
Mirror Learning: A Unifying Framework of Policy Optimisation

Jakub Grudzien Kuba¹ Christian Schroeder de Witt¹ Jakob Foerster¹

Abstract

Modern deep reinforcement learning (RL) algorithms are *motivated* by either the generalised policy iteration (GPI) or trust-region learning (TRL) frameworks. However, algorithms that *strictly respect* these theoretical frameworks have proven unscalable. Surprisingly, the only known scalable algorithms violate the GPI/TRL assumptions, e.g. due to required regularisation or other heuristics. The current explanation of their empirical success is essentially “by analogy”: they are deemed approximate adaptations of theoretically sound methods. Unfortunately, studies have shown that in practice these algorithms differ greatly from their conceptual ancestors. In contrast, in this paper we introduce a novel theoretical framework, named *Mirror Learning*, which provides theoretical guarantees to a large class of algorithms, including TRPO and PPO. While the latter two exploit the flexibility of our framework, GPI and TRL fit in merely as pathologically restrictive corner cases thereof. This suggests that the empirical performance of state-of-the-art methods is a direct consequence of their theoretical properties, rather than of aforementioned approximate analogies. Mirror learning sets us free to boldly explore novel, theoretically sound RL algorithms, a thus far uncharted wonderland.

1. Introduction

The generalised policy iteration (Sutton & Barto, 2018, GPI) and trust-region learning (Schulman et al., 2015, TRL) frameworks lay the foundations for the design of the most commonly used reinforcement learning (RL) algorithms. At each GPI iteration, an RL agent first evaluates its *policy* by computing a scalar *value function*, and then updates its policy so as to maximise the value function

¹University of Oxford. Correspondence to: Jakub Grudzien Kuba <jakub.grudzien@new.ox.ac.uk>.

at every environment state. This procedure serves well for Markov Decision Problems (Bellman, 1957, MDPs) with small state spaces, as it is guaranteed to produce a new policy whose value at every state improves monotonically over the old one. However, the sizes of state spaces that many practical problem settings constitute are intractable to exact implementations of GPI. Instead, large scale settings employ function approximation and sample based learning. Unfortunately, such an adoption of GPI has proven unstable (Mnih et al., 2015), which has necessitated a number of adjustments, such as replay buffers, target networks, etc., to stabilise learning (Van Hasselt et al., 2016). In their days, these heuristics have been empirically successful but were not backed up by any theory and required extensive hyperparameter tuning (Mnih et al., 2016).

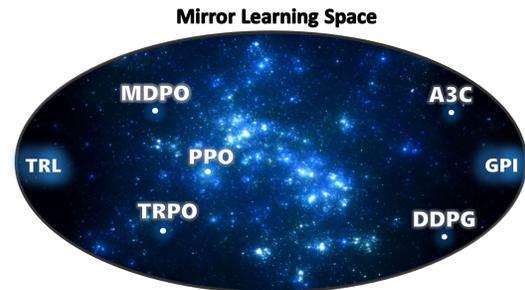


Figure 1. Known RL frameworks and algorithms as points in the infinite space of *theoretically sound* mirror learning algorithms.

Instead, TRL improves the robustness of deep RL methods by optimising a *surrogate* objective, which restricts the policy update size at each learning iteration while preserving monotonic improvement guarantees (Schulman et al., 2015). To this end it introduces a notion of distance between policies, e.g. through evaluating the maximal KL-divergence. Unfortunately, these types of measures do not scale to large MDPs that TRL was meant to tackle in the first place (since small problems can be solved by GPI).

Nevertheless, TRL’s theoretical promises and intuitive accessibility inspired a number of algorithms based on heuristic approximations thereof. This paradigm shift led to substantial empirical success: First, in *Trust-Region Policy Optimization* (Schulman et al., 2015, TRPO), a hard constraint mechanism on the policy update size was introduced. Second, *Proximal Policy Optimization* (Schulman et al., 2017, PPO) replaced the hard constraint with the *clipping objec-*

tive which, intuitively, disincentivises large policy updates. Since they were published, these algorithms have both been applied widely, resulting in state-of-the-art (SOTA) performance on a variety of benchmark, and even real-world, tasks (Schulman et al., 2015; 2017; Berner et al., 2019).

There is a stark contrast between the empirical success and the lack of theoretical understanding, which is largely limited to “proof by analogy”: For example, PPO is regarded as *sound* since it approximates an algorithm compliant with TRL. However, recent studies concluded that PPO arbitrarily exceeds policy update size constraints, thus fundamentally violating TRL principles (Wang et al., 2020; Engstrom et al., 2020), which shows that PPO’s empirical success does not follow from TRL. On a higher level, this reveals a concerning lack of theoretical understanding of the perhaps most widely used RL algorithm.

To reconnect theory with practice, in this paper we introduce a novel theoretical framework, named *Mirror Learning*, which provides theoretical guarantees to a large class of known algorithms, including TRPO (Schulman et al., 2015), PPO (Schulman et al., 2017), and MDPO (Tomar et al., 2020), to name a few, as well as to myriads of algorithms that are yet to be discovered. Mirror learning is a general, principled policy-learning framework that possesses monotonic improvement and optimal-policy convergence guarantees, under arbitrary update-size constraints. Intuitively, a mirror-learning policy update maximises the current value function while keeping the, arbitrarily specified, update cost (called *drift*) small. The update is further constrained within a *neighbourhood* of the current policy, that can also be specified arbitrarily.

Since TRL and GPI were the only anchor points for theoretically sound policy optimisation, prior unsuccessful attempts to proving the soundness of PPO and other algorithms tried to shoe-horn them into the overly narrow confines of these frameworks (Liu et al., 2019; Wang et al., 2020; Queeney et al., 2021). Mirror learning shows that this was a flawed approach: Rather than shoe-horning the algorithms, we radically expand the space of theoretically sound methods, naturally covering PPO et al in their current, practical formulations. Our work suggests that the empirical performance of these state-of-the-art methods is a direct consequence of their theoretical properties, rather than of aforementioned approximate analogies. Rather than being limited by the confines of two singular anchor points (GPI/TRL), mirror learning sets us free to explore an endless space of possible algorithms, each of which is already endowed with theoretical guarantees.

We also illustrate the explanatory power of mirror learning and use it to explain a number of thus far *unexplained* observations concerning the performance of other algorithms in the literature.

Finally, we show that mirror learning allows us to view policy optimisation as a search on a directed graph, where the total path-weight between any two nodes in an optimisation path is upper bounded by the optimality gap between them, which we confirm experimentally. We also analyse proof-of-concept instantiations of mirror learning that use a variety of different neighbourhood and drift functions.

2. Background

In this section, we introduce the RL problem formulation and briefly survey state-of-the-art learning protocols.

2.1. Preliminaries

We consider a Markov decision process (MDP) defined by a tuple $\langle \mathcal{S}, \mathcal{A}, r, P, \gamma, d \rangle$. Here, \mathcal{S} is a discrete state space, \mathcal{A} is a discrete action space¹, $r : \mathcal{S} \times \mathcal{A} \rightarrow [-R_{\max}, R_{\max}]$ is the bounded reward function, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ is the probabilistic transition function, $\gamma \in [0, 1)$ is the discount factor, and $d \in \mathcal{P}(\mathcal{S})$ (here $\mathcal{P}(X)$ denotes the set of probability distributions over a set X) is the initial state distribution. At time step $t \in \mathbb{N}$, the agent is at a state s_t , takes an action a_t according to its stationary policy $\pi(\cdot | s_t) \in \mathcal{P}(\mathcal{A})$, receives a reward $r(s_t, a_t)$, and moves to the state s_{t+1} , whose probability distribution is $P(\cdot | s_t, a_t)$. The whole experience is evaluated by the *return*, which is the discounted sum of all rewards

$$r_{\infty}^{\gamma} \triangleq \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t).$$

The state-action and state value functions that evaluate the quality of states and actions, by providing a proxy to the expected return, are given by

$$Q_{\pi}(s, a) \triangleq \mathbb{E}_{s_{1:\infty} \sim P, a_{1:\infty} \sim \pi} [r_{\infty}^{\gamma} | s_0 = s, a_0 = a],$$

$$V_{\pi}(s) \triangleq \mathbb{E}_{a_0 \sim \pi, s_{1:\infty} \sim P, a_{1:\infty} \sim \pi} [r_{\infty}^{\gamma} | s_0 = s],$$

respectively. The *advantage function*, defined as

$$A_{\pi}(s, a) \triangleq Q_{\pi}(s, a) - V_{\pi}(s),$$

estimates the advantage of selecting one action over another. The goal of the agent is to learn a policy that maximises the *expected return*, defined as a function of π ,

$$\eta(\pi) \triangleq \mathbb{E}_{s_0 \sim d, a_{0:\infty} \sim \pi, s_{0:\infty} \sim P} [r_{\infty}^{\gamma}] = \mathbb{E}_{s \sim \rho_{\pi}, a \sim \pi} [r(s, a)].$$

Here, $\rho_{\pi}(s) \triangleq \sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s | \pi)$ is the (improper) marginal state distribution. Interestingly, the set of solutions to this problem always contains the *optimal policies*—policies π^* , for which $Q_{\pi^*}(s, a) \triangleq Q^*(s, a) \geq Q_{\pi}(s, a)$ holds for any $\pi \in \Pi \triangleq \times_{s \in \mathcal{S}} \mathcal{P}(\mathcal{A})$, $s \in \mathcal{S}$, $a \in$

¹Our results extend to any compact state and action spaces. However, we work in the discrete setting in this paper for clarity.

A. Furthermore, the optimal policies are the only ones that satisfy the *optimality equation* (Sutton & Barto, 2018)

$$\pi^*(\cdot|s) = \arg \max_{\bar{\pi}(\cdot|s) \in \mathcal{P}(\mathcal{A})} \mathbb{E}_{a \sim \bar{\pi}} [Q_{\pi^*}(s, a)], \forall s \in \mathcal{S}.$$

In the next subsections, we survey the most fundamental and popular approaches of finding these optimal policies.

2.2. Generalised Policy Iteration

One key advantage of the *generalised policy iteration* (GPI) framework is its simplicity. Even though the policy influences both the rewards and state visitations, GPI guarantees that simply reacting *greedily* to the value function,

$$\pi_{\text{new}}(\cdot|s) = \arg \max_{\bar{\pi}(\cdot|s) \in \mathcal{P}(\mathcal{A})} \mathbb{E}_{a \sim \bar{\pi}} [Q_{\pi_{\text{old}}}(s, a)], \forall s \in \mathcal{S} \quad (1)$$

guarantees that the new policy obtains higher expected returns at every state, *i.e.*, $V_{\pi_{\text{new}}}(s) \geq V_{\pi_{\text{old}}}(s)$, $\forall s \in \mathcal{S}$. Moreover, this procedure converges to the set of optimal policies (Sutton & Barto, 2018), which can be seen intuitively by substituting a fixed-point policy into Equation (1).

In settings with small, discrete action spaces GPI can be executed approximately without storing the policy variable π by responding greedily to the state-action value function. This gives rise to *value-based learning* (Sutton & Barto, 2018), where the agent learns a Q-function with the Bellman-max update

$$Q_{\text{new}}(s, a) = r(s, a) + \gamma \cdot \mathbb{E}_{s' \sim \mathcal{P}} \left[\max_{a' \in \mathcal{A}} Q_{\text{old}}(s', a') \right],$$

which is known to converge to Q^* (Watkins & Dayan, 1992), and has inspired design of a number of methods (Mnih et al., 2015; Van Hasselt et al., 2016).

Another, approximate, implementation of GPI is through the policy gradient (PG) algorithms (Sutton et al., 2000). These are methods which optimise the policy π_{θ} by parameterising it with θ , and updating the parameter in the direction of the gradient of the expected return, given by

$$\nabla_{\theta} \eta(\pi_{\theta})|_{\theta=\theta_{\text{old}}} = \mathbb{E}_{s \sim \rho_{\pi_{\theta_{\text{old}}}}} \left[\nabla_{\theta} \mathbb{E}_{a \sim \pi_{\theta}} [Q_{\pi_{\theta_{\text{old}}}}(s, a)]|_{\theta=\theta_{\text{old}}} \right],$$

which is the gradient of the optimisation objective of GPI from Equation (1) weighted by $\rho_{\pi_{\theta_{\text{old}}}}$. An analogous result holds for (continuous) deterministic policies (Silver et al., 2014; Lillicrap et al., 2015). Thus, PG based algorithms approximately solve the GPI step, with a policy in the neighbourhood of $\pi_{\theta_{\text{old}}}$, provided the step-size $\alpha > 0$ in the update $\theta_{\text{new}} = \theta + \alpha \nabla_{\theta} \eta(\pi_{\theta})|_{\theta=\theta_{\text{old}}}$ is sufficiently small. PG methods have played a major role in applications of RL to the real-world settings, and especially those that involve continuous actions, where some value-based algorithms, like DQN, are intractable (Williams, 1992; Baxter & Bartlett, 2001; Mnih et al., 2016).

2.3. Trust-Region Learning

In practice, policy gradient methods may suffer from the high variance of the PG estimates and training instability (Kakade & Langford, 2002; Zhao et al., 2011). *Trust-region learning* (TRL) is a framework that aims to solve these issues. At its core lies the following policy update

$$\pi_{\text{new}} = \arg \max_{\bar{\pi} \in \Pi} \mathbb{E}_{s \sim \rho_{\pi_{\text{old}}}, a \sim \bar{\pi}} [A_{\pi_{\text{old}}}(s, a)] - C D_{\text{KL}}^{\text{max}}(\pi_{\text{old}}, \bar{\pi}),$$

where $C = 4\gamma \max_{s, a} |A_{\pi_{\text{old}}}(s, a)| / (1 - \gamma)^2$. (2)

It was shown by Schulman et al. (2015) that this update guarantees the monotonic improvement of the return, *i.e.*, $\eta(\pi_{\text{new}}) \geq \eta(\pi_{\text{old}})$. Furthermore, the KL-penalty in the above objective ensures that the new policy stays within the neighbourhood of π_{old} , referred to as the *trust region*. This is particularly important with regard to the instability issue of the PG-based algorithms (Duan et al., 2016).

Although the exact calculation of the max-KL penalty is intractable in settings with large/continuous state spaces, the algorithm can be heuristically approximated, e.g. through *Trust Region Policy Optimization* (Schulman et al., 2015, TRPO), which performs a constrained optimisation update

$$\pi_{\text{new}} = \arg \max_{\bar{\pi} \in \Pi} \mathbb{E}_{s \sim \rho_{\pi_{\text{old}}}, a \sim \bar{\pi}} [A_{\pi_{\text{old}}}(s, a)]$$

$$\text{subject to } \mathbb{E}_{s \sim \rho_{\pi_{\text{old}}}} [D_{\text{KL}}(\pi_{\text{old}}(\cdot|s), \pi_{\text{new}}(\cdot|s))] \leq \delta,$$

where $\delta > 0$. Despite its deviation from the original theory, empirical results suggest that TRPO approximately maintains the TRL properties. However, in order to achieve a simpler TRL-based heuristic, Schulman et al. (2017) introduced *Proximal Policy Optimization* (PPO) which updates the policy by

$$\pi_{\text{new}} = \arg \max_{\bar{\pi} \in \Pi} \mathbb{E}_{s \sim \rho_{\pi_{\text{old}}}, a \sim \pi_{\text{old}}} [L^{\text{clip}}], \quad (3)$$

$$L^{\text{clip}} = \min \left(r(\bar{\pi}) A_{\pi_{\text{old}}}(s, a), \text{clip}(r(\bar{\pi}), 1 \pm \epsilon) A_{\pi_{\text{old}}}(s, a) \right),$$

where for given s, a , $r(\bar{\pi}) \triangleq \bar{\pi}(a|s) / \pi(a|s)$. The *clip* operator truncates $r(\bar{\pi})$ to $1 - \epsilon$ (or $1 + \epsilon$), if it is below (or above) the threshold interval. Despite being motivated from a TRL perspective, PPO violates its very core principles by failing to constrain the update size—whether measured either by KL divergence, or by the likelihood ratios (Wang et al., 2020; Engstrom et al., 2020). Nevertheless, PPO’s ability to stabilise policy training has been demonstrated on a wide variety of tasks, frequently resulting in SOTA performance (Henderson et al., 2018; Berner et al., 2019). This begs the question how PPO’s empirical performance can be justified theoretically. Mirror learning, which we introduce in the next section, addresses this issue.

3. Mirror Learning

In this section we introduce *Mirror Learning*, state its theoretical properties, and explore its connection to prior RL theory and methodology.

3.1. The Framework

We start from the following definition.

Definition 3.1. A *drift functional* \mathfrak{D} is a map

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

such that for all $s \in \mathcal{S}$, and $\pi, \bar{\pi} \in \Pi$, writing $\mathfrak{D}_\pi(\bar{\pi}|s)$ for $\mathfrak{D}_\pi(\bar{\pi}(\cdot|s)|s)$, the following conditions are met

1. $\mathfrak{D}_\pi(\bar{\pi}|s) \geq \mathfrak{D}_\pi(\pi|s) = 0$ (*nonnegativity*),
2. $\mathfrak{D}_\pi(\bar{\pi}|s)$ has zero gradient² with respect to $\bar{\pi}(\cdot|s)$, evaluated at $\bar{\pi}(\cdot|s) = \pi(\cdot|s)$ (*zero gradient*).

Let $\nu_{\bar{\pi}} \in \mathcal{P}(\mathcal{S})$ be a state distribution that can depend on π and $\bar{\pi}$. The *drift* \mathfrak{D}^ν of $\bar{\pi}$ from π is defined as

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

$\nu_{\bar{\pi}}$ is required to be such that the above expectation is continuous in π and $\bar{\pi}$. The drift is *positive* if $\mathfrak{D}^\nu_{\bar{\pi}}(\bar{\pi}) = 0$ implies $\bar{\pi} = \pi$, and *trivial* if $\mathfrak{D}^\nu_{\bar{\pi}}(\bar{\pi}) = 0, \forall \pi, \bar{\pi} \in \Pi$.

We would like to highlight that drift is not the Bregman distance, associated with mirror descent (Nemirovskij & Yudin, 1983; Beck & Teboulle, 2003), as we do not require it to be (strongly) convex, or differentiable everywhere. All we require are, the much milder, continuity and Gâteaux-differentiability (Gateaux, 1922; Hamilton & Nashed, 1982) of the drift functional at $\bar{\pi}(\cdot|s) = \pi(\cdot|s)$.

We introduce one more concept whose role is to generally account for explicit update-size constraints. Such constraints reflect a learner’s risk-aversity towards substantial changes to its behaviour, like in TRPO, or are dictated by an algorithm design, like the learning rate in PG methods.

Definition 3.2. We say that $\mathcal{N} : \Pi \rightarrow \mathbb{P}(\Pi)$ is a *neighbourhood operator*, where $\mathbb{P}(\Pi)$ is the power set of Π , if

1. It is a continuous map (*continuity*),
2. Every $\mathcal{N}(\pi)$ is a compact set (*compactness*),
3. There exists a metric $\chi : \Pi \times \Pi \rightarrow \mathbb{R}$, such that $\forall \pi \in \Pi$, there exists $\zeta > 0$, such that $\chi(\pi, \bar{\pi}) \leq \zeta$ implies $\bar{\pi} \in \mathcal{N}(\pi)$ (*closed ball*).

The *trivial* neighbourhood operator is $\mathcal{N} \equiv \Pi$.

Let \mathfrak{D}^ν be a drift, and $\pi, \bar{\pi}$ be policies. Suppose that $\beta_\pi \in \mathcal{P}(\mathcal{S})$ is a state distribution, referred to as a *sampling distribution*, such that $\beta_\pi(s) > 0, \forall s \in \mathcal{S}$. Then, the *mirror operator* transforms the value function of π into the following functional of $\bar{\pi}$,

$$[\mathcal{M}_{\mathfrak{D}}^{\bar{\pi}} V_\pi](s) = \mathbb{E}_{a \sim \bar{\pi}} [Q_\pi(s, a)] - \frac{\nu_{\bar{\pi}}(s)}{\beta_\pi(s)} \mathfrak{D}_\pi(\bar{\pi}|s).$$

²More precisely, all its Gâteaux derivatives are zero.

Note that, in the above definition, Q_π can be equivalently replaced by A_π , as this only subtracts a constant $V_\pi(s)$ independent of $\bar{\pi}$. We will use this fact later. As it turns out, despite the appearance of the drift penalty, simply acting to increase the mirror operator suffices to guarantee the policy improvement, as summarised by the following lemma, proved in Appendix B.

Lemma 3.3. Let π_{old} and π_{new} be policies. Suppose that

$$[\mathcal{M}_{\mathfrak{D}}^{\pi_{new}} V_{\pi_{old}}](s) \geq [\mathcal{M}_{\mathfrak{D}}^{\pi_{old}} V_{\pi_{old}}](s), \forall s \in \mathcal{S}. \quad (4)$$

Then, π_{new} is better than π_{old} , so that for every state s ,

$$V_{\pi_{new}}(s) \geq V_{\pi_{old}}(s).$$

Of course, the monotonic improvement property of the expected return is a natural corollary of this lemma, as

$$\begin{aligned} \eta(\pi_{new}) &= \mathbb{E}_{s \sim d} [V_{\pi_{new}}(s)] \\ &\geq \mathbb{E}_{s \sim d} [V_{\pi_{old}}(s)] = \eta(\pi_{old}). \end{aligned}$$

Hence, optimisation of the mirror operator guarantees the improvement of the new policy’s performance. Condition (4), however, is represented by $|\mathcal{S}|$ inequalities and solving them may be intractable in large state spaces. Hence, we shall design a proxy objective whose solution simply satisfies Condition (4), and admits Monte-Carlo estimation (for practical reasons). For this purpose, we define the update rule of mirror learning.

Definition 3.4. Let π_{old} be a policy. Then, **mirror learning** updates the policy by

$$\pi_{new} = \arg \max_{\bar{\pi} \in \mathcal{N}(\pi_{old})} \mathbb{E}_{s \sim \beta_\pi} \left[[\mathcal{M}_{\mathfrak{D}}^{\bar{\pi}} V_{\pi_{old}}](s) \right]. \quad (5)$$

This, much simpler and clearer, problem formulation, is also robust to settings with large state spaces and parameterised policies. Being an expectation, it can be approximated by sampling with the unbiased “batch” estimator

$$\frac{1}{|\mathfrak{B}|} \sum_{s, a \in \mathfrak{B}} \left[\frac{\bar{\pi}(a|s)}{\pi_{old}(a|s)} Q_{\pi_{old}}(s, a) - \frac{\nu_{\bar{\pi}}(s)}{\beta_{\pi_{old}}(s)} \mathfrak{D}_{\pi_{old}}(\bar{\pi}|s) \right]. \quad (6)$$

The distribution $\nu_{\bar{\pi}}^{\pi_{old}}$ can be chosen equal to $\beta_{\pi_{old}}$, in which case the fraction from the front of the drift disappears, simplifying the estimator. Note also that one may choose $\beta_{\pi_{old}}$ independently of π_{old} . For example, it can be some fixed distribution over \mathcal{S} , like the uniform distribution. Hence, the above update supports **both on-policy and off-policy** learning. In the latter case, the fraction $\bar{\pi}(a|s)/\pi_{old}(a|s)$ is replaced by $\bar{\pi}(a|s)/\pi_{hist}(a|s)$, where π_{hist} is the policy used to sample the pair (s, a) and insert it to a replay buffer (for more details see Appendix F). Such a Monte-Carlo objective can be optimised with respect to $\bar{\pi}$ parameters by, for example, a few steps of gradient ascent. Most importantly, in its exact form, the update in Equation (5) guarantees that the resulting policy satisfies the desired Condition (4).

Lemma 3.5. *Let π_{old} be a policy and π_{new} be obtained from the mirror update of Equation (5). Then,*

$$[\mathcal{M}_{\mathcal{D}}^{\pi_{new}} V_{\pi_{old}}](s) \geq [\mathcal{M}_{\mathcal{D}}^{\pi_{old}} V_{\pi_{old}}](s), \forall s \in \mathcal{S}.$$

Hence, π_{new} attains the properties provided by Lemma 3.3.

For proof, see Appendix B. Next, we use the above results to characterise the properties of mirror learning. The improvement and convergence properties that it exhibits not only motivate its usage, but may also contribute to explaining the empirical success of various widely-used algorithms. We provide a proof sketch below, and a detailed proof in Appendix C.

Theorem 3.6 (The Fundamental Theorem of Mirror Learning). *Let \mathcal{D}^ν be a drift, \mathcal{N} be a neighbourhood operator, and the sampling distribution β_π depend continuously on π . Let $\pi_0 \in \Pi$, and the sequence of policies $(\pi_n)_{n=0}^\infty$ be obtained by mirror learning induced by \mathcal{D}^ν , \mathcal{N} , and β_π . Then, the learned policies*

1. *Attain the strict monotonic improvement property,*

$$\eta(\pi_{n+1}) \geq \eta(\pi_n) + \mathbb{E}_{s \sim d} \left[\frac{\nu_{\pi_{n+1}}(s)}{\beta_{\pi_n}(s)} \mathcal{D}_{\pi_n}(\pi_{n+1}|s) \right],$$

2. *Their value functions converge to the optimal one,*

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

3. *Their expected returns converge to the optimal return,*

$$\lim_{n \rightarrow \infty} \eta(\pi_n) = \eta^*,$$

4. *Their ω -limit set consists of the optimal policies.*

Proof sketch. We split the proof into four steps. In *Step 1*, we use Lemmas 3.3 & 3.5 to show that the sequence of value functions $(V_{\pi_k})_{k \in \mathbb{N}}$ converges. In *Step 2*, we show the existence of limit points $\bar{\pi}$ of $(\pi_k)_{k \in \mathbb{N}}$, and show that they are fixed points of the mirror learning update, which we do by contradiction. The most challenging *Step 3*, is where we prove (by contradiction) that $\bar{\pi}$ is a fixed point of GPI. In the proof, we supposed that $\bar{\pi}$ is not a fixed point of GPI, and use the drift's zero-gradient property (Definition 3.1) to show that one could slightly perturb $\bar{\pi}$ to obtain a policy π' , which corresponds to a higher mirror learning objective. This contradicts $\bar{\pi}$ simultaneously being a fixed point of mirror learning. *Step 4* finalises the proof, by recalling that fixed points of GPI are optimal policies. \square

Hence, any algorithm whose update takes the form of Equation (5) improves the return monotonically, and converges to the set of optimal policies. This result provides

RL algorithm designers with a template. New instances of it can be obtained by altering the drift \mathcal{D}^ν , the neighbourhood operator \mathcal{N} , and the sampling distribution function β_π . Indeed, in the next section, we show how some of the most well-known RL algorithms fit this template.

4. Mirror Learning View of RL Phenomena

We provide a list of RL algorithms in their mirror learning representation in Appendix E.

Generalised Policy Iteration For the trivial drift $\mathcal{D}^\nu \equiv 0$, and the trivial neighbourhood operator $\mathcal{N} \equiv \Pi$, the mirror learning update from Equation (5) is equivalent to

$$\pi_{new} = \arg \max_{\bar{\pi} \in \Pi} \mathbb{E}_{s \sim \beta_{\pi_{old}}} [\mathbb{E}_{a \sim \bar{\pi}} [Q_{\pi_{old}}(s, a)]].$$

As in this case the maximisation is unconstrained, and the expectation over $s \sim \beta_{\pi_{old}}$ is monotonically increasing in the individual conditional $\mathbb{E}_{a \sim \bar{\pi}} [Q_{\pi_{old}}(s, a)]$, the maximisation distributes across states

$$\pi_{new}(\cdot|s) = \arg \max_{\bar{\pi}(\cdot|s) \in \mathcal{P}(\mathcal{A})} \mathbb{E}_{a \sim \bar{\pi}} [Q_{\pi_{old}}(s, a)],$$

which is exactly the GPI update, as in Equation (1). Hence, all instances of GPI, *e.g.*, policy iteration (Sutton & Barto, 2018), are special cases of mirror learning, and thus inherit its improvement and convergence properties. Furthermore, neural implementations, like A3C (Mnih et al., 2016), maintain these qualities approximately.

Trust-Region Learning Let us choose the drift operator \mathcal{D} to be the scaled KL-divergence, so that $\mathcal{D}_\pi(\bar{\pi}|s) = C_\pi \mathbf{D}_{\text{KL}}(\pi(\cdot|s), \bar{\pi}(\cdot|s))$, where C_π is a constant that depends on π . Further, in the construction of the drift $\mathcal{D}_\pi^\nu(\bar{\pi})$, let us set $\nu_{\bar{\pi}}(s) = \delta(s - s_{\text{max}})$, where δ is the Dirac-delta distribution, and s_{max} is the state at which the KL-divergence between π and $\bar{\pi}$ is largest. For the neighbourhood operator, we choose the trivial one. Lastly, for the sampling distribution, we choose $\beta_\pi = \bar{\rho}_\pi$, the normalised version of ρ_π . Then, a mirror learning step maximises

$$\mathbb{E}_{s \sim \bar{\rho}_{\pi_{old}}} \left[\mathbb{E}_{a \sim \pi_{new}} [A_{\pi_{old}}(s, a)] \right] - C_{\pi_{old}} \mathbf{D}_{\text{KL}}^{\text{max}}(\pi_{old}, \pi_{new}),$$

which, for appropriately chosen $C_{\pi_{old}}$, is proportional (and thus equivalent) to the trust-region learning update from Equation (2). Hence, the monotonic improvement of trust-region learning follows from Theorem 3.6, which also implies its convergence to the set of optimal policies.

TRPO The algorithm is designed to approximate trust-region learning, and so its mirror-learning representation is similar. We make the following changes: set the drift operator to $\mathcal{D}^\nu \equiv 0$, and choose the neighbourhood operator to

the *average-KL ball*. Precisely,

$$\mathcal{N}(\pi) = \left\{ \bar{\pi} \in \Pi \mid \mathbb{E}_{s \sim \bar{\rho}_\pi} \left[\text{D}_{\text{KL}}(\pi(\cdot|s), \bar{\pi}(\cdot|s)) \right] \leq \delta \right\}.$$

The resulting mirror learning update is the learning rule of the TRPO algorithm. As a result, even though TRPO was supposed to only approximate the monotonic trust-region learning update (Schulman et al., 2015), this analysis shows that TRPO has monotonic convergence guarantees.

PPO We analyse PPO through unfolding the clipping objective L^{clip} (Equation (3)). For a given $s \in \mathcal{S}$, it equals

$$\mathbb{E}_{\mathbf{a} \sim \pi_{\text{old}}} \left[\min \left(r(\bar{\pi}) A_{\pi_{\text{old}}}(s, \mathbf{a}), \text{clip}(r(\bar{\pi}), 1 \pm \epsilon) A_{\pi_{\text{old}}}(s, \mathbf{a}) \right) \right].$$

By recalling that $r(\bar{\pi}) = \bar{\pi}(\mathbf{a}|s)/\pi_{\text{old}}(\mathbf{a}|s)$, we use importance sampling (Sutton & Barto, 2018), and the trick of “adding and subtracting”, to rewrite L^{clip} as

$$\begin{aligned} & \mathbb{E}_{\mathbf{a} \sim \bar{\pi}} \left[A_{\pi_{\text{old}}}(s, \mathbf{a}) \right] - \mathbb{E}_{\mathbf{a} \sim \pi_{\text{old}}} \left[r(\bar{\pi}) A_{\pi_{\text{old}}}(s, \mathbf{a}) \right. \\ & \left. - \min \left(r(\bar{\pi}) A_{\pi_{\text{old}}}(s, \mathbf{a}), \text{clip}(r(\bar{\pi}), 1 \pm \epsilon) A_{\pi_{\text{old}}}(s, \mathbf{a}) \right) \right]. \end{aligned}$$

Going forward, we focus on the expectation that is subtracted. First, we can replace the min operator with max, with the identity $-\min f(x) = \max[-f(x)]$, as follows

$$\begin{aligned} & \mathbb{E}_{\mathbf{a} \sim \pi_{\text{old}}} \left[r(\bar{\pi}) A_{\pi_{\text{old}}}(s, \mathbf{a}) \right. \\ & \left. + \max \left(-r(\bar{\pi}) A_{\pi_{\text{old}}}(s, \mathbf{a}), -\text{clip}(r(\bar{\pi}), 1 \pm \epsilon) A_{\pi_{\text{old}}}(s, \mathbf{a}) \right) \right]. \end{aligned}$$

Then, we move $r(\bar{\pi}) A_{\pi_{\text{old}}}(s, \mathbf{a})$ inside the max, and obtain

$$\mathbb{E}_{\mathbf{a} \sim \pi_{\text{old}}} \left[\max \left(0, [r(\bar{\pi}) - \text{clip}(r(\bar{\pi}), 1 \pm \epsilon)] A_{\pi_{\text{old}}}(s, \mathbf{a}) \right) \right],$$

which can be simplified as

$$\mathbb{E}_{\mathbf{a} \sim \pi_{\text{old}}} \left[\text{ReLU} \left([r(\bar{\pi}) - \text{clip}(r(\bar{\pi}), 1 \pm \epsilon)] A_{\pi_{\text{old}}}(s, \mathbf{a}) \right) \right] \quad (7)$$

Notice now that this expression is always non-negative, due to the presence of the ReLU function (Fukushima & Miyake, 1982). Furthermore, for $\bar{\pi}$ sufficiently close to π_{old} , *i.e.*, so that $r(\bar{\pi}) \in [1 - \epsilon, 1 + \epsilon]$, the clip operator reduces to the identity function, and so Equation (7) is constant and zero. Hence, it is zero at $\bar{\pi} = \pi_{\text{old}}$, and has a zero gradient. Therefore, the expression in Equation (7) is a drift functional of $\bar{\pi}(\cdot|s)$. This, together with taking $\beta_{\pi_{\text{old}}} \equiv \nu_{\pi_{\text{old}}}^{\bar{\pi}} \equiv \bar{\rho}_{\pi_{\text{old}}}$, and the trivial neighbourhood operator $\mathcal{N} \equiv \Pi$, shows that PPO, while it was supposed to be a heuristic approximation of trust-region learning, is by itself a rigorous instance of mirror learning. Thus, it inherits the monotonic improvement and convergence properties, which helps explain its great performance.

Interestingly, during the development of PPO, a variant more closely related to TRL was considered (Schulman

et al., 2017). Known as PPO-KL, the algorithm updates the policy to $\bar{\pi} = \pi_{\text{new}}$ to maximise

$$\mathbb{E}_{s \sim \rho_{\pi_{\text{old}}}, \mathbf{a} \sim \bar{\pi}} \left[A_{\pi_{\text{old}}}(s, \mathbf{a}) \right] - \tau \bar{\text{D}}_{\text{KL}}(\pi_{\text{old}}, \bar{\pi}),$$

and then scales τ up or down by a constant (typically 1.5), depending on whether the KL-divergence induced by the update exceeded or subceeded some target level. Intriguingly, according to the authors this approach, although more closely-related to TRL, failed in practice. Mirror learning explains this apparent paradox. Namely, the rescaling scheme of τ causes discontinuity of the penalty term, when viewed as a function of π_{old} and $\bar{\pi}$. This prevents the penalty from being a valid drift, so unlike PPO this version, PPO-KL, is not an instance of mirror learning. Recently, Hsu et al. (2020) introduced PPO-KL with a fixed τ , allowing for continuity of the KL-penalty. Such an algorithm is an instance of mirror learning, independently of whether the *forward* or *backward* KL-divergence is employed. And indeed, the authors find that both versions of the algorithms result in strong empirical performance, further validating our theory.

5. Related Work

The RL community has long worked on the development of theory to aid the development of theoretically-sound techniques. Perhaps, one of the greatest achievements along this thread is the development of trust-region learning (Schulman et al., 2015, TRL)—a framework that allows for stable training of monotonically improving policies. Unfortunately, despite its great theoretical guarantees, TRL is intractable in most practical settings. Hence, the RL community focused on the development of heuristics, like TRPO (Schulman et al., 2015) and PPO (Schulman et al., 2017), that approximate it while trading off theoretical guarantees of TRL for practicality. As these methods established new state-of-the-art performance on a variety of tasks (Duan et al., 2016; Berner et al., 2019), the conceptual connection to TRL has been considered the key to success for many algorithms (Arjona-Medina et al., 2018; Queeney et al., 2021). However, recent works have shown that PPO is prone to breaking the core TRL principles of constraining the update size (Wang et al., 2020; Engstrom et al., 2020), and thus revealed an existing chasm between RL practice and understanding—a problem that mirror learning resolves.

Trust-region algorithms, although original, inspire connections between RL and optimisation, where the method of *mirror descent* has been studied (Nemirovskij & Yudin, 1983). This idea has recently been extended upon in (Tomar et al., 2020), where *Mirror Descent Policy Optimization* (MDPO) was proposed—the algorithm optimises the policy with mirror ascent, where the role of the Bregman distance is played by the KL-divergence. The method

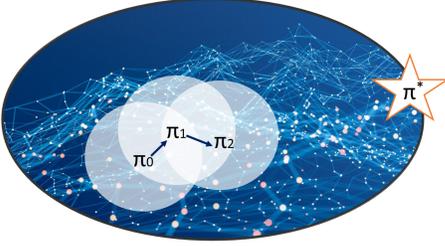


Figure 2. An intuitive view on the policy DAG and initial steps of mirror learning. A policy vertex has a neighbour, within its neighbourhood, which improves the return.

has been shown to achieve great empirical performance (which is also implied by mirror learning). Notably, a new stream of research in regularised RL is arising, where the regulariser is subsumed into the Bregman distance. For example, Neu et al. (2017) have shown that, in the average-reward setting, a variant of TRPO is an instance of mirror descent in a regularised MDP, and converges to the optimal policy. Shani et al. (2020) have generalised TRPO to handle regularised problems, and derived convergence rates for their methods in the tabular case. Lan (2021) and Zhan et al. (2021) have proposed mirror-descent algorithms that solve regularised MDPs with convex regularisers, and also provided their convergence properties. Here, we would like to highlight that mirror learning is not a method of solving regularised problems through mirror descent. Instead, it is a very general class of algorithms that solve the classical MDP. The term *mirror*, however, is inspired by the intuition behind mirror descent, which solves the image of the original problem under the *mirror map* (Nemirovskij & Yudin, 1983)—similarly, we defined the mirror operator.

Mirror learning is also related to *Probability Functional Descent* (Chu et al., 2019, PFD). Although PFD is a general framework of probability distribution optimisation in machine learning problems, in the case of RL, it is an instance of mirror learning—the one recovering GPI. Lastly, the concepts of mirror descent and functional policy representation are connected in a concurrent work of Vaswani et al. (2021), who show how the technique of mirror descent implements the functional descent of parameterised policies. This approach, although powerful, fails to capture some algorithms, including PPO, which we prove to be an instance of mirror learning. The key to the generalisation power of our theory is its abstractness, as well as simplicity.

6. Graph-theoretical Interpretation

In this section we use mirror learning to make a surprising connection between RL and graph theory. We begin by introducing the following definition of a particular *directed acyclic graph* (DAG).

Definition 6.1 (Policy DAG). Let \mathfrak{D}^ν be a positive drift, and \mathcal{N} be a neighbourhood operator. Then, the *policy DAG*

$\mathcal{G}(\Pi, \mathfrak{D}^\nu, \mathcal{N})$ is a graph where

- The vertex set is the policy space Π ,
- (π_1, π_2) is an edge if $\eta(\pi_1) < \eta(\pi_2)$ and $\pi_2 \in \mathcal{N}(\pi_1)$,
- The weight of an edge (π_1, π_2) is $\mathfrak{D}_{\pi_1}^\nu(\pi_2)$.

This graph is a valid DAG because the transitive, asymmetric “ $<$ ” relation that builds edges prevents the graph from having cycles. We also know that, for every non-optimal policy π , there in an outgoing edge (π, π') from π , as π' can be computed with a step of mirror learning (see Figure 2 for an intuitive picture).

The above definition allows us to cast mirror learning as a graph search problem. Namely, let $\pi_0 \in \Pi$ be a vertex policy at which we initialise the search, which further induces a sequence $(\pi_n)_{n=0}^\infty$. Let us define $\mathcal{U}_\beta = \min_{\pi \in \Pi, s \in \mathcal{S}} d(s)/\beta(s)$. Theorem 3.6 lets us upper-bound the weight of any traversed edge as

$$\eta(\pi_{n+1}) - \eta(\pi_n) \geq \mathcal{U}_\beta \cdot \mathfrak{D}_{\pi_n}^\nu(\pi_{n+1}). \quad (8)$$

Notice that $\eta(\pi_n) - \eta(\pi_0) = \sum_{i=1}^n [\eta(\pi_i) - \eta(\pi_{i-1})]$ converges to $\eta^* - \eta(\pi_0)$. Hence, the following series expansion

$$\eta^* - \eta(\pi_0) = \sum_{n=0}^{\infty} [\eta(\pi_{n+1}) - \eta(\pi_n)]$$

is valid. Combining it with Inequality (8), we obtain a bound on the total weight of the path induced by the search

$$\frac{\eta^* - \eta(\pi_0)}{\mathcal{U}_\beta} \geq \sum_{n=0}^{\infty} \mathfrak{D}_{\pi_n}^\nu(\pi_{n+1}), \quad (9)$$

Hence, mirror learning finds a path from π_0 to the set of graph sinks (the optimal policies), whose total weight is finite, and does not depend on the policies that appeared during the search. Note that one could also estimate the left-hand side from above by a bound $(\eta^* + V_{\max})/\mathcal{U}_\beta$ which is uniform for all initial policies π_0 . This finding may be considered counterintuitive: While we can be decreasing the number of edges in the graph by shrinking the neighbourhood operator $\mathcal{N}(\pi)$, a path to π^* still exists, and despite (perhaps) containing more edges, its total weight remains bounded.

Lastly, we believe that this inequality can be exploited for practical purposes: the drift functional \mathfrak{D} is an abstract hyper-parameter, and the choice of it is a part of algorithm design. Practitioners can choose \mathfrak{D} to describe a cost that they want to limit throughout training. For example, one can set $\mathfrak{D} = \text{risk}$, where $\text{risk}_\pi(\bar{\pi}|s)$ quantifies some notion of risk of updating $\pi(\cdot|s)$ to $\bar{\pi}(\cdot|s)$, or $\mathfrak{D} = \text{memory}$, to only make updates at a reasonable memory expense. Inequality (9) guarantees that the total expense will remain finite, and provides an upper bound for it. Thereby, we encourage employing mirror learning with drifts designed to satisfy constraints of interest.

7. Numerical Experiments

We verify the correctness of our theory with numerical experiments. Their purpose is not to establish a new state-of-the-art performance in the most challenging deep RL benchmarks. It is, instead, to demonstrate that algorithms that fall into the mirror learning framework obey Theorem 3.6 and Inequality (9). Hence, to enable a close connection between the theory and experiments we choose simple environments and for drift functionals we selected: KL-divergence, squared L2 distance, squared total variation distance, and the trivial (zero) drift. For the neighbourhood operator, in one suite of experiments, we use the expected drift ball, *i.e.*, the set of policies within some distance from the old policy, measured by the corresponding drift; in the second suite, we use the KL ball neighbourhood, like in TRPO. We represent the policies with n -dimensional action spaces as *softmax* distributions, with $n - 1$ parameters. For an exact verification of the theoretical results, we test each algorithm over *only one* random seed. In all experiments, we set the initial-state and sampling distributions to uniform. The code is available at <https://github.com/znowu/mirror-learning>.

Single-step Game. In this game, the agent chooses one of 5 actions, and receives a reward corresponding to it. These rewards are 10, 0, 1, 0, 5, respectively for each action. The optimal return in this game equals 10.

Tabular Game. This game has 5 states, lined up next to each other. At each state, an agent can choose to go left, and receive a reward of +0.1, stay at the current state and receive 0 reward, or go right and receive -0.1 reward. However, if the agent goes left at the left-most state, it receives -10 reward, and if it goes right at the right-most state, it receives +10 reward. In these two cases, the game terminates. We set $\gamma = 0.999$. Clearly, the optimal policy here ignores the small intermediate rewards and always chooses to go right. The corresponding optimal expected return is approximately 9.7.

GridWorld. We consider a 5×5 grid, with a barrier that limits the agent’s movement. The goal of the agent is to reach the top-right corner as quick as possible, which terminates the episode, and to avoid the bottom-left corner with a bomb. It receives -1 reward for every step taken, and -100 for stepping onto the bomb and terminating the game. For $\gamma = 0.999$, the optimal expected return is approximately -7 .

Figure 3 confirms that the resulting algorithms achieve the monotonic improvement property, and converges to the optimal returns—this confirms Theorem 3.6. In these simple environments, the learning curves are influenced by the choice of the drift functional more than by the neighbourhood operator, although not significantly. An exception oc-

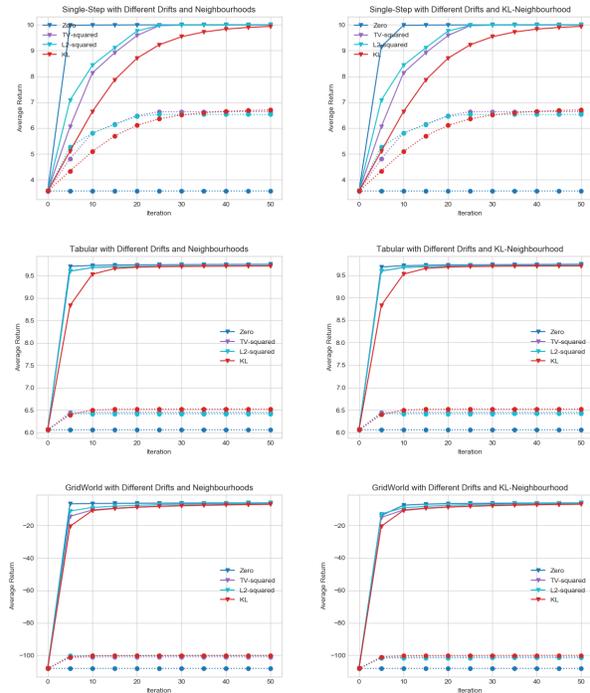


Figure 3. Mirror learning algorithms with different drifts and neighbourhood operators tested on simple environments. The solid lines represent the return, and the dotted ones represent the total drift. Algorithms in the left column use the drift ball neighbourhood, while those in the right use the KL ball neighbourhood. In both columns there are algorithms with each of the aforementioned drifts. Results are for one seed per environment and algorithm.

currs in the single-step game where, of course, one step of GPI is sufficient to solve the game. We also see that the value of the total drift (which we shift by V_{π_0} on the plots for comparison) remains well below the return in all environments, confirming Inequality (9).

8. Conclusion

In this paper, we introduced *mirror learning*—a framework which unifies existing policy iteration algorithms. We have proved Theorem 3.6, which states that any mirror learning algorithm solves the reinforcement learning problem. As a corollary to this theorem, we obtained convergence guarantees of state-of-the-art methods, including PPO and TRPO. More importantly, it provides a framework for the future development of theoretically-sound algorithms. We also proposed an interesting, graph-theoretical perspective on mirror learning, which establishes a connection between graph theory and RL. Lastly, we verified the correctness of our theoretical results through numerical experiments on a diverse family of mirror learning instances in three simple toy settings. Designing and analysing the properties of the myriads of other possible mirror learning instances is an exciting avenue for future research.

Acknowledgements

I dedicate this work to my little brother *Michał*, and thank him for persistently bringing light to my life.

I have to go now. We won't see each other for a while, but one day we'll get together again. I promise.

Kuba

References

- Arjona-Medina, J. A., Gillhofer, M., Widrich, M., Unterthiner, T., Brandstetter, J., and Hochreiter, S. Ruder: Return decomposition for delayed rewards. *arXiv preprint arXiv:1806.07857*, 2018.
- Ausubel, L. M. and Deneckere, R. J. A generalized theorem of the maximum. *Economic Theory*, 3(1):99–107, 1993.
- Baxter, J. and Bartlett, P. L. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.
- Beck, A. and Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.
- Bellman, R. A markov decision process. *journal of mathematical mechanics*. 1957.
- Berner, C., Brockman, G., Chan, B., Cheung, V., Debiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.
- Chu, C., Blanchet, J., and Glynn, P. Probability functional descent: A unifying perspective on gans, variational inference, and reinforcement learning. In *International Conference on Machine Learning*, pp. 1213–1222. PMLR, 2019.
- Duan, Y., Chen, X., Houthoofd, R., Schulman, J., and Abbeel, P. Benchmarking deep reinforcement learning for continuous control. In *International conference on machine learning*, pp. 1329–1338. PMLR, 2016.
- Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Janoos, F., Rudolph, L., and Madry, A. Implementation matters in deep policy gradients: A case study on ppo and trpo. In *International Conference on Learning Representations*, 2020.
- Fukushima, K. and Miyake, S. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pp. 267–285. Springer, 1982.
- Gateaux, R. Sur diverses questions de calcul fonctionnel. *Bulletin de la Société Mathématique de France*, 50:1–37, 1922.
- Hamilton, E. and Nashed, M. Global and local variational derivatives and integral representations of gâteaux differentials. *Journal of Functional Analysis*, 49(1):128–144, 1982.
- Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., and Meger, D. Deep reinforcement learning that matters. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
- Hsu, C. C.-Y., Mendl-Dünner, C., and Hardt, M. Revisiting design choices in proximal policy optimization. *arXiv preprint arXiv:2009.10897*, 2020.
- Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.
- Kuba, J. G., Chen, R., Wen, M., Wen, Y., Sun, F., Wang, J., and Yang, Y. Trust region policy optimisation in multi-agent reinforcement learning. *arXiv preprint arXiv:2109.11251*, 2021.
- Lan, G. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *arXiv preprint arXiv:2102.00135*, 2021.
- Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.
- Liu, B., Cai, Q., Yang, Z., and Wang, Z. Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv preprint arXiv:1906.10306*, 2019.
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.
- Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pp. 1928–1937. PMLR, 2016.
- Nemirovskij, A. S. and Yudin, D. B. Problem complexity and method efficiency in optimization. 1983.
- Neu, G., Jonsson, A., and Gómez, V. A unified view of entropy-regularized markov decision processes. *CoRR*, abs/1705.07798, 2017. URL <http://arxiv.org/abs/1705.07798>.

- Queeney, J., Paschalidis, I., and Cassandra, C. Generalized proximal policy optimization with sample reuse. *Advances in Neural Information Processing Systems*, 34, 2021.
- Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347, 2017.
- Shani, L., Efroni, Y., and Mannor, S. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5668–5675, 2020.
- Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. Deterministic policy gradient algorithms. In *International conference on machine learning*, pp. 387–395. PMLR, 2014.
- Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. 2018.
- Sutton, R. S., Mcallester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems 12*, volume 12, pp. 1057–1063. MIT Press, 2000.
- Tomar, M., Shani, L., Efroni, Y., and Ghavamzadeh, M. Mirror descent policy optimization. *arXiv preprint arXiv:2005.09814*, 2020.
- Van Hasselt, H., Guez, A., and Silver, D. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.
- Vaswani, S., Bachem, O., Totaro, S., Mueller, R., Geist, M., Machado, M. C., Castro, P. S., and Roux, N. L. A functional mirror ascent view of policy gradient methods with function approximation. *arXiv preprint arXiv:2108.05828*, 2021.
- Wang, Y., He, H., and Tan, X. Truly proximal policy optimization. In *Uncertainty in Artificial Intelligence*, pp. 113–122. PMLR, 2020.
- Watkins, C. J. C. H. and Dayan, P. Q-learning. *Machine Learning*, 8(3):279–292, May 1992. ISSN 1573-0565. doi: 10.1007/BF00992698. URL <https://doi.org/10.1007/BF00992698>.
- Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.
- Zhan, W., Cen, S., Huang, B., Chen, Y., Lee, J. D., and Chi, Y. Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence. *arXiv preprint arXiv:2105.11066*, 2021.
- Zhao, T., Hachiya, H., Niu, G., and Sugiyama, M. Analysis and improvement of policy gradient estimation. In *NIPS*, pp. 262–270. Citeseer, 2011.

A. Definition Details

Definition 3.1. A *drift functional* \mathfrak{D} is a map

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

such that for all $s \in \mathcal{S}$, and $\pi, \bar{\pi} \in \Pi$, writing $\mathfrak{D}_\pi(\bar{\pi}|s)$ for $\mathfrak{D}_\pi(\bar{\pi}(\cdot|s)|s)$, the following conditions are met

1. $\mathfrak{D}_\pi(\bar{\pi}|s) \geq \mathfrak{D}_\pi(\pi|s) = 0$ (*nonnegativity*),
2. $\mathfrak{D}_\pi(\bar{\pi}|s)$ has zero gradient³ with respect to $\bar{\pi}(\cdot|s)$, evaluated at $\bar{\pi}(\cdot|s) = \pi(\cdot|s)$ (*zero gradient*).

Let $\nu_{\bar{\pi}} \in \mathcal{P}(\mathcal{S})$ be a state distribution that can depend on π and $\bar{\pi}$. The *drift* \mathfrak{D}^ν of $\bar{\pi}$ from π is defined as

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

$\nu_{\bar{\pi}}$ is required to be such that the above expectation is continuous in π and $\bar{\pi}$. The drift is *positive* if $\mathfrak{D}_\pi^\nu(\bar{\pi}) = 0$ implies $\bar{\pi} = \pi$, and *trivial* if $\mathfrak{D}_\pi^\nu(\bar{\pi}) = 0, \forall \pi, \bar{\pi} \in \Pi$.

In this definition, the notion of the gradient of $\mathfrak{D}_\pi(\bar{\pi}|s)$ with respect to $\bar{\pi}(\cdot|s) \in \mathcal{P}(\mathcal{A})$ is rather intuitive. As the set $\mathcal{P}(\mathcal{A})$ is not a subset of \mathbb{R}^n , the gradient is not necessarily defined—however, we do not require its existence. Instead, as $\mathcal{P}(\mathcal{A})$ is a convex statistical manifold, we consider its Gâteaux derivatives (Gateaux, 1922; Hamilton & Nashed, 1982). That is, for any $p \in \mathcal{P}(\mathcal{A})$, writing $v = p - \pi(\cdot|s)$, we consider the Gâteaux derivative, given as the limit

$$\delta \mathfrak{D}_\pi(\bar{\pi}|s)[v, \pi(\cdot|s)] \triangleq \lim_{\epsilon \rightarrow 0} \frac{\mathfrak{D}_\pi(\pi + \epsilon \cdot p|s) - \mathfrak{D}_\pi(\pi|s)}{\epsilon} \quad (10)$$

and require them to be zero at $\bar{\pi}(\cdot|s) = \pi(\cdot|s)$. That is, we require that $\delta \mathfrak{D}_\pi(\bar{\pi}|s)[v, \pi(\cdot|s)] = 0$.

Definition 3.2. We say that $\mathcal{N} : \Pi \rightarrow \mathbb{P}(\Pi)$ is a *neighbourhood operator*, where $\mathbb{P}(\Pi)$ is the power set of Π , if

1. It is a continuous map (*continuity*),
2. Every $\mathcal{N}(\pi)$ is a compact set (*compactness*),
3. There exists a metric $\chi : \Pi \times \Pi \rightarrow \mathbb{R}$, such that $\forall \pi \in \Pi$, there exists $\zeta > 0$, such that $\chi(\pi, \bar{\pi}) \leq \zeta$ implies $\bar{\pi} \in \mathcal{N}(\pi)$ (*closed ball*).

The *trivial* neighbourhood operator is $\mathcal{N} \equiv \Pi$.

We specify that a metric $\chi : \Pi \times \Pi \rightarrow \mathbb{R}$ is a metric between elements of a product of statistical manifolds, i.e., $\Pi = \times_{s \in \mathcal{S}} \mathcal{P}(\mathcal{A})$. Therefore, we carefully require that it is a non-negative, monotonically non-decreasing function of individual statistical divergence metrics $\chi_s(\pi(\cdot|s), \bar{\pi}(\cdot|s))$, $\forall s \in \mathcal{S}$, and that $\chi(\pi, \bar{\pi}(\cdot|s)) = 0$ only if $\chi_s(\pi(\cdot|s), \bar{\pi}(\cdot|s)) = 0, \forall s \in \mathcal{S}$.

³More precisely, all its Gâteaux derivatives are zero.

B. Proofs of Lemmas

Lemma 3.3. *Let π_{old} and π_{new} be policies. Suppose that*

$$[\mathcal{M}_{\mathfrak{D}}^{\pi_{new}} V_{\pi_{old}}](s) \geq [\mathcal{M}_{\mathfrak{D}}^{\pi_{old}} V_{\pi_{old}}](s), \quad \forall s \in \mathcal{S}. \quad (4)$$

Then, π_{new} is better than π_{old} , so that for every state s ,

$$V_{\pi_{new}}(s) \geq V_{\pi_{old}}(s).$$

Proof. Let us, for brevity, write $V_{new} = V_{\pi_{new}}$, $V_{old} = V_{\pi_{old}}$, $\beta = \beta_{\pi_{old}}$, $\nu = \nu_{\pi_{old}}^{\pi_{new}}$, and $\mathbb{E}_{\pi}[\cdot] = \mathbb{E}_{\mathbf{a} \sim \pi, s' \sim P}[\cdot]$. For every state $s \in \mathcal{S}$, we have

$$\begin{aligned} & V_{new}(s) - V_{old}(s) \\ &= \mathbb{E}_{\pi_{new}} [r(s, \mathbf{a}) + \gamma V_{new}(s')] - \mathbb{E}_{\pi_{old}} [r(s, \mathbf{a}) + \gamma V_{old}(s')] \\ &= \mathbb{E}_{\pi_{new}} [r(s, \mathbf{a}) + \gamma V_{old}(s')] - \frac{\nu(s)}{\beta(s)} \mathfrak{D}_{\pi_{old}}(\pi_{new}|s) \\ &\quad - \mathbb{E}_{\pi_{old}} [r(s, \mathbf{a}) + \gamma V_{old}(s')] \\ &\quad + \gamma \mathbb{E}_{\pi_{new}} [V_{new}(s') - V_{old}(s')] + \frac{\nu(s)}{\beta(s)} \mathfrak{D}_{\pi_{old}}(\pi_{new}|s) \\ &= \mathbb{E}_{\pi_{new}} [Q_{\pi_{old}}(s, \mathbf{a})] - \frac{\nu(s)}{\beta(s)} \mathfrak{D}_{\pi_{old}}(\pi_{new}|s) \\ &\quad - \mathbb{E}_{\pi_{old}} [Q_{\pi_{old}}(s, \mathbf{a})] \\ &\quad + \gamma \mathbb{E}_{\pi_{new}} [V_{new}(s') - V_{old}(s')] + \frac{\nu(s)}{\beta(s)} \mathfrak{D}_{\pi_{old}}(\pi_{new}|s) \\ &= [\mathcal{M}_{\mathfrak{D}}^{\pi_{new}} V_{old}](s) - [\mathcal{M}_{\mathfrak{D}}^{\pi_{old}} V_{old}](s) \\ &\quad + \gamma \mathbb{E}_{\pi_{new}} [V_{new}(s') - V_{old}(s')] + \frac{\nu(s)}{\beta(s)} \mathfrak{D}_{\pi_{old}}(\pi_{new}|s) \\ &\geq \gamma \mathbb{E}_{\pi_{new}} [V_{new}(s') - V_{old}(s')] + \frac{\nu(s)}{\beta(s)} \mathfrak{D}_{\pi_{old}}(\pi_{new}|s), \end{aligned}$$

by Inequality (4). Taking infimum within the expectation,

$$\begin{aligned} & V_{new}(s) - V_{old}(s) \\ &\geq \gamma \cdot \inf_{s'} [V_{new}(s') - V_{old}(s')] + \frac{\nu(s)}{\beta(s)} \mathfrak{D}_{\pi_{old}}(\pi_{new}|s). \end{aligned}$$

Taking infimum over s

$$\begin{aligned} & \inf_s [V_{new}(s) - V_{old}(s)] \\ &\geq \gamma \inf_{s'} [V_{new}(s') - V_{old}(s')] + \inf_s \frac{\nu(s)}{\beta(s)} \mathfrak{D}_{\pi_{old}}(\pi_{new}|s). \end{aligned}$$

We conclude that

$$\inf_s [V_{new}(s) - V_{old}(s)] \geq \inf_s \frac{\nu(s)}{\beta(s)} \mathfrak{D}_{\pi_{old}}(\pi_{new}|s) / (1 - \gamma),$$

which is non-negative, and concludes the proof. \square

Lemma 3.5. *Let π_{old} be a policy and π_{new} be obtained from the mirror update of Equation (5). Then,*

$$[\mathcal{M}_{\mathfrak{D}}^{\pi_{new}} V_{\pi_{old}}](s) \geq [\mathcal{M}_{\mathfrak{D}}^{\pi_{old}} V_{\pi_{old}}](s), \quad \forall s \in \mathcal{S}.$$

Hence, π_{new} attains the properties provided by Lemma 3.3.

Proof. We will prove the statement by contradiction. Suppose that

$$\pi_{\text{new}} = \arg \max_{\bar{\pi} \in \mathcal{N}(\pi_{\text{old}})} \mathbb{E}_{s \sim \beta_{\pi_{\text{old}}}} \left[[\mathcal{M}_{\mathcal{D}}^{\bar{\pi}} V_{\pi_{\text{old}}}] (s) \right], \quad (11)$$

and that there exists a state s_0 , such that

$$[\mathcal{M}_{\mathcal{D}}^{\pi_{\text{new}}} V_{\pi_{\text{old}}}] (s_0) < [\mathcal{M}_{\mathcal{D}}^{\pi_{\text{old}}} V_{\pi_{\text{old}}}] (s_0). \quad (12)$$

Let us define a policy $\hat{\pi}$, so that $\hat{\pi}(\cdot|s_0) = \pi_{\text{old}}(\cdot|s_0)$, and $\hat{\pi}(\cdot|s) = \pi_{\text{new}}(\cdot|s)$ for $s \neq s_0$. As the distance between of $\hat{\pi}$ from π_{old} is the same as of π_{new} at $s \neq s_0$, and possibly smaller at s_0 , we have that $\hat{\pi}$ is not further from π_{old} than π_{new} . Hence, $\hat{\pi} \in \mathcal{N}(\pi_{\text{old}})$. Furthermore, Inequality (12) implies

$$[\mathcal{M}_{\mathcal{D}}^{\hat{\pi}} V_{\pi_{\text{old}}}] (s_0) > [\mathcal{M}_{\mathcal{D}}^{\pi_{\text{new}}} V_{\pi_{\text{old}}}] (s_0).$$

Hence,

$$\begin{aligned} & \mathbb{E}_{s \sim \beta_{\pi_{\text{old}}}} \left[[\mathcal{M}_{\mathcal{D}}^{\hat{\pi}} V_{\pi_{\text{old}}}] (s) \right] - \mathbb{E}_{s \sim \beta_{\pi_{\text{old}}}} \left[[\mathcal{M}_{\mathcal{D}}^{\pi_{\text{new}}} V_{\pi_{\text{old}}}] (s) \right] \\ &= \beta_{\text{old}}(s_0) \left([\mathcal{M}_{\mathcal{D}}^{\hat{\pi}} V_{\pi_{\text{old}}}] (s_0) - [\mathcal{M}_{\mathcal{D}}^{\pi_{\text{new}}} V_{\pi_{\text{old}}}] (s_0) \right) > 0. \end{aligned}$$

This is a contradiction with Equation (11), which finishes the proof. □

C. The Proof of Theorem 3.6

Theorem 3.6 (The Fundamental Theorem of Mirror Learning). *Let \mathcal{D}^ν be a drift, \mathcal{N} be a neighbourhood operator, and the sampling distribution β_π depend continuously on π . Let $\pi_0 \in \Pi$, and the sequence of policies $(\pi_n)_{n=0}^\infty$ be obtained by mirror learning induced by \mathcal{D}^ν , \mathcal{N} , and β_π . Then, the learned policies*

1. *Attain the strict monotonic improvement property,*

$$\eta(\pi_{n+1}) \geq \eta(\pi_n) + \mathbb{E}_{s \sim d} \left[\frac{\nu_{\pi_n}^{\pi_{n+1}}(s)}{\beta_{\pi_n}(s)} \mathcal{D}_{\pi_n}(\pi_{n+1}|s) \right],$$

2. *Their value functions converge to the optimal one,*

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

3. *Their expected returns converge to the optimal return,*

$$\lim_{n \rightarrow \infty} \eta(\pi_n) = \eta^*,$$

4. *Their ω -limit set consists of the optimal policies.*

Proof. We split the proof of the whole theorem into two parts, each of which proves different groups of properties stated in the theorem.

Strict monotonic improvement (Property 1)

By Lemma 3.3, we have that $\forall n \in \mathbb{N}, s \in \mathcal{S}$,

$$V_{\pi_{n+1}}(s) \geq V_{\pi_n}(s). \quad (13)$$

Let us write $\beta = \beta_{\pi_n}$ and $\nu = \nu_{\pi_n}^{\pi_{n+1}}$. With this inequality, as well as Lemma 3.5, we have

$$\begin{aligned} & V_{\pi_{n+1}}(s) - V_{\pi_n}(s) \\ &= \mathbb{E}_{\mathbf{a} \sim \pi_{n+1}, s' \sim P} [r(s, \mathbf{a}) + \gamma V_{\pi_{n+1}}(s')] \\ &\quad - \mathbb{E}_{\mathbf{a} \sim \pi_n, s' \sim P} [r(s, \mathbf{a}) + \gamma V_{\pi_n}(s')] \\ &= \frac{\nu(s)}{\beta(s)} \mathcal{D}_{\pi_n}(\pi_{n+1}|s) + \mathbb{E}_{\mathbf{a} \sim \pi_{n+1}, s' \sim P} [r(s, \mathbf{a}) + \gamma V_{\pi_{n+1}}(s')] \\ &\quad - \frac{\nu(s)}{\beta(s)} \mathcal{D}_{\pi_n}(\pi_{n+1}|s) - \mathbb{E}_{\mathbf{a} \sim \pi_n, s' \sim P} [r(s, \mathbf{a}) + \gamma V_{\pi_n}(s')]. \end{aligned}$$

Now, by Inequality (13),

$$\begin{aligned} & \geq \frac{\nu(s)}{\beta(s)} \mathcal{D}_{\pi_n}(\pi_{n+1}|s) + \mathbb{E}_{\mathbf{a} \sim \pi_{n+1}} [r(s, \mathbf{a}) + \gamma V_{\pi_n}(s')] \\ &\quad - \frac{\nu(s)}{\beta(s)} \mathcal{D}_{\pi_n}(\pi_{n+1}|s) - \mathbb{E}_{\mathbf{a} \sim \pi_n} [r(s, \mathbf{a}) + \gamma V_{\pi_n}(s')] \\ &= \frac{\nu(s)}{\beta(s)} \mathcal{D}_{\pi_n}(\pi_{n+1}|s) + \mathbb{E}_{\mathbf{a} \sim \pi_{n+1}} [Q_{\pi_n}(s, \mathbf{a})] \\ &\quad - \frac{\nu(s)}{\beta(s)} \mathcal{D}_{\pi_n}(\pi_{n+1}|s) - \mathbb{E}_{\mathbf{a} \sim \pi_n} [Q_{\pi_n}(s, \mathbf{a})] \\ &= \frac{\nu(s)}{\beta(s)} \mathcal{D}_{\pi_n}(\pi_{n+1}|s) + [\mathcal{M}_{\mathcal{D}}^{\pi_{n+1}} V_{\pi_n}](s) - [\mathcal{M}_{\mathcal{D}}^{\pi_n} V_{\pi_n}](s) \\ &\geq \frac{\nu(s)}{\beta(s)} \mathcal{D}_{\pi_n}(\pi_{n+1}|s), \end{aligned}$$

where the last inequality holds by Lemma 3.5. To summarise, we proved that

$$V_{\pi_{n+1}}(s) - V_{\pi_n}(s) \geq \frac{\nu(s)}{\beta(s)} \mathfrak{D}_{\pi_n}(\pi_{n+1}|s). \quad (14)$$

Taking the expectation over $s \sim d$, we obtain

$$\eta(\pi_{n+1}) - \eta(\pi_n) \geq \mathbb{E}_{s \sim d} \left[\frac{\nu(s)}{\beta(s)} \mathfrak{D}_{\pi_n}(\pi_{n+1}|s) \right],$$

which finishes this part of the proof.

Optimality (Properties 2, 3, & 4)

Step 1 (convergence of value functions).

We consider the sequence $(V_{\pi_n})_{n=0}^{\infty}$ of associated value functions. By Lemma 3.5, we know that $\forall n \in \mathbb{N}$, $V_{\pi_{n+1}} \geq V_{\pi_n}$ uniformly. However, as for every $n \in \mathbb{N}$ and $s \in \mathcal{S}$, the uniform bound $V_{\pi_n}(s) \leq V_{\max}$ holds, the sequence of value functions must converge. We denote its limit by V .

Step 2 (characterisation of policy limit points).

Let $L\Pi$ be the set of all limit points of $(\pi_n)_{n=0}^{\infty}$. As the sequence $(\pi_n)_{n=0}^{\infty}$ is bounded, Bolzano-Weierstrass guarantees that there exists $\bar{\pi} \in L\Pi$, and a subsequence, say $(\pi_{n_i})_{i=1}^{\infty}$, which converges to it. Writing $\beta_k = \beta_{\pi_k}$, we consider the optimisation problem that mirror learning induces at every element π_{n_i} of this subsequence,

$$\max_{\pi \in \mathcal{N}(\pi_{n_i})} \mathbb{E}_{s \sim \beta_{n_i}} \left[[\mathcal{M}_{\mathfrak{D}}^{\pi} V_{\pi_{n_i}}](s) \right] \quad (15)$$

Note that by the continuity of the value function (Kuba et al., 2021, Appendix A), as well as the drift, the neighbourhood operator, and the sampling distribution, we obtain that the above expression is continuous in π_{n_i} . Hence, by Berge's Maximum Theorem (Ausubel & Deneckere, 1993), writing $\bar{\beta} = \beta_{\bar{\pi}}$, as $i \rightarrow \infty$, the above expression converges to the following,

$$\max_{\pi \in \mathcal{N}(\bar{\pi})} \mathbb{E}_{s \sim \bar{\beta}} \left[[\mathcal{M}_{\mathfrak{D}}^{\pi} V_{\bar{\pi}}](s) \right]. \quad (16)$$

Note that $V_{\bar{\pi}} = V$, as the sequence of value functions converges (has a unique limit point V). Furthermore, for all $i \in \mathbb{N}$, π_{n_i+1} is the argmax (precisely, is an element of the upper hemicontinuous argmax correspondence) of Equation (15). Hence, there exists a subsequence $(\pi_{n_{i_k}+1})$ of $(\pi_{n_i+1})_{i \in \mathbb{N}}$ that converges to a policy π' , which is a solution to Equation (16).

Claim: The solution to Equation (16) is $\pi' = \bar{\pi}$.

We prove the above claim by contradiction. Suppose that $\pi' \neq \bar{\pi}$. As π' can be obtained from $\bar{\pi}$ via mirror learning, Inequality (13) implies

$$\begin{aligned} Q_{\pi'}(s, a) &= \mathbb{E}_{s' \sim P} [r(s, a) + \gamma V_{\pi'}(s')] \\ &\geq \mathbb{E}_{s' \sim P} [r(s, a) + \gamma V_{\bar{\pi}}(s')] = Q_{\bar{\pi}}(s, a). \end{aligned} \quad (17)$$

If we have

$$\mathbb{E}_{s \sim \bar{\beta}} \left[[\mathcal{M}_{\mathfrak{D}}^{\pi'} V_{\bar{\pi}}](s) \right] > \mathbb{E}_{s \sim \bar{\beta}} \left[[\mathcal{M}_{\mathfrak{D}}^{\bar{\pi}} V_{\bar{\pi}}](s) \right],$$

then for some state s ,

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

which can be written as

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

Using Inequality (17), and non-negativity of drifts,

$$\begin{aligned} V_{\pi'}(s) &= \mathbb{E}_{\mathbf{a} \sim \pi'} [Q_{\pi'}(s, \mathbf{a})] \\ &\geq \mathbb{E}_{\mathbf{a} \sim \pi'} [Q_{\bar{\pi}}(s, \mathbf{a})] - \frac{\nu(s)}{\bar{\beta}(s)} \mathfrak{D}_{\bar{\pi}}(\pi'|s) > V(s). \end{aligned}$$

However, as $V_{\pi'}$ is the limit of $V_{\pi_{n_{i_k}+1}}$, we have $V_{\pi'} = V$ (uniqueness of the value limit) which yields a contradiction, proving the claim. Hence, the limit point $\bar{\pi}$ of the sequence $(\pi_n)_{n=0}^{\infty}$ satisfies

$$\bar{\pi} = \arg \max_{\pi \in \mathcal{N}(\bar{\pi})} \mathbb{E}_{\mathbf{s} \sim \bar{\beta}} \left[\left[\mathcal{M}_{\bar{\mathfrak{D}}}^{\pi} V_{\bar{\pi}} \right](s) \right]. \quad (18)$$

Step 3 (dropping the drift).

Let $\bar{\pi}$ be a limit point of $(\pi_n)_{n=0}^{\infty}$. From Equation (18), and the definition of the mirror operator, we know that

$$\bar{\pi} = \arg \max_{\pi \in \mathcal{N}(\bar{\pi})} \mathbb{E}_{\mathbf{s} \sim \bar{\beta}, \mathbf{a} \sim \pi} [A_{\bar{\pi}}(s, \mathbf{a})] - \mathfrak{D}_{\bar{\pi}}^{\nu}(\pi). \quad (19)$$

Suppose that there exists a policy π' , and a state s , such that

$$\mathbb{E}_{\mathbf{a} \sim \pi'} [A_{\bar{\pi}}(s, \mathbf{a})] > \mathbb{E}_{\mathbf{a} \sim \bar{\pi}} [A_{\bar{\pi}}(s, \mathbf{a})] = 0. \quad (20)$$

For any policy π , consider the canonical parametrisation $\pi(\cdot|s) = (x_1, \dots, x_{m-1}, 1 - \sum_{i=1}^{m-1} x_i)$, where m is the size of the action space. We have that

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

This means that $\mathbb{E}_{\mathbf{a} \sim \pi} [A_{\bar{\pi}}(s, \mathbf{a})]$ is an affine function of $\pi(\cdot|s)$, and thus, its Gâteaux derivatives are constant in $\mathcal{P}(\mathcal{A})$ for fixed directions. Together with Inequality (20), this implies that the Gâteaux derivative, in the direction from $\bar{\pi}$ to π' , is strictly positive. Furthermore, the Gâteaux derivatives of $\frac{\nu_{\bar{\pi}}^{\pi}(s)}{\bar{\beta}(s)} \mathfrak{D}_{\bar{\pi}}(\pi|s)$ are zero at $\pi(\cdot|s) = \bar{\pi}(\cdot|s)$, as

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

and

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

and both the lower and upper of the above bounds have zero derivatives. Hence, the Gâteaux derivative of $\mathbb{E}_{\mathbf{a} \sim \pi} [A_{\bar{\pi}}(s, \mathbf{a})] - \frac{\nu_{\bar{\pi}}^{\pi}(s)}{\bar{\beta}(s)} \mathfrak{D}_{\bar{\pi}}(\pi|s)$ is strictly positive. Therefore, for conditional policies $\hat{\pi}(\cdot|s)$ sufficiently close to $\bar{\pi}(\cdot|s)$ in the direction towards $\pi'(\cdot|s)$, we have (with a slight abuse of notation for ν)

$$\mathbb{E}_{\mathbf{a} \sim \hat{\pi}} [A_{\bar{\pi}}(s, \mathbf{a})] - \frac{\nu_{\bar{\pi}}^{\hat{\pi}}(s)}{\bar{\beta}(s)} \mathfrak{D}_{\bar{\pi}}(\hat{\pi}|s) > 0. \quad (21)$$

Let us construct a policy $\tilde{\pi}$ as follows. For all states $y \neq s$, we set $\tilde{\pi}(\cdot|y) = \bar{\pi}(\cdot|y)$. Moreover, for $\tilde{\pi}(\cdot|s)$ we choose $\hat{\pi}(\cdot|s)$ as in Inequality (21), sufficiently close to $\bar{\pi}(\cdot|s)$, so that $\tilde{\pi} \in \mathcal{N}(\bar{\pi})$. Then, we have

$$\begin{aligned} &\mathbb{E}_{\mathbf{s} \sim \bar{\beta}, \mathbf{a} \sim \tilde{\pi}} [A_{\bar{\pi}}(s, \mathbf{a})] - \mathfrak{D}_{\bar{\pi}}^{\nu}(\tilde{\pi}) \\ &= \bar{\beta}(s) [\mathbb{E}_{\mathbf{a} \sim \hat{\pi}} [A_{\bar{\pi}}(s, \mathbf{a})] - \frac{\nu_{\bar{\pi}}^{\hat{\pi}}(s)}{\bar{\beta}(s)} \mathfrak{D}_{\bar{\pi}}(\hat{\pi}|s)] > 0. \end{aligned} \quad (22)$$

The above contradicts Equation (19). Hence, the assumption made in Inequality (20) was false. Thus, we have proved that, for every state s ,

$$\begin{aligned}\bar{\pi} &= \arg \max_{\pi} \mathbb{E}_{a \sim \pi} [A_{\bar{\pi}}(s, a)] \\ &= \arg \max_{\pi} \mathbb{E}_{a \sim \pi} [Q_{\bar{\pi}}(s, a)].\end{aligned}\tag{23}$$

Step 4 (optimality).

Equation (23) implies that $\bar{\pi}$ is an optimal policy (Sutton & Barto, 2018), and so the value function $V = V_{\bar{\pi}}$ is the optimal value function V^* (Property 2). Consequently, the expected return that the policies converge to, $\eta = \mathbb{E}_{s \sim d}[V(s)] = \mathbb{E}_{s \sim d}[V^*(s)] = \eta^*$ is optimal (Property 3). Lastly, as $\bar{\pi}$ was an arbitrary limit point, any element of the ω -limit set is an optimal policy (Property 4). This finishes the proof. \square

D. Extension to Continuous State and Action Spaces

The results on Mirror Learning extend to continuous state and action spaces through general versions of our claims. These require a little more care in their formulation, but their validity holds as a corollary to our proofs.

In general, the state and the action spaces \mathcal{S} and \mathcal{A} must be assumed to be compact and measurable. For the state space, we introduce a reference probability measure $\mu_{\mathcal{S}} : \mathbb{P}(\mathcal{S}) \rightarrow \mathbb{R}$ that is strictly positive, *i.e.*, $\mu_{\mathcal{S}}(s) > 0, \forall s \in \mathcal{S}$. Under such setting, a policy π^* is optimal if it satisfies the Bellman optimality equation

$$\pi^*(\cdot|s) = \arg \max_{\pi(\cdot|s) \in \mathcal{P}(\mathcal{A})} \mathbb{E}_{a \sim p} [Q_{\pi^*}(s, a)],$$

at states that form a set of measure 1 with respect to $\mu_{\mathcal{S}}$. In other words, a policy is optimal if it obeys the Bellman optimality equation almost surely.

As for the results, the inequality provided by Lemma 3.3 (the state value function improvement) holds almost surely with respect to $\mu_{\mathcal{S}}$ as long as the policy π_{new} satisfies Inequality (4), also almost surely with respect to this measure. Of course, the corollary on the monotonic improvement remains valid. Next, the inequality introduced in Lemma 3.5 must now be stated almost surely, again with respect to $\mu_{\mathcal{S}}$. Lastly, the entire statement of Theorem 3.6 remains unchanged—it has the same formulation for all compact .

E. The Listing of RL Approaches as Instances of Mirror Learning

Generalised Policy Iteration (Sutton & Barto, 2018)

$$\pi_{\text{new}}(\cdot|s) = \arg \max_{\bar{\pi}(\cdot|s) \in \mathcal{P}(\mathcal{A})} \mathbb{E}_{\mathbf{a} \sim \bar{\pi}} [Q_{\pi_{\text{old}}}(s, \mathbf{a})], \quad \forall s \in \mathcal{S}$$

- Drift functional: trivial $\mathfrak{D} \equiv 0$.
- Neighbourhood operator: trivial $\mathcal{N} \equiv \Pi$.
- Sampling distribution: arbitrary.

Trust-Region Learning (Schulman et al., 2015)

$$\pi_{\text{new}} = \arg \max_{\bar{\pi} \in \Pi} \mathbb{E}_{s \sim \rho_{\pi_{\text{old}}}, \mathbf{a} \sim \bar{\pi}} [A_{\pi_{\text{old}}}(s, \mathbf{a})] - CD_{\text{KL}}^{\max}(\pi_{\text{old}}, \bar{\pi}),$$

$$\text{where } C = \frac{4\gamma \max_{s, a} |A_{\pi_{\text{old}}}(s, a)|}{(1 - \gamma)^2}$$

- Drift functional:

$$\mathfrak{D}_{\pi}(\bar{\pi}|s) = (1 - \gamma)CD_{\text{KL}}(\pi(\cdot|s), \bar{\pi}(\cdot|s)), \quad \text{with } \nu_{\pi}^{\bar{\pi}}(s_{\text{max}}) = 1, \quad \text{where } s_{\text{max}} = \arg \max_{s \in \mathcal{S}} \mathfrak{D}_{\pi}(\bar{\pi}|s).$$

- Neighbourhood operator: trivial $\mathcal{N} \equiv \Pi$.
- Sampling distribution: $\beta_{\pi} = \bar{\rho}_{\pi}$, the normalised marginal discounted state distribution.

TRPO (Schulman et al., 2015)

$$\pi_{\text{new}} = \arg \max_{\bar{\pi} \in \Pi} \mathbb{E}_{s \sim \rho_{\pi_{\text{old}}}, \mathbf{a} \sim \bar{\pi}} [A_{\pi_{\text{old}}}(s, \mathbf{a})]$$

$$\text{subject to } \mathbb{E}_{s \sim \rho_{\pi_{\text{old}}}} [D_{\text{KL}}(\pi_{\text{old}}(\cdot|s), \pi_{\text{new}}(\cdot|s))] \leq \delta.$$

- Drift functional: trivial $\mathfrak{D} \equiv 0$.
- Neighbourhood operator:

$$\mathcal{N}(\pi) = \{\bar{\pi} \in \Pi : \mathbb{E}_{s \sim \rho_{\pi}} [D_{\text{KL}}(\pi(\cdot|s), \bar{\pi}(\cdot|s))] \leq \delta\}.$$

- Sampling distribution: $\beta_{\pi} = \bar{\rho}_{\pi}$.

PPO (Schulman et al., 2017)

$$\pi_{\text{new}} = \arg \max_{\bar{\pi} \in \Pi} \mathbb{E}_{s \sim \bar{\rho}_{\pi_{\text{old}}}, \mathbf{a} \sim \pi_{\text{old}}} [L^{\text{clip}}],$$

$$\text{where for given } s, a, \text{ and } r(\bar{\pi}) = \frac{\bar{\pi}(a|s)}{\pi_{\text{old}}(a|s)}$$

$$L^{\text{clip}} = \min \left(r(\bar{\pi})A_{\pi_{\text{old}}}(s, a), \text{clip}(r(\bar{\pi}), 1 \pm \epsilon)A_{\pi_{\text{old}}}(s, a) \right).$$

- Drift functional:

$$\mathfrak{D}_{\pi}(\bar{\pi}|s) = \mathbb{E}_{\mathbf{a} \sim \pi} \left[\text{ReLU} \left([r(\bar{\pi}) - \text{clip}(r(\bar{\pi}), 1 \pm \epsilon)] A_{\pi}(s, \mathbf{a}) \right) \right], \quad \text{with } \nu_{\pi}^{\bar{\pi}} = \bar{\rho}_{\pi}.$$

- Neighbourhood operator: trivial $\mathcal{N} \equiv \Pi$. Note, however, that in practical implementations the policy is updated with ϵ_{PPO} steps of gradient ascent with gradient-clipping threshold M . This corresponds to a neighbourhood of an L2-ball of radius $M\epsilon_{\text{PPO}}$ in the policy parameter space.
- Sampling distribution: $\beta_{\pi} = \bar{\rho}_{\pi}$.

PPO-KL (Hsu et al., 2020)

$$\pi_{\text{new}} = \arg \max_{\bar{\pi} \in \Pi} \mathbb{E}_{s \sim \bar{\rho}_{\pi_{\text{old}}}, a \sim \bar{\pi}} [A_{\pi_{\text{old}}}(s, a)] - \tau \bar{D}_{\text{KL}}(\pi_{\text{old}}, \bar{\pi}),$$

- Drift functional:

$$\mathfrak{D}_{\pi}(\bar{\pi}|s) = \tau D_{\text{KL}}(\pi(\cdot|s), \bar{\pi}(\cdot|s)), \quad \text{with } \nu_{\bar{\pi}}^{\pi} = \bar{\rho}_{\pi}.$$

- Neighbourhood operator: the same as in PPO.
- Sampling distribution: $\beta_{\pi} = \bar{\rho}_{\pi}$.

MDPO (Tomar et al., 2020)

$$\pi_{\text{new}} = \arg \max_{\bar{\pi} \in \Pi} \mathbb{E}_{s \sim \beta_{\pi_{\text{old}}}, a \sim \bar{\pi}} [A_{\pi_{\text{old}}}(s, a)] - \frac{1}{t_{\pi_{\text{old}}}} \bar{D}_{\text{KL}}(\bar{\pi}, \pi_{\text{old}}).$$

- Drift functional:

$$\mathfrak{D}_{\pi}(\bar{\pi}|s) = \frac{1}{t_{\pi}} D_{\text{KL}}(\bar{\pi}(\cdot|s), \pi(\cdot|s)), \quad \text{with } \nu_{\bar{\pi}}^{\pi} = \beta_{\pi}.$$

- Neighbourhood operator: trivial $\mathcal{N} = \Pi$.
- Sampling distribution: $\beta_{\pi} = \bar{\rho}_{\pi}$ for on-policy MDPO, and $\beta_{\pi_n}(s) = \frac{1}{n+1} \sum_{i=0}^n \bar{\rho}_{\pi_i}(s)$ for off-policy MDPO.

F. Instructions for Implementation of Off-Policy Mirror Learning

In the case of off-policy learning, estimating $\mathbb{E}_{\mathbf{a} \sim \bar{\pi}} [Q_{\pi_{\text{old}}}(s, \mathbf{a})]$ is not as straightforward as in Equation (6), since sampling actions from the replay buffer is not equivalent to sampling actions from $\pi_{\text{old}}(\cdot|s)$ anymore. The reason for this is that while sampling an action from the buffer, we also sample a past policy, π_{hist} , which was used to insert the action to the buffer. To formalise it, we draw a past policy from some distribution dictated by the buffer, $\pi_{\text{hist}} \sim h \in \mathcal{P}(\Pi)$, and then draw an action $\mathbf{a} \sim \pi_{\text{hist}}$. To account for this, we use the following estimator,

$$\frac{\bar{\pi}(\mathbf{a}|s)}{\pi_{\text{hist}}(\mathbf{a}|s)} Q_{\pi_{\text{old}}}(s, \mathbf{a}).$$

Note that this requires that the value $\pi_{\text{hist}}(\mathbf{a}|s)$ has also been inserted in the buffer. The expectation of the new estimator can be computed as

$$\begin{aligned} & \mathbb{E}_{\pi_{\text{hist}} \sim h, \mathbf{a} \sim \pi_{\text{hist}}} \left[\frac{\bar{\pi}(\mathbf{a}|s)}{\pi_{\text{hist}}(\mathbf{a}|s)} Q_{\pi_{\text{old}}}(s, \mathbf{a}) \right] \\ &= \sum_{\pi_{\text{hist}}} h(\pi_{\text{hist}}) \sum_{\mathbf{a} \in \mathcal{A}} \pi_{\text{hist}}(\mathbf{a}|s) \frac{\bar{\pi}(\mathbf{a}|s)}{\pi_{\text{hist}}(\mathbf{a}|s)} Q_{\pi_{\text{old}}}(s, \mathbf{a}) \\ &= \sum_{\mathbf{a} \in \mathcal{A}} \bar{\pi}(\mathbf{a}|s) Q_{\pi_{\text{old}}}(s, \mathbf{a}) \sum_{\pi_{\text{hist}}} h(\pi_{\text{hist}}) \\ &= \sum_{\mathbf{a} \in \mathcal{A}} \bar{\pi}(\mathbf{a}|s) Q_{\pi_{\text{old}}}(s, \mathbf{a}) = \mathbb{E}_{\mathbf{a} \sim \bar{\pi}} [Q_{\pi_{\text{old}}}(s, \mathbf{a})]. \end{aligned}$$

Hence, the estimator has the desired mean.