in \arg \max_{\pi \in \Pi} \left[\mathbb{E}_{S \sim d^{\bar{\pi}}, A \sim \pi} \left[Q_{\bar{\pi}}(S, A) - \frac{\nu_{\bar{\pi}}^{\pi}(S)}{d^{\bar{\pi}}(S)} \mathfrak{D}_{\bar{\pi}}(\pi | S) \right] \right] \\ &= \arg \max_{\pi \in \Pi} \left[\mathbb{E}_{S \sim d^{\bar{\pi}}, A \sim \pi} \left[A_{\bar{\pi}}(S, A) - \frac{\nu_{\bar{\pi}}^{\pi}(S)}{d^{\bar{\pi}}(S)} \mathfrak{D}_{\bar{\pi}}(\pi | S) \right] \right] \end{aligned} \quad (34)$$

as subtracting an action-independent baseline does not affect the argmax. Now, we assume the existence of a policy π' and state s with

$$\mathbb{E}_{A \sim \pi'} [A_{\bar{\pi}}(s, A)] > \mathbb{E}_{A \sim \bar{\pi}} [A_{\bar{\pi}}(s, A)] = 0. \quad (35)$$

Let $m = |\mathcal{A}|$ denote the size of the action space. Then, we can write for any policy π , $\pi(\cdot | s) = (x_1, \dots, x_{m-1}, 1 - \sum_{i=1}^{m-1} x_i)$. With this notation, we have

$$\begin{aligned} \mathbb{E}_{A \sim \pi} [A_{\bar{\pi}}(s, A)] &= \sum_{i=1}^m \pi(a_i | s) A_{\bar{\pi}}(s, a_i) \\ &= \sum_{i=1}^{m-1} x_i A_{\bar{\pi}}(s, a_i) + \left(1 - \sum_{i=1}^{m-1} x_i\right) A_{\bar{\pi}}(s, a_m) \\ &= \sum_{i=1}^{m-1} x_i (A_{\bar{\pi}}(s, a_i) - A_{\bar{\pi}}(s, a_m)) + A_{\bar{\pi}}(s, a_m). \end{aligned}$$

This shows that $\mathbb{E}_{A \sim \pi} [A_{\bar{\pi}}(s, A)]$ is an affine function of $\pi(\cdot | s)$, which implies that all its Gâteaux derivatives are constant in $\Delta(\mathcal{A})$ for fixed directions. Due to Inequality (35), this further implies that the Gâteaux derivatives in direction from $\bar{\pi}$ to π' are strictly positive. Additionally, we have that the Gâteaux derivatives of $\frac{\nu_{\bar{\pi}}^{\pi}(s)}{d^{\bar{\pi}}(s)} \mathfrak{D}_{\bar{\pi}}(\pi | s)$ are zero at $\pi = \bar{\pi}$. We see this by establishing lower and upper bounds, which both have derivatives of zero due to the independence of π and the zero-gradient property of the drift:

$$\frac{1}{d^{\bar{\pi}}(s)} \mathfrak{D}_{\bar{\pi}}(\bar{\pi} | s) = \frac{\nu_{\bar{\pi}}^{\bar{\pi}}(s)}{d^{\bar{\pi}}(s)} \mathfrak{D}_{\bar{\pi}}(\bar{\pi} | s) = 0 \leq \frac{\nu_{\bar{\pi}}^{\pi}(s)}{d^{\bar{\pi}}(s)} \mathfrak{D}_{\bar{\pi}}(\pi | s) \leq \frac{1}{d^{\bar{\pi}}(s)} \mathfrak{D}_{\bar{\pi}}(\pi | s)$$

recalling that $\mathfrak{D}_{\bar{\pi}}(\bar{\pi} | s) = 0$ for any $s \in \mathcal{S}$ and using $\nu_{\bar{\pi}}^{\pi}(s) \leq 1$. In combination, we obtain that the Gâteaux derivative of $\mathbb{E}_{A \sim \pi} [A_{\bar{\pi}}(s, A)] - \frac{\nu_{\bar{\pi}}^{\pi}(s)}{d^{\bar{\pi}}(s)} \mathfrak{D}_{\bar{\pi}}(\pi | s)$ is strictly positive as well. Therefore, we can find some policy $\hat{\pi}(\cdot | s)$ by taking a sufficiently small step from $\bar{\pi}(\cdot | s)$ in the direction of $\pi'(\cdot | s)$ such that $\hat{\pi} \in \mathcal{N}(\bar{\pi})$ and

$$\mathbb{E}_{A \sim \hat{\pi}} [A_{\bar{\pi}}(s, A)] - \frac{\nu_{\bar{\pi}}^{\hat{\pi}}(s)}{d^{\bar{\pi}}(s)} \mathfrak{D}_{\bar{\pi}}(\hat{\pi} | s) > \mathbb{E}_{A \sim \bar{\pi}} [A_{\bar{\pi}}(s, A)] - \frac{\nu_{\bar{\pi}}^{\bar{\pi}}(s)}{d^{\bar{\pi}}(s)} \mathfrak{D}_{\bar{\pi}}(\bar{\pi} | s) = 0.$$

With this, we can construct a policy which contradicts Equation (34). Let $\tilde{\pi}$ be defined such that

$$\tilde{\pi}(\cdot | x) = \begin{cases} \bar{\pi}(\cdot | x) & \text{if } x \neq s, \\ \hat{\pi}(\cdot | x) & \text{if } x = s. \end{cases}$$

This guarantees $\tilde{\pi} \in \mathcal{N}(\bar{\pi})$ and

$$\begin{aligned} \mathbb{E}_{S \sim d^{\bar{\pi}}} \left[\mathbb{E}_{A \sim \tilde{\pi}} [A_{\bar{\pi}}(S, A)] - \frac{\nu_{\bar{\pi}}^{\tilde{\pi}}(S)}{d^{\bar{\pi}}(S)} \mathfrak{D}_{\bar{\pi}}(\tilde{\pi} | S) \right] \\ = d^{\bar{\pi}}(s) \left(\mathbb{E}_{A \sim \tilde{\pi}} [A_{\bar{\pi}}(s, A)] - \frac{\nu_{\bar{\pi}}^{\tilde{\pi}}(s)}{d^{\bar{\pi}}(s)} \mathfrak{D}_{\bar{\pi}}(\tilde{\pi} | s) \right) \\ = d^{\bar{\pi}}(s) \left(\mathbb{E}_{A \sim \hat{\pi}} [A_{\bar{\pi}}(s, A)] - \frac{\nu_{\bar{\pi}}^{\hat{\pi}}(s)}{d^{\bar{\pi}}(s)} \mathfrak{D}_{\bar{\pi}}(\hat{\pi} | s) \right) \\ > 0, \end{aligned}$$

which contradicts Equation (34), so the assumption (35) must be wrong, proving

$$\bar{\pi} = \arg \max_{\pi \in \Pi} \mathbb{E}_{A \sim \pi} [A_{\bar{\pi}}(s, A)] = \arg \max_{\pi \in \Pi} \mathbb{E}_{A \sim \pi} [Q_{\bar{\pi}}(s, A)].$$

Step 5

The main result (33) from step 4 shows that any limit point $\bar{\pi}$ of $(\pi_n)_{n \in \mathbb{N}}$ is also a fixed point of GPI. Thus, as corollaries all properties induced by GPI (see Theorem 2.1) apply to $\bar{\pi} \in L\Pi$. Particularly, we have the optimality of $\bar{\pi}$, the value function optimality $V = V_{\bar{\pi}} = V^*$ and thereby also the maximality of returns as

$$\lim_{n \rightarrow \infty} J(\pi_n) = \lim_{n \rightarrow \infty} \mathbb{E}_{S \sim p_0} [V_{\pi_n}(S)] = \mathbb{E}_{S \sim p_0} [V^*(S)] = \max_{\pi \in \Pi} J(\pi).$$

Thus, we have shown all properties as claimed by Theorem 5.4. \square

We close this section with some remarks. In practice, exact updates according to the mirror learning update rule (26) are generally infeasible. Instead, we can sample the expectation to obtain batch estimators over a batch \mathcal{D} of transitions

$$\frac{1}{|\mathcal{D}|} \sum_{s, a \in \mathcal{D}} \left(Q_{\pi_{\text{old}}}(s, a) - \frac{\nu_{\pi_{\text{old}}}^{\pi_{\text{new}}}}{d^{\pi_{\text{old}}}} \mathfrak{D}_{\pi_{\text{old}}}(\pi_{\text{new}} | s) \right),$$

where $Q_{\pi_{\text{old}}}$ has to be estimated as well. These batch estimators can also only be approximately optimized each iteration via gradient ascent to update the policy. Given these approximations and the at-best local convergence of gradient ascent, the outlaid convergence properties remain theoretical.

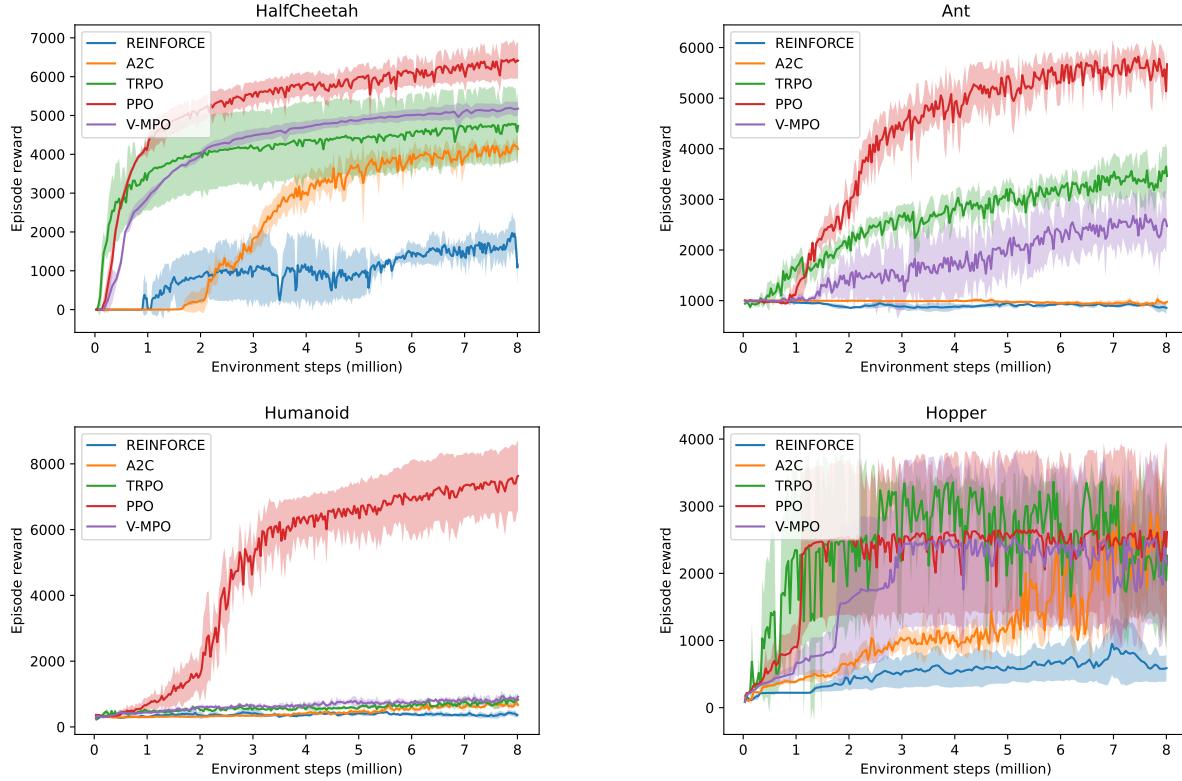


Figure 4: Comparison of rewards per episode during training on several MuJoCo tasks. For each algorithm, we report means and standard deviations of three runs with different random seeds.

6 Numerical Experiments

Now, we empirically compare the discussed policy gradient algorithms. Consistent with the original works [54, 69, 71, 73], we compare them on the established MuJoCo task suite [80], accessed through the Gymnasium library [81]. MuJoCo features robotics simulations, where the tasks are to control and move robots of different shapes by applying torques to each joint.

Our implementations build on the PPO implementation from the BRAX library [23] and are written in JAX [13]. For enhanced comparability, all algorithms that estimate advantages use GAE similarly to PPO. Instead of A3C, we use its synchronous variant A2C due to its simpler implementation. Note that A2C exhibits comparable performance as A3C [85] and only differs in that it waits for all actors to collect transitions to update them synchronously. We modify REINFORCE to average gradients over batches of transitions similarly as in the other algorithms since computing one update per environment step is computationally very costly. Note that this is however likely to improve the performance compared to a naive implementation of REINFORCE. We do not tune hyperparameters and keep choices consistent across algorithms where possible. See Appendix Appendix A for the hyperparameters we use. The experiments were run on a standard consumer CPU. All our implemented algorithms and the code for running the experiments can be found at <https://github.com/Matt00n/PolicyGradientsJax>.

In our main experiment, we compare the performance of the algorithms in terms of the achieved episodic rewards over the course of training. The performances in different MuJoCo tasks are presented in Figure 4. We observe that PPO outperforms the other algorithms in three of four tasks by achieving higher episodic rewards while learning good policies quickly. The performance difference is most prevalent on the *Humanoid*-task, the most challenging of the four, where PPO learns much stronger policies than the other algorithms. In addition, we find our implementation of PPO to be competitive with common RL libraries as shown in Appendix B.1. V-MPO and TRPO are comparable in performance, with each of the two slightly outperforming the other on two out of four environments. We note that V-MPO is intended for training for billions of environment steps, such that its lower performance compared to PPO in our experiments is expected¹⁰ [73]. A2C requires more interactions with the environment to reach similar performance

¹⁰Also see the discussions at <https://openreview.net/forum?id=SylOlp4FvH> on this.

levels as V-MPO and TRPO but fails to learn any useful policy in the *Ant*-task. This slower learning¹¹ is at least partially caused by A2C only using a single update epoch per batch. REINFORCE performance worst on all environments, which is unsurprising giving the high variance of gradients in REINFORCE [75]. This also highlights the benefits of the bias-variance trade-off by the other algorithms as discussed in Section 4.6. We find our performance-based ranking of the algorithms to be consistent with literature (e.g., [71, 73, 5]).

Moreover, we remark that A2C is the only algorithm for which we used an entropy bonus because the learned policies collapsed without it. We showcase this in our expended experiments in Appendix B.2. This underlines the usefulness of the (heuristic) constraints of V-MPO, PPO and TRPO on the KL divergence, which avoid such collapses even without any entropy bonuses. To further investigate this, we show the average KL divergences between consecutive policies throughout training in Figure 5. Here, we approximated the KL divergence using the unbiased estimator [68]

$$\hat{D}_{KL}(\pi_{\text{old}}(\cdot | s) \| \pi_{\text{new}}(\cdot | s)) = \mathbb{E}_{A \sim \pi_{\text{old}}} \left[\frac{\pi_{\text{new}}(A | s)}{\pi_{\text{old}}(A | s)} - 1 - \ln \frac{\pi_{\text{new}}(A | s)}{\pi_{\text{old}}(A | s)} \right]$$

for all algorithms except TRPO, which analytically calculates the exact KL divergence since it is used within the algorithm. We see that the KL divergences remain relatively constant for all algorithms after some initial movement. TRPO displays the most constant KL divergence, which is explained by its hard constraint. With the chosen hyperparameters, V-MPO uses the same bound on the KL divergence as TRPO, however without strictly enforcing it as outlined in the derivation of V-MPO. Thus, V-MPO’s KL divergence exhibits slightly more variance than TRPO and also frequently exceeds this bound. PPO’s clipping heuristic achieves a similar effect resulting in a comparable picture. Due to the lack of constraints on the KL divergence, A2C and REINFORCE show slightly more variance. Interestingly, their KL divergences are orders of magnitudes lower than for the other algorithms, especially for REINFORCE (note the logarithmic scale in Figure 5). We reason this with A2C and REINFORCE using only a singly update epoch per batch, whereas the PPO and V-MPO use multiple epochs and TRPO uses a different update scheme via line search. In Appendix B.3, we provide experimental evidence for this hypothesis. Additionally, we note again that the entropy bonus also stabilizes and limits the KL divergence for A2C as shown in Appendix B.2.

These findings highlight the benefits of regularization through constraining the KL divergence and incentivizing entropy. Regularization stabilizes learning and prevents a collapse of the policy. At the same time, it allows more frequent updates through multiple epochs per batch, which drastically increases the sample efficiency of the algorithms and speeds up learning.

7 Conclusion

In this work, we presented a holistic overview of on-policy policy gradient methods in reinforcement learning. We derived the theoretical foundations of policy gradient algorithms, primarily in the form of the Policy Gradient Theorem. We have shown how the most prominent policy gradient algorithms can be derived based on this theorem. We discussed common techniques used by these algorithms to stabilize training including learning an advantage function to limit the variance of estimated policy gradients, constraining the divergence between policies and regularizing the policy through entropy bonuses. Subsequently, we presented evidence from literature on the convergence behavior of policy gradient algorithms, which suggest that they may find at least locally optimal policies. Finally, we conducted numerical experiments on well-established benchmarks to further compare the behavior of the discussed algorithms. Here, we found that PPO outperforms the other algorithms in the majority of the considered tasks and we provided evidence for the necessity of regularization, by constraining KL divergence or by incentivizing entropy, to stabilize training.

We acknowledge several limitations of our work. First, we deliberately limited our scope to on-policy algorithms, which excludes closely related off-policy policy gradient algorithms and the novelties introduced by them. Second, we presented an incomplete overview of on-policy policy gradient algorithms as other, albeit less established, algorithms exist (e.g., [61, 16]) and the development of further algorithms remains an active research field. Here, we focused on the, in our view, most prominent algorithms as determined by their impact, usage and introduced novelties. Third, the convergence results we referenced rest on assumptions that are quickly violated in practice. In particular, we want to underline that the results based mirror learning rely on the infeasible assumption of finding a global maximizer each iteration. Fourth, while we compared the discussed algorithms empirically and found results to be consistent with existing literature, our analysis is limited to the specific setting we used. Different results may arise on other benchmarks, with different hyperparameters or generally different implementations.

Finally, we note that still many questions remain to be answered in the field of on-policy policy gradient algorithm. So far, our understanding of which algorithm performs best under which circumstances is still limited. Moreover,

¹¹Slow in terms of the required environment steps. Note however that A2C runs significantly faster than PPO, TRPO and V-MPO in absolute time due to using less epochs per batch.

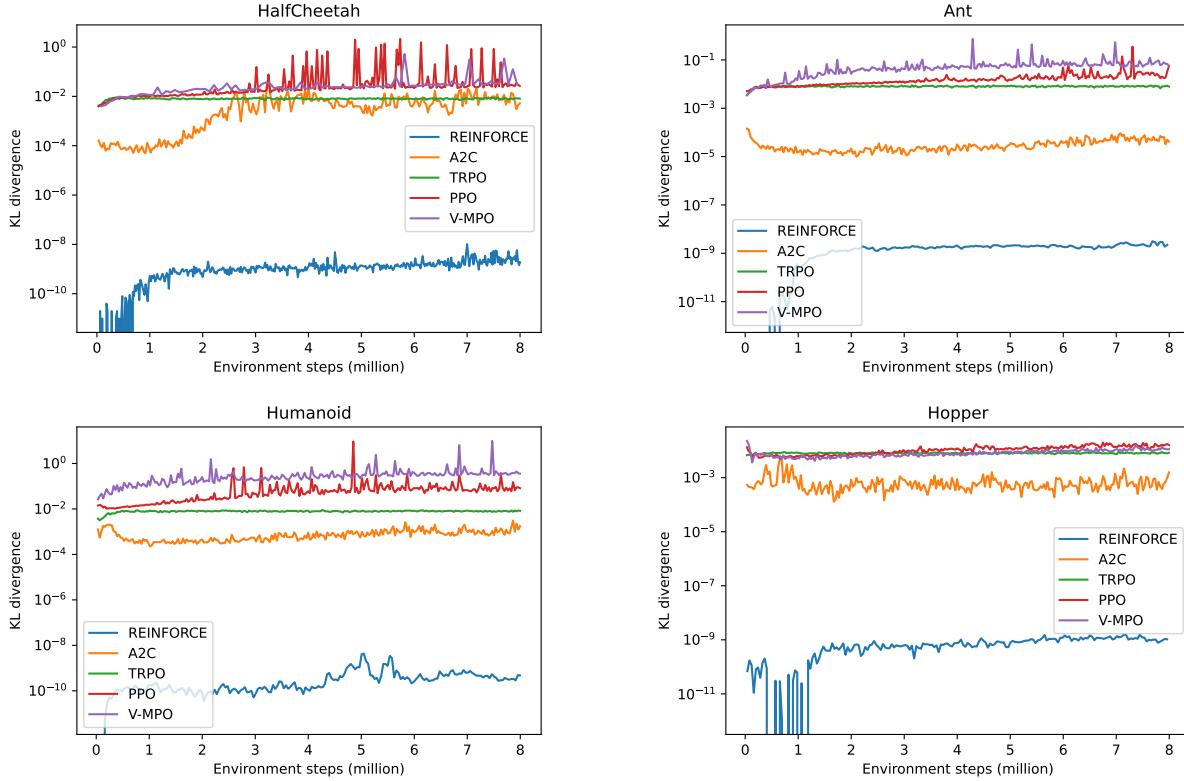


Figure 5: Comparison of the average KL divergence across policies during training.

it is unclear whether the best possible policy gradient algorithm has yet been discovered, which is why algorithm development remains of interest. Similarly, comprehensive empirical comparisons with other classes of RL algorithms may yield further insights on the practical advantages and disadvantages of policy gradient algorithms and how their performance depends on the problem settings. Finally, we observe that still only a limited number of convergence results exist and not even all discussed algorithms are covered by these, e.g., no convergence results exist for V-MPO to the best of our knowledge. Here, further research is needed to enhance our understanding of the convergence behavior of policy gradient algorithms.

References

- [1] Abbas Abdolmaleki, Jost Tobias Springenberg, Jonas Degrave, Steven Bohez, Yuval Tassa, Dan Belov, Nicolas Heess, and Martin Riedmiller. Relative entropy regularized policy iteration. *arXiv preprint arXiv:1812.02256*, 2018.
- [2] Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Remi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation. *arXiv preprint arXiv:1806.06920*, 2018.
- [3] Joshua Achiam. Spinning Up in Deep Reinforcement Learning. 2018.
- [4] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 64–66. PMLR, 09–12 Jul 2020.
- [5] Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphael Marinier, Léonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, et al. What matters in on-policy reinforcement learning? a large-scale empirical study. *arXiv preprint arXiv:2006.05990*, 2020.
- [6] Lawrence M Ausubel and Raymond J Deneckere. A generalized theorem of the maximum. *Economic Theory*, 3(1):99–107, 1993.
- [7] Andrew G Barto, Richard S Sutton, and Charles W Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on systems, man, and cybernetics*, (5):834–846, 1983.
- [8] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.
- [9] Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.
- [10] Julius Berner, Philipp Grohs, Gitta Kutyniok, and Philipp Petersen. The modern mathematics of deep learning. *arXiv preprint arXiv:2105.04026*, pages 86–114, 2021.
- [11] Jalaj Bhandari and Daniel Russo. On the linear convergence of policy gradient methods for finite mdps. In *International Conference on Artificial Intelligence and Statistics*, pages 2386–2394. PMLR, 2021.
- [12] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- [13] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018.
- [14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [15] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial intelligence and statistics*, pages 192–204. PMLR, 2015.
- [16] Karl W Cobbe, Jacob Hilton, Oleg Klimov, and John Schulman. Phasic policy gradient. In *International Conference on Machine Learning*, pages 2020–2027. PMLR, 2021.
- [17] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.
- [18] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in neural information processing systems*, 27, 2014.
- [19] Thomas Degris, Martha White, and Richard S Sutton. Off-policy actor-critic. *arXiv preprint arXiv:1205.4839*, 2012.
- [20] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.
- [21] Prafulla Dhariwal, Christopher Hesse, Oleg Klimov, Alex Nichol, Matthias Plappert, Alec Radford, John Schulman, Szymon Sidor, Yuhuai Wu, and Peter Zhokhov. Openai baselines. <https://github.com/openai/baselines>, 2017.

- [22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [23] C. Daniel Freeman, Erik Frey, Anton Raichuk, Sertan Girgin, Igor Mordatch, and Olivier Bachem. Brax - a differentiable physics engine for large scale rigid body simulation, 2021.
- [24] Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR, 2018.
- [25] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011.
- [26] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.
- [27] Evan Greensmith, Peter L Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5(9), 2004.
- [28] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.
- [29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [30] Magnus R Hestenes, Eduard Stiefel, et al. Methods of conjugate gradients for solving linear systems. *Journal of research of the National Bureau of Standards*, 49(6):409–436, 1952.
- [31] Timothy Classen Hesterberg. *Advances in importance sampling*. Stanford University, 1988.
- [32] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [33] Matthew W. Hoffman, Bobak Shahriari, John Aslanides, Gabriel Barth-Maron, Nikola Momchev, Danila Sinopalnikov, Piotr Stańczyk, Sabela Ramos, Anton Raichuk, Damien Vincent, Léonard Hussenot, Robert Dadashi, Gabriel Dulac-Arnold, Manu Orsini, Alexis Jacq, Johan Ferret, Nino Vieillard, Seyed Kamyar Seyed Ghasemipour, Sertan Girgin, Olivier Pietquin, Feryal Behbahani, Tamara Norman, Abbas Abdolmaleki, Albin Cassirer, Fan Yang, Kate Baumli, Sarah Henderson, Abe Friesen, Ruba Haroun, Alex Novikov, Sergio Gómez Colmenarejo, Serkan Cabi, Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Andrew Cowie, Ziyu Wang, Bilal Piot, and Nando de Freitas. Acme: A research framework for distributed reinforcement learning. *arXiv preprint arXiv:2006.00979*, 2020.
- [34] Markus Holzleitner, Lukas Gruber, José Arjona-Medina, Johannes Brandstetter, and Sepp Hochreiter. Convergence proof for actor-critic methods applied to ppo and rudder. In *Transactions on Large-Scale Data-and Knowledge-Centered Systems XLVIII: Special Issue In Memory of Univ. Prof. Dr. Roland Wagner*, pages 105–130. Springer, 2021.
- [35] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.
- [36] Shengyi Huang, Rousslan Fernand Julien Dossa, Chang Ye, Jeff Braga, Dipam Chakraborty, Kinal Mehta, and João G.M. Araújo. Cleanrl: High-quality single-file implementations of deep reinforcement learning algorithms. *Journal of Machine Learning Research*, 23(274):1–18, 2022.
- [37] David R Hunter and Kenneth Lange. A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37, 2004.
- [38] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 267–274, 2002.
- [39] Prasenjit Karmakar and Shalabh Bhatnagar. Two time-scale stochastic approximation with controlled markov noise and off-policy temporal-difference learning. *Mathematics of Operations Research*, 43(1):130–151, 2018.
- [40] Henry J Kelley. Gradient theory of optimal flight paths. *Ars Journal*, 30(10):947–954, 1960.
- [41] Jack Kiefer and Jacob Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, pages 462–466, 1952.
- [42] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

- [43] Vijay R Konda and John N Tsitsiklis. On actor-critic algorithms. *SIAM journal on Control and Optimization*, 42(4):1143–1166, 2003.
- [44] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [45] Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. Trust region policy optimisation in multi-agent reinforcement learning. *arXiv preprint arXiv:2109.11251*, 2021.
- [46] Jakub Grudzien Kuba, Christian Schroeder de Witt, and Jakob Foerster. Mirror learning: A unifying framework of policy optimisation. *arXiv preprint arXiv:2201.02373*, 2022.
- [47] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
- [48] Johannes Lederer. Activation functions in artificial neural networks: A systematic overview. *arXiv preprint arXiv:2101.09957*, 2021.
- [49] Eric Liang, Richard Liaw, Robert Nishihara, Philipp Moritz, Roy Fox, Ken Goldberg, Joseph E. Gonzalez, Michael I. Jordan, and Ion Stoica. RLlib: Abstractions for distributed reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2018.
- [50] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.
- [51] Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv preprint arXiv:1906.10306*, 2019.
- [52] Peter Marbach and John N Tsitsiklis. Simulation-based optimization of markov reward processes. *IEEE Transactions on Automatic Control*, 46(2):191–209, 2001.
- [53] Charles C Margossian. A review of automatic differentiation and its efficient implementation. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 9(4):e1305, 2019.
- [54] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR, 2016.
- [55] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.
- [56] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2 edition, 2018.
- [57] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.
- [58] David Pollard. Asymptopia: an exposition of statistical asymptotic theory. In *Asymptopia: an exposition of statistical asymptotic theory*, 2000.
- [59] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.
- [60] 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.
- [61] Md Masudur Rahman and Yexiang Xue. Robust policy optimization in deep reinforcement learning. *arXiv preprint arXiv:2212.07536*, 2022.
- [62] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.
- [63] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.
- [64] Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pages 233–257. Elsevier, 1971.

- [65] Reuven Y. Rubinstein. *Simulation and the Monte Carlo Method*. Wiley, New York, first edition, 1981.
- [66] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning internal representations by error propagation, 1985.
- [67] Gavin A Rummery and Mahesan Niranjan. *On-line Q-learning using connectionist systems*, volume 37. University of Cambridge, Department of Engineering Cambridge, UK, 1994.
- [68] John Schulman. Approximating kl divergence. *John Schulman's Homepage*, 2020.
- [69] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.
- [70] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.
- [71] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [72] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.
- [73] H Francis Song, Abbas Abdolmaleki, Jost Tobias Springenberg, Aidan Clark, Hubert Soyer, Jack W Rae, Seb Noury, Arun Ahuja, Siqi Liu, Dhruva Tirumala, et al. V-mpo: On-policy maximum a posteriori policy optimization for discrete and continuous control. *arXiv preprint arXiv:1909.12238*, 2019.
- [74] Richard S Sutton and Andrew G Barto. Toward a modern theory of adaptive networks: expectation and prediction. *Psychological review*, 88(2):135, 1981.
- [75] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.
- [76] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.
- [77] Richard S Sutton, Satinder Singh, and David McAllester. Comparing policy-gradient algorithms. *IEEE Transactions on Systems, Man, and Cybernetics*, 2000.
- [78] Richard Stuart Sutton. *Temporal credit assignment in reinforcement learning*. University of Massachusetts Amherst, 1984.
- [79] Gerald Tesauro et al. Temporal difference learning and td-gammon. *Communications of the ACM*, 38(3):58–68, 1995.
- [80] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.
- [81] Mark Towers, Jordan K. Terry, Ariel Kwiatkowski, John U. Balis, Gianluca de Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Arjun KG, Markus Krimmel, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Andrew Tan Jin Shen, and Omar G. Younis. Gymnasium, March 2023.
- [82] Hado van Hasselt. Reinforcement learning lecture 5: Model-free prediction, October 2021.
- [83] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [84] Christopher John Cornish Hellaby Watkins. Learning from delayed rewards. 1989.
- [85] Lilian Weng. Policy gradient algorithms. *lilianweng.github.io*, 2018.
- [86] Robert Edwin Wengert. A simple automatic derivative evaluation program. *Communications of the ACM*, 7(8):463–464, 1964.
- [87] Ronald J Williams. *Reinforcement-learning connectionist systems*. College of Computer Science, Northeastern University, 1987.

- [88] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Reinforcement learning*, pages 5–32, 1992.
- [89] Ronald J Williams and Jing Peng. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268, 1991.
- [90] Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Basar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6):3586–3612, 2020.

Hyperparameter	Value				
	REINFORCE	A2C	TRPO	PPO	V-MPO
Learning rate	$3 \cdot 10^{-4}$				
Num. minibatches	1	8	8	8	8
Num. epochs	1	1	1^{12}	10	10
Discount (γ)	—	0.99	0.99	0.99	0.99
GAE parameter (λ)	—	0.95	0.95	0.95	0.95
Normalize advantages	—	True	True	True	False
Entropy bonus coef.	0	0.1	0	0	0
Max. grad. norm	0.5	0.5	0.5	0.5	0.5
Unroll length	—	2048	2048	2048	2048
KL target (δ)	—	—	0.01	—	—
CG damping	—	—	0.1	—	—
CG max. iterations	—	—	10	—	—
Line search max. iterations	—	—	10	—	—
Line search shrinkage factor	—	—	0.8	—	—
PPO clipping (ε)	—	—	—	0.2	—
Min. temp. (η_{\min})	—	—	—	—	10^{-8}
Min. KL pen. (ν_{\min})	—	—	—	—	10^{-8}
Init. temp. (η_{init})	—	—	—	—	1
Init. KL pen. (mean) ($\nu_{\mu_{\text{init}}}$)	—	—	—	—	1
Init. KL pen. (std) ($\nu_{\sigma_{\text{init}}}$)	—	—	—	—	1
KL target (mean) ($\varepsilon_{\nu_{\mu}}$)	—	—	—	—	0.01
KL target (std) ($\varepsilon_{\nu_{\sigma}}$)	—	—	—	—	$5 \cdot 10^{-5}$
KL target (temp.) (ε_{η})	—	—	—	—	0.01

Table 2: Algorithm hyperparameters.

Appendices

A Hyperparameters

We report the hyperparameters we use in our main experiments in Table 2. All algorithms use separate policy and value networks. Policy networks use 4 hidden layers with 32 neurons respectively. Value networks use 5 layers with 256 neurons each. We use swish-activation functions [62] throughout both networks. Policy outputs are transformed to fit the bounds of the actions spaces via a squashing function. We use the Adam optimizer [42] with gradient clipping and a slight linear decay of the learning rates. Further, we preprocess observations and rewards by normalizing them using running means and standard deviations and clipping them to the interval $[-10, 10]$. All algorithms except REINFORCE use 8 parallel environments to collect experience. We use independent environments to evaluate the agents throughout training. In the evaluations, agents select actions deterministically as the mode of the constructed distribution.

B Extended Experiments

Here, we present results from further experiments. Unless indicated otherwise, we use the hyperparameters as reported in Appendix A.

B.1 Comparison to RL frameworks

In Table 3, we compare the performance of our implementation of PPO with popular RL frameworks. Note that we did not tune any hyperparameters for our implementations such that the reported scores should be understood as lower bounds. We compare PPO since it is the most popular and commonly implemented of the discussed algorithms across frameworks. In contrast, especially TRPO and V-MPO are rarely found.

¹²TRPO uses one epoch for its policy updates but 10 epochs per batch for updating the value network.

¹³Numbers read approximately from plots in the paper.

	Framework						
	CleanRL [36]	Baselines [21]	SB3 [60]	RLLib [49]	ACME ¹³ [33]	Ours	Ours
MuJoCo version	v4	v1	v3	v2	v2	v4	v4
Steps in million	1	1	1	44	10	1	8
HalfCheetah	2906	1669	5819	9664	6800	4332	6414
Hopper	2052	2316	2410	—	2550	895	2616
Humanoid	742	—	—	—	6600	700	7633
Ant	—	—	1327	—	5200	1258	5671

Table 3: Comparison of the mean performance of our PPO implementation with popular RL frameworks. Scores for the frameworks are shown as reported in the respective paper or documentation.

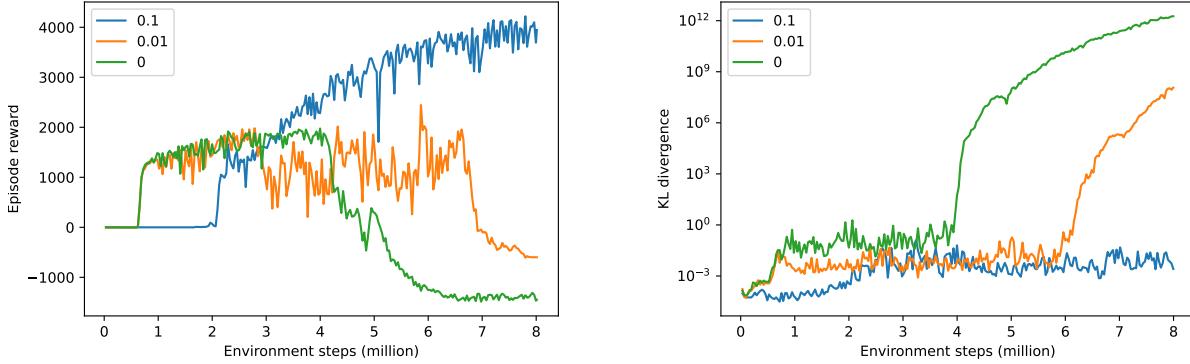


Figure 6: We compare the episode reward (left) and KL divergence (right) for different values of the entropy coefficient for A2C on HalfCheetah.

B.2 Entropy Bonus in A2C

In Figure 6, we show that using an entropy bonus improves the performance of A2C by stabilizing learning. In particular, insufficiently low values of the entropy coefficient result in a collapse of the policy after some time. This is visible in a drastic increase in the KL divergences (note the logarithmic scale).

B.3 A2C and REINFORCE with Multiple Update Epochs

In Figure 7, we showcase that the KL divergence is low for A2C and REINFORCE due to using only a single update epoch per batch. On the contrary, when using multiple epochs, the policies collapse for both algorithms as visible by the diverging KL divergence and abrupt performance loss. Note, that here we show this behavior for five epochs, however in our tests A2C and REINFORCE display similar behaviors already when only using two epochs, albeit the policies then only collapse after an extended period of time. Further, note that over the displayed range of environment steps, the algorithms do not yet learn any useful policies when using a single epoch. However, performance improves for both A2C and REINFORCE when given more time as depicted in Figure 4.

C V-MPO: Derivation Details

In the following, we provide a more detailed derivation of the objective function of V-MPO

$$J_{\text{V-MPO}}(\theta, \eta, \nu) = \mathcal{L}_\pi(\theta) + \mathcal{L}_\eta(\eta) + \mathcal{L}_\nu(\theta, \nu),$$

where \mathcal{L}_π is the policy loss

$$\mathcal{L}_\pi(\theta) = - \sum_{a,s \in \mathcal{D}} \frac{\exp\left(\frac{\hat{A}_\phi(s,a)}{\eta}\right)}{\sum_{a',s' \in \mathcal{D}} \exp\left(\frac{\hat{A}_\phi(s',a')}{\eta}\right)} \ln \pi_\theta(a | s), \quad (36)$$

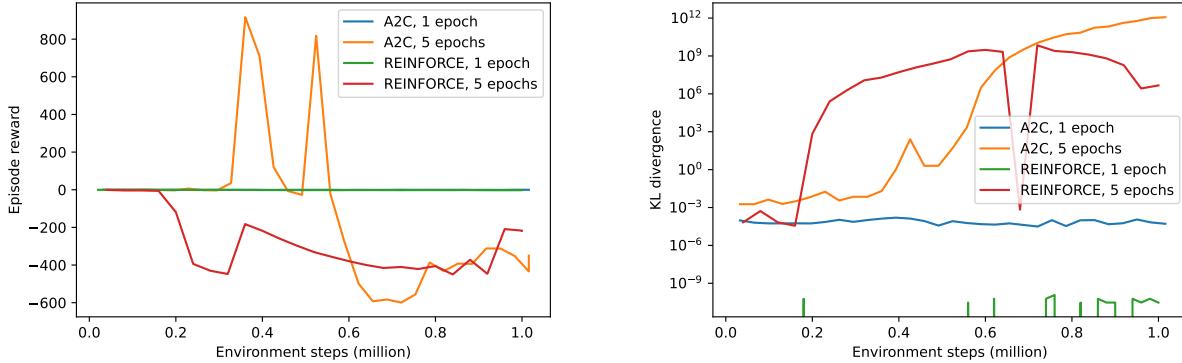


Figure 7: We compare the episode reward (left) and KL divergence (right) for different numbers of update epochs for A2C and REINFORCE on HalfCheetah.

\mathcal{L}_η is the temperature loss

$$\mathcal{L}_\eta(\eta) = \eta \varepsilon_\eta + \eta \ln \left[\frac{1}{|\tilde{\mathcal{D}}|} \sum_{a,s \in \tilde{\mathcal{D}}} \exp \left(\frac{\hat{A}_\phi(s, a)}{\eta} \right) \right] \quad (37)$$

and \mathcal{L}_ν is the trust-region loss

$$\mathcal{L}_\nu(\theta, \nu) = \frac{1}{|\mathcal{D}|} \sum_{s \in \mathcal{D}} \left(\nu \left(\varepsilon_\nu - \text{sg} \left[\left[D_{\text{KL}}(\pi_{\text{old}}(\cdot | s) \| \pi_\theta(\cdot | s)) \right] \right] \right) + \text{sg}[\nu] D_{\text{KL}}(\pi_{\text{old}}(\cdot | s) \| \pi_\theta(\cdot | s)) \right). \quad (38)$$

Let $p_\theta(s, a) = \pi_\theta(a | s) d\pi_\theta(s)$ denote the joint state-action distribution under policy π_θ conditional on the parameters θ . Let \mathcal{I} be a binary random variable whether the updated policy π_θ is an improvement over the old policy π_{old} , i.e. $\mathcal{I} = 1$ if it is an improvement. We assume the probability of π_θ being an improvement is proportional to the following expression

$$p_\theta(\mathcal{I} = 1 | s, a) \propto \exp \left(\frac{A_{\pi_{\text{old}}}(s, a)}{\eta} \right) \quad (39)$$

Given the desired outcome $\mathcal{I} = 1$, we seek the posterior distribution conditioned on this event. Specifically, we seek the maximum a posteriori estimate

$$\begin{aligned} \theta^* &= \arg \max_{\theta} [p_\theta(\mathcal{I} = 1) \rho(\theta)] \\ &= \arg \max_{\theta} [\ln p_\theta(\mathcal{I} = 1) + \ln \rho(\theta)], \end{aligned} \quad (40)$$

where ρ is some prior distribution to be specified. Using Theorem D.7, we obtain

$$\ln p_\theta(\mathcal{I} = 1) = \mathbb{E}_{S, A \sim \psi} \left[\ln \frac{p_\theta(\mathcal{I} = 1, S, A)}{\psi(S, A)} \right] + D_{\text{KL}}(\psi \| p_\theta(\cdot, \cdot | \mathcal{I} = 1)), \quad (41)$$

where ψ is a distribution over $\mathcal{S} \times \mathcal{A}$. Observe that, since the KL-divergence is non-negative, the first term is a lower bound for $\ln p_\theta(\mathcal{I} = 1)$. Akin to EM algorithms, V-MPO now iterates between an expectation (E) and a maximization (M) step. In the E-step we choose the variational distribution ψ to minimize the KL divergence in Equation (41) to make the lower bound as tight as possible. In the M-step, we maximize this lower bound and the prior $\ln \rho(\theta)$ to obtain a new estimate of θ^* via Equation (40).

First, we consider the E-step. Minimizing $D_{\text{KL}}(\psi \| p_{\theta_{\text{old}}}(\cdot, \cdot | \mathcal{I} = 1))$ w.r.t. ψ leads to

$$\begin{aligned} \psi(s, a) &= p_{\theta_{\text{old}}}(s, a | \mathcal{I} = 1) \\ &= \frac{p_{\theta_{\text{old}}}(s, a) p_{\theta_{\text{old}}}(\mathcal{I} = 1 | s, a)}{p_{\theta_{\text{old}}}(\mathcal{I} = 1)} \\ &= \frac{p_{\theta_{\text{old}}}(s, a) p_{\theta_{\text{old}}}(\mathcal{I} = 1 | s, a)}{\int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} p_{\theta_{\text{old}}}(s, a) p_{\theta_{\text{old}}}(\mathcal{I} = 1 | s, a) da ds} \end{aligned}$$

using Bayes' Theorem (Theorem D.2). Sampling from right-hand side of (39) thus yields

$$\hat{\psi}(s, a) = \frac{\exp\left(\frac{A_{\pi_{\text{old}}}(s, a)}{\eta}\right)}{\sum_{a, s \in \mathcal{D}} \exp\left(\frac{A_{\pi_{\text{old}}}(s, a)}{\eta}\right)},$$

which is the variational distribution found in the policy loss (37). [73] find that using only the highest 50 % of advantages per batch, i.e. replacing \mathcal{D} with $\tilde{\mathcal{D}}$, substantially improves the algorithm. The advantage function A_π is estimated by \hat{A}_ϕ , which is learned identically as in A3C.

We derive the temperature loss to automatically adjust the temperature η by applying (39) to the KL term in (41), which we want to minimize:

$$\begin{aligned} D_{KL}\left(\psi \| p(\cdot, \cdot | \mathcal{I} = 1)\right) &= D_{KL}\left(\psi \| \frac{p_{\theta_{\text{old}}}(S, A)p_{\theta_{\text{old}}}(\mathcal{I} = 1 | S, A)}{p_{\theta_{\text{old}}}(\mathcal{I} = 1)}\right) \\ &= D_{KL}\left(\psi \| \frac{p_{\theta_{\text{old}}}(S, A) \exp\left(\frac{A_{\pi_{\text{old}}}(S, A)}{\eta}\right)}{p_{\theta_{\text{old}}}(\mathcal{I} = 1)}\right) \\ &= - \int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \psi(s, a) \ln \left(\frac{p_{\theta_{\text{old}}}(s, a) \exp\left(\frac{A_{\pi_{\text{old}}}(s, a)}{\eta}\right)}{\psi(s, a) p_{\theta_{\text{old}}}(\mathcal{I} = 1)} \right) da ds \end{aligned}$$

By applying the logarithm to the individual terms, rearranging and multiplying through by η we get

$$\begin{aligned} D_{KL}\left(\psi \| p(\cdot, \cdot | \mathcal{I} = 1)\right) &= - \int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \psi(s, a) \left(\frac{A_{\pi_{\text{old}}}(s, a)}{\eta} + \ln p_{\theta_{\text{old}}}(s, a) \right. \\ &\quad \left. - \ln p_{\theta_{\text{old}}}(\mathcal{I} = 1) - \ln \psi(s, a) \right) da ds \\ &\propto - \int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \psi(s, a) \left(A_{\pi_{\text{old}}}(s, a) + \eta \ln p_{\theta_{\text{old}}}(s, a) - \eta \ln p_{\theta_{\text{old}}}(\mathcal{I} = 1) \right. \\ &\quad \left. - \eta \ln \psi(s, a) \right) da ds \\ &= - \int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \psi(s, a) A_{\pi_{\text{old}}}(s, a) da ds + \eta \int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \psi(s, a) \ln \frac{\psi(s, a)}{p_{\theta_{\text{old}}}(s, a)} da ds \\ &\quad + \lambda \int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \psi(s, a) da ds \end{aligned}$$

with $\lambda = \eta \ln p_{\theta_{\text{old}}}(\mathcal{I} = 1)$. To optimize η while minimizing the KL term, we transform this into a constrained optimization problem with a bound on the KL divergence

$$\begin{aligned} \arg \max_{\psi} & \int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \psi(s, a) A_{\pi_{\text{old}}}(s, a) da ds \\ \text{subject to } & \int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \psi(s, a) \ln \frac{\psi(s, a)}{p_{\theta_{\text{old}}}(s, a)} da ds \leq \varepsilon_\eta, \\ & \int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \psi(s, a) da ds = 1 \end{aligned}$$

and then back into an unconstrained problem via Lagrangian relaxation, yielding the objective function

$$\begin{aligned} \mathcal{J}(\psi, \eta, \lambda) &= \int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \psi(s, a) A_{\pi_{\text{old}}}(s, a) da ds + \eta \left(\varepsilon_\eta \right. \\ &\quad \left. - \int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \psi(s, a) \ln \frac{\psi(s, a)}{p_{\theta_{\text{old}}}(s, a)} da ds \right) + \lambda \left(1 - \int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \psi(s, a) da ds \right). \end{aligned}$$

Differentiating w.r.t. $\psi(s, a)$ and setting to zero yields

$$\psi(s, a) = p_{\theta_{\text{old}}}(s, a) \exp\left(\frac{A_{\pi_{\text{old}}}(s, a)}{\eta}\right) \exp\left(-1 - \frac{\lambda}{\eta}\right)$$

Normalizing over s and a confirms the already attained solution

$$\psi(s, a) = \frac{p_{\theta_{\text{old}}}(s, a) \exp\left(\frac{A_{\pi_{\text{old}}}(s, a)}{\eta}\right)}{\int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} p_{\theta_{\text{old}}}(s, a) \exp\left(\frac{A_{\pi_{\text{old}}}(s, a)}{\eta}\right) da ds}, \quad (42)$$

but now we can also find the optimal η by substituting this solution into $\mathcal{J}(\psi, \eta, \lambda)$. Doing so and dropping terms independent of η leads to

$$\begin{aligned} & \eta \left(\varepsilon_\eta - \int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \psi(s, a) \ln \frac{\psi(s, a)}{p_{\theta_{\text{old}}}(s, a)} da ds \right) \\ &= \eta \varepsilon_\eta + \eta \int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \psi(s, a) \ln p_{\theta_{\text{old}}}(s, a) da ds - \eta \int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \psi(s, a) \ln \psi(s, a) da ds. \end{aligned} \quad (43)$$

Because of Equation (42), we have

$$\begin{aligned} \eta \psi(s, a) \ln \psi(s, a) &= \eta \psi(s, a) \ln \frac{p_{\theta_{\text{old}}}(s, a) \exp\left(\frac{A_{\pi_{\text{old}}}(s, a)}{\eta}\right)}{\int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} p_{\theta_{\text{old}}}(s, a) \exp\left(\frac{A_{\pi_{\text{old}}}(s, a)}{\eta}\right) da ds} \\ &= \psi(s, a) \left(\eta \ln p_{\theta_{\text{old}}}(s, a) + A_{\pi_{\text{old}}}(s, a) - \eta \ln \int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} p_{\theta_{\text{old}}}(s, a) \exp\left(\frac{A_{\pi_{\text{old}}}(s, a)}{\eta}\right) da ds \right), \end{aligned}$$

where the first summand cancels out the second term in (43) and the second summand no longer depends on η and thus can be dropped. Hence, we obtain the temperature loss function

$$\mathcal{L}_\eta(\eta) = \eta \varepsilon_\eta + \eta \ln \left(\int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \exp\left(\frac{A_{\pi_{\text{old}}}(s, a)}{\eta}\right) da ds \right) \quad (44)$$

through which we can optimize η using gradient descent.

Given the non-parametric sample-based variational distribution $\psi(s, a)$, the M-step now optimizes the policy parameters θ . Based on (40), we want to maximize the discussed lower bound, i.e. minimize

$$-\int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \psi(s, a) \ln \frac{p_\theta(\mathcal{I} = 1, s, a)}{\psi(s, a)} da ds - \ln p(\theta)$$

to find new policy parameters θ . Using Equations (42) and (39), the first term becomes

$$\begin{aligned} & -\int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \psi(s, a) \ln \frac{p_\theta(\mathcal{I} = 1, s, a)}{\psi(s, a)} da ds \\ &= -\int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \psi(s, a) \ln \frac{p_\theta(\mathcal{I} = 1 | s, a) p_\theta(s, a)}{\psi(s, a)} da ds \\ &= -\int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \psi(s, a) \ln \left(\frac{\exp\left(\frac{A_{\pi_{\text{old}}}(s, a)}{\eta}\right) p_\theta(s, a)}{p_{\theta_{\text{old}}}(s, a) \exp\left(\frac{A_{\pi_{\text{old}}}(s, a)}{\eta}\right)} \frac{1}{\int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} p_{\theta_{\text{old}}}(s, a) \exp\left(\frac{A_{\pi_{\text{old}}}(s, a)}{\eta}\right) da ds} \right) da ds \\ &= -\int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \psi(s, a) \ln \left(\frac{p_\theta(s, a)}{p_{\theta_{\text{old}}}(s, a)} \frac{1}{\int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} p_{\theta_{\text{old}}}(s, a) \exp\left(\frac{A_{\pi_{\text{old}}}(s, a)}{\eta}\right) da ds} \right) da ds. \end{aligned}$$

Using $p_\theta(s, a) = \pi_\theta(a | s)d^{\pi_\theta}(s)$, assuming the state distribution d^π to be independent of θ and dropping terms that do not depend on θ yields

$$\begin{aligned} \arg \min_{\theta} \left(- \int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \psi(s, a) \ln \frac{p_\theta(\mathcal{I} = 1, s, a)}{\psi(s, a)} da ds \right) &= \arg \min_{\theta} \left(- \int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \psi(s, a) \ln p_\theta(s, a) da ds \right) \\ &= \arg \min_{\theta} \left(- \int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \psi(s, a) \ln \pi_\theta(a | s) da ds \right), \end{aligned}$$

which is the weighted maximum likelihood policy loss as in (36), that we compute on sampled transitions, effectively assigning out-of-sample transitions a weight of zero.

A useful prior $\rho(\theta)$ in Equation (40) is to keep the new policy close to the previous one as in TRPO and PPO. This translates to

$$\rho(\theta) \approx -\nu \mathbb{E}_{S \sim d^{\pi_{\text{old}}}} [D_{KL}(\pi_{\text{old}}(\cdot | S) \| \pi_\theta(\cdot | S))].$$

Since optimizing the resulting sample-based maximum likelihood objective directly tends to result in overfitting, this prior is instead transformed into a constraint on the KL-divergence with bound ε_ν , i.e.

$$\begin{aligned} \arg \min_{\theta} \quad & \left(- \int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \psi(s, a) \ln \frac{p_\theta(\mathcal{I} = 1, s, a)}{\psi(s, a)} da ds \right) \\ \text{subject to} \quad & \mathbb{E}_{S \sim d^{\pi_{\text{old}}}} [D_{KL}(\pi_{\text{old}}(\cdot | S) \| \pi_\theta(\cdot | S))] \leq \varepsilon_\nu. \end{aligned}$$

To employ gradient-based optimization, we use Lagrangian relaxation to transform this constraint optimization problem back into the unconstrained problem

$$\mathcal{J}(\theta, \nu) = \mathcal{L}_\pi(\theta) + \nu (\varepsilon_\nu - \mathbb{E}_{S \sim d^{\pi_{\text{old}}}} [D_{KL}(\pi_{\text{old}}(\cdot | S) \| \pi_\theta(\cdot | S))]). \quad (45)$$

This problem is solved by alternating between optimizing for θ and ν via gradient descent in a coordinate-descent strategy. Using the stop-gradient operator $\text{sg}[[\cdot]]$, the objective can equivalently to this strategy be rewritten for as

$$\mathcal{L}_\nu(\theta, \nu) = \nu \left(\varepsilon_\nu - \mathbb{E}_{S \sim d^{\pi_{\text{old}}}} \left[\text{sg}[[D_{KL}(\pi_{\theta_{\text{old}}}(\cdot | S) \| \pi_\theta(\cdot | S))]] \right] \right) + \text{sg}[[\nu]] \mathbb{E}_{S \sim d^{\pi_{\text{old}}}} [D_{KL}(\pi_{\theta_{\text{old}}}(\cdot | S) \| \pi_\theta(\cdot | S))].$$

Sampling this gives Equation (38). η and ν are Lagrangian multipliers and hence must be positive. We enforce this by projecting the computed values to small positive values η_{\min} and ν_{\min} respectively if necessary.

D Auxiliary Theory

Here, we list a range of well-known definitions and results that we use in our work.

Definition D.1. (*Compact Space*) A topological space X is called compact if for every set S of open covers of X , there exists a finite subset $S' \subset S$ that also is an open cover of X .

Theorem D.2. (*Bayes' Theorem*) Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $\bigcup_{i \in I} B_i$ be a disjoint and finite partition of Ω with $B_i \in \mathcal{A}$ and $\mathbb{P}(B_i) > 0$ for $i \in I$. Then, for all $A \in \mathcal{A}$ and all $k \in I$

$$\mathbb{P}(B_k | A) = \frac{\mathbb{P}(A | B_k)\mathbb{P}(B_k)}{\sum_{i \in I} \mathbb{P}(A | B_i)\mathbb{P}(B_i)}.$$

Theorem D.3. Let X be a random variable. Then,

$$\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Definition D.4. (*Entropy*) Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $X \sim \mathbb{P}$ be a random variable. The entropy of X is given by

$$H(X) := \mathbb{E}_{X \sim \mathbb{P}} [-\ln \mathbb{P}(X)].$$

Definition D.5. (*Kullback-Leibler Divergence*) For any measurable space \mathcal{A} and probability densities p and q of the respective distributions P and Q , the Kullback-Leibler divergence or relative entropy from Q to P is given by

$$D_{KL}(p\|q) := \int_{a \in \mathcal{A}} p(a) \ln \frac{p(a)}{q(a)} da.$$

Definition D.6. (*Total Variation Divergence*) For any measurable space \mathcal{A} and probability densities p and q of the respective distributions P and Q , the total variation variance from Q to P is given by

$$D_{TV}(p\|q) := \frac{1}{2} \int_{a \in \mathcal{A}} |p(a) - q(a)| da.$$

Theorem D.7. Let (Ω, \mathcal{A}) be a measurable space and p and ψ be probability measures on that space. Let $X \in \mathcal{A}$ and $Z \in \mathcal{A}$. Then,

$$\ln p(X) = \mathbb{E}_{Z \sim \psi} \left[\ln \frac{p(X, Z)}{\psi(Z)} \right] + D_{KL}(\psi \| p(\cdot | X)).$$

Theorem D.8. Let X be a random variable. Then,

$$\min_a \mathbb{E}[(X - a)^2] = \mathbb{E}[X].$$

Theorem D.9. Let (\mathcal{A}, Σ) be a measurable space with σ -finite measures μ and ν such that ν is absolutely continuous in μ . Let g be a Radon-Nikodym derivative of ν w.r.t. μ , i.e. $\nu(A) = \int_A g d\mu$ for all $A \in \Sigma$. Let, f be a ν -integrable function. Then,

$$\int_{\mathcal{A}} f d\nu = \int_{\mathcal{A}} (f \cdot g) d\mu.$$

Theorem D.10. (*Leibniz Integral Rule*) Let X be an open subset of \mathbb{R}^d , $d \in \mathbb{N}$. Let \mathcal{A} be a measurable set and $f: X \times \mathcal{A} \rightarrow \mathbb{R}$ be a function which satisfies

1. $f(x, a)$ is a Lebesgue-integrable function of a for all $x \in X$.
2. For almost all $a \in \mathcal{A}$, all partial derivatives exist for all $x \in X$.
3. There exists some integrable function $g: \mathcal{A} \rightarrow \mathbb{R}$ with $|\nabla_x f(x, a)| \leq g(a)$ for all $x \in X$ and almost all $a \in \mathcal{A}$.

Then, for all $x \in X$ we have

$$\nabla_x \int_{a \in \mathcal{A}} f(x, a) da = \int_{a \in \mathcal{A}} \nabla_x f(x, a) da$$

Theorem D.11. (*Fubini's Theorem*) Let \mathcal{A}_1 and \mathcal{A}_2 be measurable spaces with measures μ_1 and μ_2 and $f: \mathcal{A}_1 \times \mathcal{A}_2 \rightarrow \mathbb{R}$ be measurable and integrable w.r.t. the product measure $\mu_1 \otimes \mu_2$, i.e. $\int_{\mathcal{A}_1 \times \mathcal{A}_2} |f| d(\mu_1 \otimes \mu_2) < \infty$ or $f \geq 0$ almost everywhere. Then, $f(x, y)$ is integrable for almost all x and y and

$$\int_{\mathcal{A}_1} \int_{\mathcal{A}_2} f(x, y) d\mu_1(x) d\mu_2(y) = \int_{\mathcal{A}_2} \int_{\mathcal{A}_1} f(x, y) d\mu_2(y) d\mu_1(x)$$

Theorem D.12. (*Taylor's Theorem - one-dimensional*) Let $k \in \mathbb{N}$ and let $f: \mathbb{R} \rightarrow \mathbb{R}$ be k -times differentiable at $a \in \mathbb{R}$. Then, there exists a function $h_k: \mathbb{R} \rightarrow \mathbb{R}$ such that

$$f(x) = \sum_{i=0}^k \frac{f^{(i)}(a)}{i!} (x-a)^i + h_k(x)(x-a)^k.$$

Theorem D.13. (*Monotone Convergence Theorem*) Let $(x_n)_{n=0}^\infty \subset \mathbb{R}$ be a bounded and monotonically increasing sequence. Then, the sequence converges, i.e. $\lim_{n \rightarrow \infty} x_n$ exists and is finite.

Theorem D.14. (*Bolzano-Weierstrass Theorem*) Let $(x_n)_{n=0}^\infty \subset \mathbb{R}^d$, $d \in \mathbb{N}$ be a bounded sequence. Then, there exists some convergent subsequence $(x_{n_i})_{i=0}^\infty$.

Theorem D.15. (*Berge's Maximum Theorem*) Let X and Θ be topological spaces, $f: X \times \Theta \rightarrow \mathbb{R}$ be continuous on $X \times \Theta$ and $C: \Theta \rightrightarrows X$ be a compact-valued correspondence with $C(\theta) \neq \emptyset$ for all $\theta \in \Theta$. Let

$$f^*(\theta) = \sup \{ f(x, \theta) \mid x \in C(\theta) \}$$

and

$$C^*(\theta) = \arg \max \{ f(x, \theta) \mid x \in C(\theta) \} = \{ x \in C(\theta) \mid f(x, \theta) = f^*(\theta) \}.$$

If C is continuous at θ , then f^* is continuous and C^* is upper hemicontinuous with nonempty and compact values.

Definition D.16. (*Gâteaux Derivative*) Let X and Y be locally convex topological spaces, let U be an open subset of X and $F: U \rightarrow Y$. The Gâteaux derivative of F at $x \in U$ in the direction $d \in X$ is defined as

$$dF(x, d) = \lim_{h \rightarrow 0} \frac{F(x + rd) - F(x)}{r}.$$