Meta-Gradient Reinforcement Learning

Zhongwen Xu
DeepMind
zhongwen@google.com

Hado van Hasselt
DeepMind
hado@google.com

David Silver
DeepMind
davidsilver@google.com

Abstract

The goal of reinforcement learning algorithms is to estimate and/or optimise the value function. However, unlike supervised learning, no teacher or oracle is available to provide the true value function. Instead, the majority of reinforcement learning algorithms estimate and/or optimise a proxy for the value function. This proxy is typically based on a sampled and bootstrapped approximation to the true value function, known as a *return*. The particular choice of return is one of the chief components determining the nature of the algorithm: the rate at which future rewards are discounted; when and how values should be bootstrapped; or even the nature of the rewards themselves. It is well-known that these decisions are crucial to the overall success of RL algorithms. We introduce a novel, gradient-based meta-learning algorithm that is able to adapt the nature of the return, online, whilst interacting and learning from the environment. When applied to 57 games on the Atari 2600 environment over 200 million frames, our algorithm achieved a new state-of-the-art.

The central goal of reinforcement learning (RL) is to optimise the agent's *return* (cumulative reward); this is typically achieved by a combination of prediction and control. The prediction subtask is to estimate the value function – the expected return from any given state. Ideally, this would be achieved by updating an approximate value function towards the true value function. The control subtask is to optimise the agent's policy for selecting actions, so as to maximise the value function. Ideally, the policy would simply be updated in the direction that increases the true value function. However, the true value function is unknown and therefore, for both prediction and control, a sampled return is instead used as a proxy. A large family of RL algorithms [Sutton, 1988, Rummery and Niranjan, 1994, van Seijen et al., 2009, Sutton and Barto, 2018], including several state-of-the-art deep RL algorithms [Mnih et al., 2015, van Hasselt et al., 2016, Harutyunyan et al., 2016, Hessel et al., 2017, Espeholt et al., 2018], are characterised by different choices of the return.

The *discount factor* γ determines the time-scale of the return. A discount factor close to $\gamma = 1$ provides a long-sighted goal that accumulates rewards far into the future, while a discount factor close to $\gamma = 0$ provides a short-sighted goal that prioritises short-term rewards. Even in problems where long-sightedness is clearly desired, it is frequently observed that discounts $\gamma < 1$ achieve better results [Prokhorov and Wunsch, 1997], especially during early learning. It is known that many algorithms converge faster with lower discounts [Bertsekas and Tsitsiklis, 1996], but of course too low a discount can lead to highly sub-optimal policies that are too myopic. In practice it can be better to first optimise for a myopic horizon, e.g., with $\gamma = 0$ at first, and then to repeatedly increase the discount only after learning is somewhat successful [Prokhorov and Wunsch, 1997].

The return may also be *bootstrapped* at different time horizons. An n*-step return* accumulates rewards over n time-steps and then adds the value function at the nth time-step. The λ-return [Sutton, 1988, Sutton and Barto, 2018] is a geometrically weighted combination of n-step returns. In either case, the meta-parameter n or λ can be important to the performance of the algorithm, trading off bias and variance. Many researchers have sought to automate the selection of these parameters [Kearns and Singh, 2000, Downey and Sanner, 2010, Konidaris et al., 2011, White and White, 2016].

Preprint. Work in progress.

There are potentially many other design choices that may be represented in the return, including off-policy corrections [Espeholt et al., 2018, Munos et al., 2016], target networks [Mnih et al., 2015], emphasis on certain states [Sutton et al., 2016], reward clipping [Mnih et al., 2013], or even the nature of the rewards themselves [Randløv and Alstrøm, 1998, Singh et al., 2005, Zheng et al., 2018].

In this work, we are interested in one of the fundamental problems in reinforcement learning: what would be the best form of return for the agent to maximise? Specifically, we propose to *learn* the return function by treating it as a parametric function with tunable meta-parameters η, for instance including the discount factor γ, or the bootstrapping parameter λ [Sutton, 1988]. The meta-parameters η are adjusted *online* during the agent's interaction with the environment, allowing the return to both adapt to the specific problem, and also to dynamically adapt over time to the changing context of learning. We derive a practical gradient-based meta-learning algorithm and show that this can significantly improve performance on large-scale deep reinforcement learning applications.

1 Meta-Gradient Reinforcement Learning Algorithms

In deep reinforcement learning, the value function and policy are approximated by a neural network with parameters θ, denoted by $v_\theta(S)$ and $\pi_\theta(A|S)$ respectively. At the core of the algorithm is an *update function*,

$$\theta' = \theta + f(\tau, \theta, \eta)\,, \tag{1}$$

that adjusts parameters from a sequence of experience $\tau_t = \{S_t, A_t, R_{t+1}, \ldots\}$ consisting of states S, actions A and rewards R. The nature of the function is determined by *meta-parameters* η.

Our meta-gradient RL approach is based on the principle of online cross-validation [Sutton, 1992], using successive samples of experience. The underlying RL algorithm is applied to the first sample (or samples), and its performance is measured in a subsequent sample. Specifically, the algorithm starts with parameters θ, and applies the update function to the first sample(s), resulting in new parameters θ'; the gradient $\mathrm{d}\theta'/\mathrm{d}\eta$ of these updates indicates how the meta-parameters affected these new parameters. The algorithm then measures the performance of the new parameters θ' on a subsequent, independent sample τ', utilising a differentiable *meta-objective* $J'(\tau', \theta', \eta')$. When validating the performance on the second sample, we use a fixed meta-parameter η' in J' as a reference value. In this way, we form a differentiable function of the meta-parameters, and obtain the gradient of η by taking the derivative of meta-objective J' w.r.t. η and applying the chain rule:

$$\frac{\partial J'(\tau', \theta', \eta')}{\partial \eta} = \frac{\partial J'(\tau', \theta', \eta')}{\partial \theta'} \frac{\mathrm{d}\theta'}{\mathrm{d}\eta}\,. \tag{2}$$

To compute the gradient of the updates, $\mathrm{d}\theta'/\mathrm{d}\eta$, we note that the parameters form an additive sequence, and the gradient can therefore be accumulated online,

$$\frac{\mathrm{d}\theta'}{\mathrm{d}\eta} = \frac{\mathrm{d}\theta}{\mathrm{d}\eta} + \frac{\partial f(\tau, \theta, \eta)}{\partial \eta} + \frac{\partial f(\tau, \theta, \eta)}{\partial \theta}\frac{\mathrm{d}\theta}{\mathrm{d}\eta} = \left(I + \frac{\partial f(\tau, \theta, \eta)}{\partial \theta}\right)\frac{\mathrm{d}\theta}{\mathrm{d}\eta} + \frac{\partial f(\tau, \theta, \eta)}{\partial \eta} \tag{3}$$

The gradient $\partial f(\tau, \theta, \eta)/\partial\theta$ is large and challenging to compute in practice. Instead we approximate the gradient using an accumulative trace $z \approx \mathrm{d}\theta/\mathrm{d}\eta$,

$$z' = \mu z + \frac{\partial f(\tau, \theta, \eta)}{\partial \eta} \tag{4}$$

The gradient in Equation (3) is defined for fixed meta-parameters η. In practice, the meta-parameters will adapt online. To allow for this adaptation, the parameter $\mu \in [0, 1]$ decays the trace and focuses on recent updates. Choosing $\mu = 0$ results in a greedy meta-gradient that considers only the immediate effect of the meta-parameters η on the parameters θ; this may often be sufficient.

Finally, the meta-parameters η are updated to optimise the meta-objective, for example by applying stochastic gradient descent (SGD) to update η in the direction of the meta-gradient,

$$\Delta\eta = -\beta \frac{\partial J'(\tau', \theta', \eta')}{\partial \theta'} z', \tag{5}$$

where β is the learning rate for updating meta parameter η.

In the following sections we instantiate this idea more specifically to RL algorithms based on predicting or controlling returns. We begin with a pedagogical example of using meta-gradients for

prediction using a temporal-difference update. We then consider meta-gradients for control, using a canonical actor-critic update function and a policy gradient meta-objective. Many other instantiations of meta-gradient RL would be possible, since the majority of deep reinforcement learning updates are differentiable functions of the return, including, for instance, value-based methods like $Sarsa(\lambda)$ [Sutton and Barto, 2018] and DQN [Mnih et al., 2015], policy-gradient methods [Williams, 1992], or actor-critic algorithms like A3C [Mnih et al., 2016] and IMPALA [Espeholt et al., 2018].

1.1 Applying Meta-Gradients to Returns

We define the return $g_\eta(\tau_t)$ to be a function of an episode or a truncated n-step sequence of experience $\tau_t = \{S_t, A_t, R_{t+1}, \ldots, S_{t+n}\}$. The nature of the return is determined by the meta-parameters η.

The n-step return [Sutton and Barto, 2018] accumulates rewards over the sequence and then bootstraps from the value function,

$$g_\eta(\tau_t) = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots, + \gamma^{n-1} R_{t+n} + \gamma^n v_\theta(S_{t+n}) \tag{6}$$

where $\eta = \{\gamma, n\}$.

The λ-return is a geometric mixture of n-step returns, [Sutton, 1988]

$$g_\eta(\tau_t) = R_{t+1} + \gamma(1-\lambda) v_\theta(S_{t+1}) + \gamma\lambda g_\eta(\tau_{t+1}) \tag{7}$$

where $\eta = \{\gamma, \lambda\}$. The λ-return has the advantage of being fully differentiable with respect to the meta-parameters. The meta-parameters η may be viewed as gates that cause the return to terminate ($\gamma = 0$) or bootstrap ($\lambda = 0$), or to continue onto the next step ($\gamma = 1$ or $\lambda = 1$). The n-step or λ-return can be augmented with off-policy corrections [Precup et al., 2000, Sutton et al., 2014, Espeholt et al., 2018] if it is necessary to correct for the distribution used to generate the data.

A typical RL algorithm would hand-select the meta-parameters, such as the discount factor γ and bootstrapping parameter λ, and these would be held fixed throughout training. Instead, we view the return g as a function parameterised by meta-parameters η, which may be differentiated to understand its dependence on η. This in turn allows us to compute the gradient $\partial f / \partial \eta$ of the update function with respect to the meta-parameters η, and hence the meta-gradient $\partial J'(\tau', \theta', \eta') / \partial \eta$. In essence, our agent asks itself the question, "which return results in the best performance?", and adjusts its meta-parameters accordingly.

1.2 Meta-Gradient Prediction

We begin with a simple instantiation of the idea, based on the canonical TD(λ) algorithm for prediction. The objective of the TD(λ) algorithm (according to the forward view [Sutton and Barto, 2018]) is to minimise the squared error between the value function approximator $v_\theta(S)$ and the λ-return $g_\eta(\tau)$,

$$J(\tau, \theta, \eta) = (g_\eta(\tau) - v_\theta(S))^2 \qquad \frac{\partial J(\tau, \theta, \eta)}{\partial \theta} = -2(g_\eta(\tau) - v_\theta(S)) \frac{\partial v_\theta(S)}{\partial \theta} \tag{8}$$

where τ is a sampled trajectory starting with state S, and $\partial J(\tau, \theta, \eta) / \partial \theta$ is a semi-gradient [Sutton and Barto, 2018], i.e. the λ-return is treated as a constant.

The TD(λ) update function $f(\cdot)$ applies SGD to update the agent's parameters θ to descend the gradient of the objective with respect to the parameters,

$$f(\tau, \theta, \eta) = -\frac{\alpha}{2} \frac{\partial J(\tau, \theta, \eta)}{\partial \theta} = \alpha(g_\eta(\tau) - v_\theta(S)) \frac{\partial v_\theta(S)}{\partial \theta} \tag{9}$$

where α is the learning rate for updating agent θ. We note that this update is itself a differentiable function of the meta-parameters η,

$$\frac{\partial f(\tau, \theta, \eta)}{\partial \eta} = -\frac{\alpha}{2} \frac{\partial^2 J(\tau, \theta, \eta)}{\partial \theta \, \partial \eta} = \alpha \frac{\partial g_\eta(\tau)}{\partial \eta} \frac{\partial v_\theta(S)}{\partial \theta} \tag{10}$$

The key idea of the meta-gradient prediction algorithm is to adjust meta-parameters η in the direction that achieves the best predictive accuracy. This is measured by cross-validating the new parameters θ' on a second trajectory τ' that starts from state S', using a mean squared error (MSE) meta-objective and taking its semi-gradient,

$$J'(\tau', \theta', \eta') = (g_{\eta'}(\tau') - v_{\theta'}(S'))^2 \qquad \frac{\partial J'(\tau', \theta', \eta')}{\partial \theta'} = -2(g_{\eta'}(\tau') - v_{\theta'}(S')) \frac{\partial v_{\theta'}(S')}{\partial \theta'} \tag{11}$$

The meta-objective in this case would typically make use of an unbiased and long-sighted return[1], for example using $\eta' = \{\gamma', \lambda'\}$ where $\gamma' = 1$ and $\lambda' = 1$.

1.3 Meta-Gradient Control

We now provide a practical example of meta-gradients applied to control. We focus on the A2C algorithm – an actor-critic update function that combines both prediction and control into a single update. This update function is widely used in several state-of-the-art agents [Mnih et al., 2016, Jaderberg et al., 2017b, Espeholt et al., 2018]. The semi-gradient of the A2C objective, $\partial J(\tau;\theta,\eta)/\partial\theta$, is defined as follows,

$$-\frac{\partial J(\tau,\theta,\eta)}{\partial\theta} = (g_\eta(\tau) - v_\theta(S))\frac{\partial \log \pi_\theta(A|S)}{\partial\theta} + c(g_\eta(\tau) - v_\theta(S))\frac{\partial v_\theta(S)}{\partial\theta} + d\frac{\partial H(\pi_\theta(\cdot|S))}{\partial\theta}. \quad (12)$$

The first term represents a control objective, encouraging the policy π_θ to select actions that maximise the return. The second term represents a prediction objective, encouraging the value function approximator v_θ to more accurately estimate the return $g_\eta(\tau)$. The third term regularises the policy according to its entropy $H(\pi_\theta)$, and c, d are coefficients that weight the different components in the objective function.

The A2C update function $f(\cdot)$ applies SGD to update the agent's parameters θ. This update function is a differentiable function of the meta-parameters η,

$$f(\tau,\theta,\eta) = -\alpha\frac{\partial J(\tau,\theta,\eta)}{\partial\theta} \qquad \frac{\partial f(\tau,\theta,\eta)}{\partial\eta} = \alpha\frac{\partial g_\eta(\tau)}{\partial\eta}\left[\frac{\partial \log \pi_\theta(A|S)}{\partial\theta} + c\frac{\partial v_\theta(S)}{\partial\theta}\right] \quad (13)$$

Now we come to the choice of meta-objective J' to use for control. Our goal is to identify the return function that maximises overall performance in our agents. This may be directly measured by a meta-objective focused exclusively on optimising returns – in other words a policy gradient objective,

$$\frac{\partial J'(\tau',\theta',\eta')}{\partial\theta'} = (g_{\eta'}(\tau') - v_{\theta'}(S'))\frac{\partial \log \pi_{\theta'}(A'|S')}{\partial\theta'}. \quad (14)$$

This equation evaluates how good the updated policy θ' is in terms of returns computed under η', when measured on "held-out" experiences τ', e.g. the subsequent n-step trajectory. When cross-validating performance using this meta-objective, we use fixed meta-parameters η', ideally representing a good proxy to the true objective of the agent. In practice this typically means selecting reasonable values of η'; the agent is free to adapt its meta-parameters η and choose values that perform better in practice.

We now put the meta-gradient control algorithm together. First, the parameters θ are updated on a sample of experience τ using the A2C update function (Equation (12)), and the gradient of the update (Equation (13)) is accumulated into trace z. Second, the performance is cross-validated on a subsequent sample of experience τ' using the policy gradient meta-objective (Equation (14)). Finally, the meta-parameters η are updated according to the gradient of the meta-objective (Equation (5)).

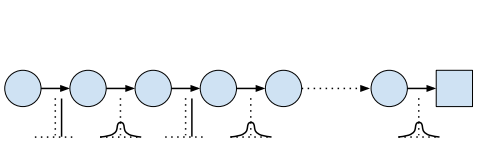
1.4 Conditioned Value and Policy Functions

One complication of the approach outlined above is that the return function $g_\eta(\tau)$ is non-stationary, adapting along with the meta-parameters throughout the training process. As a result, there is a danger that the value function v_θ becomes inaccurate, since it may be approximating old returns. For example, the value function may initially form a good approximation of a short-sighted return with $\gamma = 0$, but if γ subsequently adapts to $\gamma = 1$ then the value function may suddenly find its approximation is rather poor. The same principle applies for the policy π, which again may have specialised to old returns.

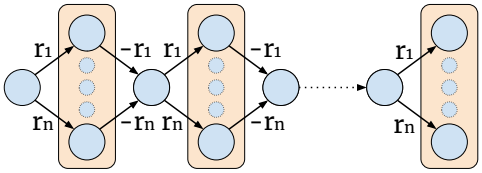
To deal with non-stationarity in the value function and policy, we utilise an idea similar to universal value function approximation (UVFA) [Schaul et al., 2015]. The key idea is to provide the meta-parameters η as an additional input to condition the value function and policy, as follows:

$$v_\theta^\eta(S) = v_\theta([S;\mathbf{e}_\eta]), \qquad \pi_\theta^\eta(S) = \pi_\theta([S;\mathbf{e}_\eta]), \qquad \mathbf{e}_\eta = \mathbf{W}_\eta\eta,$$

[1] The meta-objective could even use a discount factor that is longer-sighted than the original problem, perhaps spanning over many episodes.



(a) Rewards alternate between $+0.1$ or randomly sampling from a zero-mean Gaussian.



(b) Bottleneck states transition randomly to one of n states; these states (shaded rectangles) are aliased.

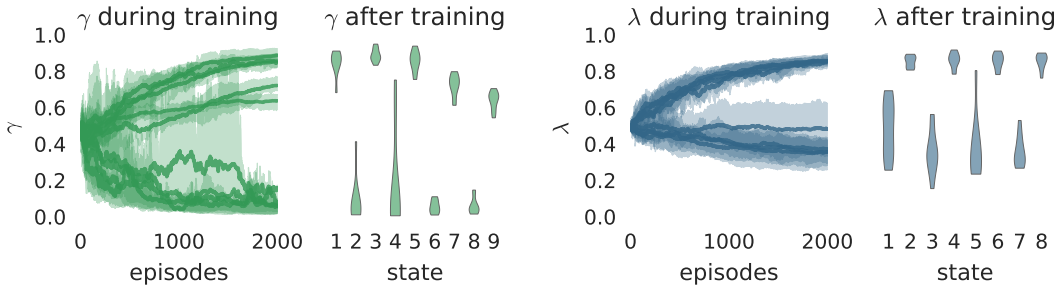


Figure 1: Illustrative results of meta-gradient learning of a state-dependent (a) bootstrapping parameter λ or (b) discount factor γ, in the respective Markov reward processes (top). In each of the subplot shown in the bottom, the first one shows how the meta-parameter γ or λ adapts over the course of training (averaged over 10 seeds - shaded regions cover 20%–80% percentiles). The second plot shows the final value of γ or λ in each state, identifying appropriately high/low values for odd/even states respectively (violin plots show distribution over seeds).

where $\mathbf{e}_\eta$ is the embedding of η, $[s; \mathbf{e}_\eta]$ denotes concatenation of vectors s and $\mathbf{e}_\eta$, $\mathbf{W}_\eta$ is the embedding matrix (or row vector, for scalar η) that is updated by backpropagation during training.

In this way, the agent explicitly learns value functions and policies that are appropriate for various η. The approximation problem becomes a little harder, but the payoff is that the algorithm can freely shift the meta-parameters without needing to wait for the approximator to "catch up".

1.5 Meta-Gradient Reinforcement Learning in Practice

To scale up the meta-gradient approach, several additional steps were taken. For efficiency, the A2C objective and meta-objective were accumulated over all time-steps within an n-step trajectory of experience. The A2C objective was optimised by RMSProp without momentum [Mnih et al., 2015, 2016, Espeholt et al., 2018]. This is a differentiable function of the meta-parameters, and can therefore be substituted similarly to SGD (see Equation (13)); this process may be simplified by automatic differentiation (Appendix B.2). As in IMPALA, an off-policy correction was used, based on a V-trace return (see Appendix B.1). For efficient implementation, mini-batches of trajectories were computed in parallel; trajectories were reused twice for both the update function and for cross-validation (see Appendix B.3).

2 Illustrative Examples

To illustrate the key idea of our meta-gradient approach, we provide two examples that show how the discount factor γ and temporal difference parameter λ, respectively, can be meta-learned. We focus on meta-gradient prediction using the TD(λ) algorithm and a MSE meta-objective with $\gamma' = 1$ and $\lambda' = 1$, as described in Section 1.2. For these illustrative examples, we consider *state-dependent* meta-parameters that can take on a different value in each state.

The first example is a 10-step Markov reward process (MRP), that alternates between "signal" and "noise" transitions. Transitions from odd-numbered "signal" states receive a small positive reward, $R = +0.1$. Transitions from even-numbered "noise" states receive a random reward, $R \sim \mathcal{N}(0, 1)$. To ensure that the signal can overwhelm the noise, it is beneficial to terminate the return (low γ) in "noise" states, but to continue the return (high γ) in "signal" states.

The second example is a 9-step MRP, that alternates between "fan-out" and "fan-in" transitions. At odd-numbered "fan-out" time-steps, the state transitions to a randomly sampled successor, all of which are aliased. At even-numbered "fan-in" time-steps, the state deterministically transitions back to a bottleneck state. The transition to a "fan-out" state yields a deterministic reward unique to that state. The transition away from that state yields the negation of that same reward. Each pair of "fan-in" and "fan-out" time-steps is zero-mean in expected reward. The fully observed bottleneck states are visited frequently and tend to have more accurate value functions. To predict accurately, it is therefore beneficial to bootstrap (low λ) in bottleneck states for which the value function is well-known, but to avoid bootstrapping (high λ) in the noisier, partially observed, fanned-out states.

Figure 1 shows the results of meta-gradient prediction using the TD(λ) algorithm. The meta-gradient algorithm was able to adapt both λ and γ to form returns that alternate between high or low values in odd or even states respectively.

3 Deep Reinforcement Learning Experiments

In this section, we demonstrate the advantages of the proposed meta-gradient learning approach using a state-of-the-art actor-critic framework IMPALA [Espeholt et al., 2018]. We focused on adapting the discount factor $\eta = \{\gamma\}$ (which we found to be the most effective meta-parameter in preliminary experiments), and the bootstrapping parameter λ. For these experiments, the meta-parameters were state-independent, adapting one scalar value for γ and λ respectively (state-dependent meta-parameters did not provide significant benefit in preliminary experiments).[2]

3.1 Experiment Setup

We validate the proposed approach on Atari 2600 video games from Arcade Learning Environment (ALE) [Bellemare et al., 2013], a standard benchmark for deep reinforcement learning algorithms. We build our agent with the IMPALA framework [Espeholt et al., 2018], an efficient distributed implementation of actor-critic architecture [Sutton and Barto, 2018, Mnih et al., 2016]. We utilise the deep ResNet architecture [He et al., 2016] specified in Espeholt et al. [2018], which has shown great advantages over the shallow architecture [Mnih et al., 2015]. Following Espeholt et al., we train our agent for 200 million frames. Our algorithm does not require extra data compared to the baseline algorithms, as each experience can be utilised in both training the agent itself and training the meta parameters η (i.e., each experience can serve as validation data of other experiences). We describe the detailed implementation in the Appendix B.3. For full details about the IMPALA implementation and the specific off-policy correction $g_\eta(\tau)$, please refer to [Espeholt et al., 2018].

The agents are evaluated on 57 different Atari games and the median of human-normalised scores [Nair et al., 2015, van Hasselt et al., 2016, Wang et al., 2016b, Mnih et al., 2016] are reported. There are two different evaluation protocols. The first protocol is is "human starts" [Nair et al., 2015, Wang et al., 2016b, van Hasselt et al., 2016], which initialises episodes to a state that is randomly sampled from human play. The second protocol is "no-ops starts", which initialises each episode with a random sequence of no-op actions; this protocol is also used during training.

We keep all of the hyper-parameters (e.g., batch size, unroll length, learning rate, entropy cost) the same as specified in [Espeholt et al., 2018] for a fair comparison. For self-contained purpose, we provide all of the important hyper-parameters used in this paper, including the ones following [Espeholt et al., 2018] and the additional meta-learning optimisation hyper-parameters (i.e., meta batch size, meta learning rate α', embedding size for η), in Appendix A. The meta-learning hyper-parameters are chosen according to the performance of six Atari games as common practice in Deep RL Atari experiments [van Hasselt et al., 2016, Mnih et al., 2016, Wang et al., 2016b]. Additional implementation details are provided in Appendix B.

3.2 Experiment Results

We compared four variants of the IMPALA algorithm: the original baseline algorithm without meta-gradients, i.e. $\eta = \{\}$; using meta-gradients with $\eta = \{\lambda\}$; using meta-gradients with $\eta = \{\gamma\}$; and

[2] In practice we parameterise $\eta = \sigma(x)$, where σ is the logistic function $\sigma(x) = \frac{1}{1+e^{-x}}$; i.e. the meta-parameters are actually the logits of γ and λ.

	η	Human starts		No-op starts	
		$\gamma = 0.99$	$\gamma = 0.995$	$\gamma = 0.99$	$\gamma = 0.995$
IMPALA	$\{\}$	144.4%	211.9%	191.8%	257.1%
Meta-gradient	$\{\lambda\}$	156.6%	214.2%	185.5%	246.5%
		$\gamma' = 0.99$	$\gamma' = 0.995$	$\gamma' = 0.99$	$\gamma' = 0.995$
Meta-gradient	$\{\gamma\}$	233.2%	267.9%	280.9%	275.5%
Meta-gradient	$\{\gamma, \lambda\}$	221.6%	292.9%	242.6%	287.6%

Table 1: Results of meta-learning the discount parameter γ, the temporal-difference learning parameter λ, or both γ and λ, compared to the baseline IMPALA algorithm which meta-learns neither. Results are given both for the discount factor $\gamma = 0.99$ originally reported in [Espeholt et al., 2018] and also for a tuned discount factor $\gamma = 0.995$ (see Appendix C); the cross-validated discount factor γ' in the meta-objective was set to the same value for a fair comparison.

using meta-gradients with $\eta = \{\gamma, \lambda\}$. The original IMPALA algorithm used a discount factor of $\gamma = 0.99$; however, when we manually tuned the discount factor and found that a discount factor of $\gamma = 0.995$ performed considerably better (see Appendix C). For a fair comparison, we tested our meta-gradient algorithm in both cases. When the discount factor is not adapted, $\eta = \{\}$ or $\eta = \{\lambda\}$, we used a fixed value of $\gamma = 0.99$ or $\gamma = 0.995$. When the discount factor is adapted, $\eta = \{\gamma\}$ or $\eta = \{\gamma, \lambda\}$, we cross-validate with a meta-parameter of $\gamma' = 0.99$ or $\gamma' = 0.995$ accordingly in the meta-objective J' (Equation (14)). Manual tuning of the λ parameter did not have a significant impact on performance and we therefore compared only to the original value of $\lambda = 1$.

We summarise the median human-normalised scores in Table 1; individual improvements on each game, compared to the IMPALA baseline, are given in Appendix D.1; and individual plots demonstrating the adaptation of γ and λ are provided in Appendix D.2. The meta-gradient RL algorithm increased the median performance, compared to the baseline algorithm, by a margin between 30% and 80% across "human starts" and "no-op starts" conditions, and with both $\gamma = 0.99$ and $\gamma = 0.995$.

We also verified the architecture choice of conditioning the value function v and policy π on the meta-parameters η. We compared the proposed algorithm with an identical meta-gradient algorithm that adapts the discount factor $\eta = \{\gamma\}$, but does not provide an embedding of the discount factor as an input to π and v. For this experiment, we used a cross-validation discount factor of $\gamma' = 0.995$. The human-normalised median score was only 183%, well below the IMPALA baseline with $\gamma = 0.995$ (211.9%), and much worse than the full meta-gradient algorithm that includes the discount factor embedding (267.9%).

Finally, we compare against the state-of-the-art agent trained on Atari games, namely Rainbow [Hessel et al., 2017], which combines DQN [Mnih et al., 2015] with double Q-learning [van Hasselt et al., 2016, van Hasselt, 2010], prioritised replay [Schaul et al., 2016], dueling networks [Wang et al., 2016b], multi-step targets [Sutton, 1988, Sutton and Barto, 2018], distributional RL [Bellemare et al., 2017], and parameter noise for exploration [Fortunato et al., 2018]. Rainbow obtains median human-normalised score of 153% on the human starts protocol and 223% on the no-ops protocol. In contrast, the meta-gradient agent achieved a median score of 292.9% on human starts and 287.6% on no-ops, with the same number (200M) of frames. We note, however, that there are many differences between the two algorithms, including the deeper neural network architecture used in our work.

4 Related Work

Among the earliest studies on meta learning (or learning to learn [Thrun and Pratt, 1998]), Schmidhuber [1987] applied genetic programming to itself to evolve better genetic programming algorithms. Hochreiter et al. [2001] used recurrent neural networks like Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] as meta-learners. A recent direction of research has been to meta-learn an *optimiser* using a recurrent parameterisation [Andrychowicz et al., 2016, Wichrowska et al., 2017]. Duan et al. [2016] and Wang et al. [2016a] proposed to learn a recurrent meta-policy that itself learns to solve the reinforcement learning problem, so that the recurrent policy can generalise into new tasks faster than learning the policy from scratch. Model-Agnostic

Meta-Learning (MAML) [Finn et al., 2017a, Finn and Levine, 2018, Finn et al., 2017b, Grant et al., 2018, Al-Shedivat et al., 2018] learns a good initialisation of the model that can adapt quickly to other tasks within a few gradient update steps. These works focus on a multi-task setting in which meta-learning takes place on a distribution of training tasks, to facilitate fast adaptation on an unseen test task. In contrast, our work emphasises the (arguably) more fundamental problem of meta-learning within a single task. In other words we return to the standard formulation of RL as maximising rewards during a single lifetime of interactions with an environment.

Contemporaneously with our own work, Zheng et al. [2018] also propose a similar algorithm to learn meta-parameters of the return: in their case an auxiliary reward function that is added to the external rewards. They do not condition their value function or policy, and reuse the same samples for both the update function and the cross-validation step – which may be problematic in stochastic domains when the noise these updates becomes highly correlated.

There are many works focusing on adapting learning rate through gradient-based methods [Sutton, 1992, Schraudolph, 1999, Maclaurin et al., 2015, Pedregosa, 2016, Franceschi et al., 2017], Bayesian optimisation methods [Snoek et al., 2012], or evolution based hyper-parameter tuning [Jaderberg et al., 2017a, Elfwing et al., 2017]. In particular, [Sutton, 1992], introduced the idea of online cross-validation; however, this method was limited in scope to adapting the learning rate for linear updates in supervised learning (later extended to non-linear updates by Schraudolph [1999]); whereas we focus on the fundamental problem of reinforcement learning, i.e., adapting the return function to maximise the proxy returns we can achieve from the environment.

There has also been significant prior work on automatically adapting the bootstrapping parameter λ. Singh and Dayan [1998] empirically analyse the effect of λ in terms of bias, variance and MSE. Kearns and Singh [2000] derive upper bounds on the error of temporal-difference algorithms, and use these bounds to derive schedules for λ. Downey and Sanner [2010] introduced a Bayesian model averaging approach to scheduling λ. Konidaris et al. [2011] derive a maximum-likelihood estimator, TD(γ), that weights the n-step returns according to the discount factor, leading to a parameter-free algorithm for temporal-difference learning with linear function approximation. White and White [2016] introduce an algorithm that explicitly estimates the bias and variance, and greedily adapts λ to locally minimise the MSE of the λ-return. Unlike our meta-gradient approach, these prior approaches exploit i.i.d. assumptions on the trajectory of experience that are not realistic in many applications.

5 Conclusion

In this work, we discussed how to learn the meta-parameters of a return function. Our meta-learning algorithm runs online, while interacting with a single environment, and successfully adapts the return to produce better performance. We demonstrated, by adjusting the meta-parameters of a state-of-the-art deep learning algorithm, that we could achieve much higher performance than previously observed on 57 Atari 2600 games from the Arcade Learning Environment.

Our proposed method is more general, and can be applied not just to the discount factor or bootstrapping parameter, but also to other components of the return, and even more generally to the learning update itself. Hyper-parameter tuning has been a thorn in the side of reinforcement learning research for several decades. Our hope is that this approach will allow agents to automatically tune their own hyper-parameters, by exposing them as meta-parameters of the learning update. This may also result in better performance because the parameters can change over time and adapt to novel environments.

References

M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

M. Al-Shedivat, T. Bansal, Y. Burda, I. Sutskever, I. Mordatch, and P. Abbeel. Continuous adaptation via meta-learning in nonstationary and competitive environments. In *ICLR*, 2018.

M. Andrychowicz, M. Denil, S. Gomez, M. W. Hoffman, D. Pfau, T. Schaul, and N. de Freitas. Learning to learn by gradient descent by gradient descent. In *NIPS*, pages 3981–3989, 2016.

M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling. The arcade learning environment: An evaluation platform for general agents. *J. Artif. Intell. Res.(JAIR)*, 47:253–279, 2013.

M. G. Bellemare, W. Dabney, and R. Munos. A distributional perspective on reinforcement learning. In *ICML*, 2017.

D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

C. Downey and S. Sanner. Temporal difference bayesian model averaging: A bayesian perspective on adapting lambda. In *ICML*, pages 311–318. Citeseer, 2010.

Y. Duan, J. Schulman, X. Chen, P. L. Bartlett, I. Sutskever, and P. Abbeel. RL^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

S. Elfwing, E. Uchibe, and K. Doya. Online meta-learning by parallel algorithm competition. *CoRR*, abs/1702.07490, 2017.

L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, Y. Doron, V. Firoiu, T. Harley, I. Dunning, et al. IMPALA: Scalable distributed Deep-RL with importance weighted actor-learner architectures. *arXiv preprint arXiv:1802.01561*, 2018.

C. Finn and S. Levine. Meta-learning and universality: Deep representations and gradient descent can approximate any learning algorithm. *ICLR*, 2018.

C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017a.

C. Finn, T. Yu, T. Zhang, P. Abbeel, and S. Levine. One-shot visual imitation learning via meta-learning. In *CoRL*, 2017b.

M. Fortunato, M. G. Azar, B. Piot, J. Menick, I. Osband, A. Graves, V. Mnih, R. Munos, D. Hassabis, O. Pietquin, et al. Noisy networks for exploration. In *ICLR*, 2018.

L. Franceschi, M. Donini, P. Frasconi, and M. Pontil. Forward and reverse gradient-based hyperparameter optimization. In *ICML*, 2017.

E. Grant, C. Finn, S. Levine, T. Darrell, and T. Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. *ICLR*, 2018.

A. Harutyunyan, M. G. Bellemare, T. Stepleton, and R. Munos. Q(λ) with off-policy corrections. In *ALT*, pages 305–320. Springer, 2016.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver. Rainbow: Combining improvements in deep reinforcement learning. In *AAAI*, 2017.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

S. Hochreiter, A. S. Younger, and P. R. Conwell. Learning to learn using gradient descent. In *ICANN*, pages 87–94. Springer, 2001.

M. Jaderberg, V. Dalibard, S. Osindero, W. M. Czarnecki, J. Donahue, A. Razavi, O. Vinyals, T. Green, I. Dunning, K. Simonyan, et al. Population based training of neural networks. *arXiv preprint arXiv:1711.09846*, 2017a.

M. Jaderberg, V. Mnih, W. M. Czarnecki, T. Schaul, J. Z. Leibo, D. Silver, and K. Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. In *ICLR*, 2017b.

M. J. Kearns and S. P. Singh. Bias-variance error bounds for temporal difference updates. In *COLT*, pages 142–147, 2000.

G. Konidaris, S. Niekum, and P. S. Thomas. TD_γ: Re-evaluating complex backups in temporal difference learning. In *NIPS*, pages 2402–2410, 2011.

D. Maclaurin, D. Duvenaud, and R. Adams. Gradient-based hyperparameter optimization through reversible learning. In *ICML*, pages 2113–2122, 2015.

A. Mahmood. *Incremental Off-policy Reinforcement Learning Algorithms*. PhD thesis, University of Alberta, 2017.

V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning. *NIPS workshop*, 2013.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *ICML*, pages 1928–1937, 2016.

R. Munos, T. Stepleton, A. Harutyunyan, and M. Bellemare. Safe and efficient off-policy reinforcement learning. In *NIPS*, pages 1054–1062, 2016.

A. Nair, P. Srinivasan, S. Blackwell, C. Alcicek, R. Fearon, A. De Maria, V. Panneershelvam, M. Suleyman, C. Beattie, S. Petersen, et al. Massively parallel methods for deep reinforcement learning. *arXiv preprint arXiv:1507.04296*, 2015.

F. Pedregosa. Hyperparameter optimization with approximate gradient. In *ICML*, pages 737–746, 2016.

D. Precup, R. S. Sutton, and S. P. Singh. Eligibility traces for off-policy policy evaluation. In *ICML*, pages 759–766, 2000.

D. V. Prokhorov and D. C. Wunsch. Adaptive critic designs. *TNN*, 8(5):997–1007, 1997.

J. Randløv and P. Alstrøm. Learning to drive a bicycle using reinforcement learning and shaping. In *ICML*, volume 98, pages 463–471, 1998.

G. A. Rummery and M. Niranjan. On-line Q-learning using connectionist sytems. Technical Report CUED/F-INFENG-TR 166, Cambridge University, UK, 1994.

T. Schaul, D. Horgan, K. Gregor, and D. Silver. Universal value function approximators. In *ICML*, pages 1312–1320, 2015.

T. Schaul, J. Quan, I. Antonoglou, and D. Silver. Prioritized experience replay. In *ICLR*, 2016.

J. Schmidhuber. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. PhD thesis, Technische Universität München, 1987.

N. N. Schraudolph. Local gain adaptation in stochastic gradient descent. In *ICANN*. IET, 1999.

S. Singh and P. Dayan. Analytical mean squared error curves for temporal difference learning. *Machine Learning*, 32(1):5–40, 1998.

S. P. Singh, A. G. Barto, and N. Chentanez. Intrinsically motivated reinforcement learning. In *Advances in neural information processing systems*, pages 1281–1288, 2005.

J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In *NIPS*, pages 2951–2959, 2012.

R. S. Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1): 9–44, 1988.

R. S. Sutton. Adapting bias by gradient descent: An incremental version of delta-bar-delta. In *AAAI*, pages 171–176, 1992.

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press Cambridge, 2018.

R. S. Sutton, A. R. Mahmood, D. Precup, and H. van Hasselt. A new Q(λ) with interim forward view and Monte Carlo equivalence. In *ICML*, pages 568–576, 2014.

R. S. Sutton, A. R. Mahmood, and M. White. An emphatic approach to the problem of off-policy temporal-difference learning. *JMLR*, 17(1):2603–2631, 2016.

S. Thrun and L. Pratt. *Learning to learn*. Springer Science & Business Media, 1998.

H. van Hasselt. Double Q-learning. In *NIPS*, pages 2613–2621, 2010.

H. van Hasselt, A. Guez, and D. Silver. Deep reinforcement learning with double Q-learning. In *AAAI*, volume 16, pages 2094–2100, 2016.

H. H. van Seijen, H. P. van Hasselt, S. Whiteson, and M. A. Wiering. A theoretical and empirical analysis of Expected Sarsa. In *Proceedings of the IEEE International Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, pages 177–184, 2009.

J. X. Wang, Z. Kurth-Nelson, D. Tirumala, H. Soyer, J. Z. Leibo, R. Munos, C. Blundell, D. Kumaran, and M. Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016a.

Z. Wang, T. Schaul, M. Hessel, H. Van Hasselt, M. Lanctot, and N. De Freitas. Dueling network architectures for deep reinforcement learning. *ICML*, 2016b.

M. White and A. White. A greedy approach to adapting the trace parameter for temporal difference learning. In *AAMAS*, pages 557–565, 2016.

O. Wichrowska, N. Maheswaranathan, M. W. Hoffman, S. G. Colmenarejo, M. Denil, N. de Freitas, and J. Sohl-Dickstein. Learned optimizers that scale and generalize. In *ICML*, 2017.

R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256, May 1992.

Z. Zheng, J. Oh, and S. Singh. On learning intrinsic rewards for policy gradient methods. *arXiv preprint arXiv:1804.06459*, 2018.

A Detailed Hyper-Parameters used in the Atari Experiments

In Table 2, we describe the details of the important hyper-parameters used in the Atari experiments. The IMPALA hyper-parameter section is following [Espeholt et al., 2018], which is provided here for self-contained purpose. The hyper-parameters in meta-gradient section are obtained by a search on six games (Beamrider, Breakout, Pong, Q*bert, Seaquest and Space Invaders) following common practice in Deep RL Atari experiments [van Hasselt et al., 2016, Mnih et al., 2016, Wang et al., 2016b]. All of the hyper-parameters are fixed across all Atari games.

IMPALA hyper-parameter	**Value**
Network architecture	Deep ResNet
Unroll length	20
Batch size	32
Baseline loss scaling (c)	0.5
Entropy cost (d)	0.01
Learning rate (α)	0.0006
RMSProp momentum	0.0
RMSProp decay	0.99
RMSProp ϵ	0.1
Clip global gradient norm	40.0
Learning rate schedule	Anneal linearly to 0 from beginning to end of training.
Number of learners	1 (NVIDIA P100)
Number of actors	80
Meta-gradient hyper-parameter	**Value**
Trace decay (μ)	0
Meta learning rate (β)	0.001
Meta optimiser	ADAM
Meta batch size	8
Meta update frequency	along with every agent update
Embedding size for η	16

Table 2: Detailed hyper-parameters for Atari experiments.

B Implementation Details

B.1 V-trace Return

The λ-return [Sutton, 1988] is defined as

$$G_t^\lambda = R_{t+1} + \gamma_{t+1}(1-\lambda_{t+1})v(S_{t+1}) + \gamma_{t+1}\lambda_{t+1}G_{t+1}^\lambda\,.$$

This can be rewritten [Sutton et al., 2014] as

$$\begin{aligned} G_t^\lambda &= v(S_t) + \delta_t + \gamma_{t+1}\lambda_{t+1}\delta_{t+1} + \dots \\ &= v(S_t) + \sum_{k=1}^{\infty}\left(\prod_{j=1}^{k-1}\gamma_{t+j}\lambda_{t+j}\right)\delta_{t+k}\,. \end{aligned}$$

This return is on-policy. For some algorithms, especially policy-gradient methods, it is important that we have an estimate for the current policy. But in the IMPALA architecture, the data may be slightly stale before the learning algorithms consumes it. Then, off-policy corrections can be applied to make the data on-policy again. In particular, IMPALA uses a *v-trace* return, defined by

$$G_t^\lambda = v(S_t) + \sum_{k=1}^{\infty}\left(\prod_{j=1}^{k-1}\gamma_{t+j}c_{t+j}\right)\delta_{t+k}\,,$$

where $c_t = \min\left(1, \frac{\pi(A_t|S_t)}{\pi'(A_t|S_t)}\right)$, where π is the current policy and π' is the (older) policy that was used to generate the data. Note that this return can be interpreted as an adaptive-λ return, with a fixed adaptation scheme that depends only on the off-policy nature of the trajectory. A similar scheme was proposed by Mahmood [2017].

B.2 Calculate the Meta-Gradient with Auto-Diff

An important fact to note in the proposed approach is that, the update rule for $\theta \rightarrow \theta'$ in first-order optimiser like Equation (9) is linear and differentiable. In modern machine learning frameworks like TensorFlow [Abadi et al., 2016], we can alternatively obtain the meta-gradient specified in Equation (2), by utilising the automatic differentiation functionality in the framework. The only requirement is to rewrite the update operations so that the agent update can allow the gradient to flow through, since the build-in update operations are typically not differentiable in the common implementations.

B.3 Data Efficiency

In order to reduce the data we needed for meta learning, we can reuse the experiences for both agent training and meta learning. For example, we can use experiences τ for updating θ into θ', validate the performance of this update via evaluating J' on experiences τ'. Vice versa, we can swap the roles of τ and τ', then use experiences τ' for updating θ, and validate the performance of this update via evaluating J' on experiences τ. In this way, the proposed algorithm does not require extra data other than the ones used to train the agent parameter θ to conduct the meta learning update to η.

B.4 Running Speed

As for running speed, with one learner on NVIDIA P100 GPU and 80 actors on 80 CPU cores, our method runs around 13k environment steps/second, compared to around 20k environment steps/second of IMPALA baseline on the same hardware and software environments. We introduce around 35% additional compute overhead from the meta-gradient updates, however with this minor overhead we can boost the performance significantly. We'd like to highlight that the total wall clock time for finishing 200 Million frames in each game is about 4 hours.

B.5 IMPALA

We used the IMPALA algorithm with the *deep* architecture and the *experts* mode of training. In this mode, a separate agent is trained on each environment (i.e. the standard RL setting), as opposed to a multi-task setting. Population-based training was not utilised by the IMPALA experts in [Espeholt et al., 2018]; we follow this convention. In principle, γ and λ could be exposed to population-based training (PBT) [Jaderberg et al., 2017a], however, this would blow up the computation time by the size of the population (24 in [Espeholt et al., 2018]), which is beyond reach for typical experiments; furthermore adaptation by PBT does not exploit the gradient and is therefore perhaps less likely scale to larger meta-parameterisations.

C Results of Grid Search of Discount Factor γ on Atari Experiments

We conduct simple grid search on the discount factor γ, i.e., let $\gamma = 0.99, 0.995, 0.998, 0.999$ respectively, and apply it in the IMPALA framework. The grid search of discount factor γ is to find some good γ' ($\eta' = \gamma'$ in this case) to be used in the meta-objective $J'(\tau', \theta_{t+1}, \eta')$, so that the meta learning approach can have a good proxy to the true return to learn from.

As we can see from Table 3, the discount factor γ has huge impact on the agent performance.

Table 3: Performance comparison of IMPALA baseline with different discount factor γ, all scores are human-normalised [Mnih et al., 2015, Wang et al., 2016b, van Hasselt et al., 2016].

	Human starts	No-ops starts
Agents	median	median
$\gamma = 0.99$ [Espeholt et al., 2018]	144.4%	191.8%
$\gamma = 0.995$	211.9%	257.1%
$\gamma = 0.998$	208.5%	210.7%
$\gamma = 0.999$	114.9%	153.0%

D Additional Experiment Results

D.1 Relative Performance Improvement in Individual Games

In this section, we provide the relative performance improvement of the meta-gradient algorithm compared to the IMPALA baselines in individual Atari 2600 games. We show the results of adapting discount factor, i.e., $\eta = \{\gamma\}$, in Figure 2, and results of adapting both discount factor and bootstrapping parameter, i.e., $\eta = \{\gamma, \lambda\}$, in Figure 3.

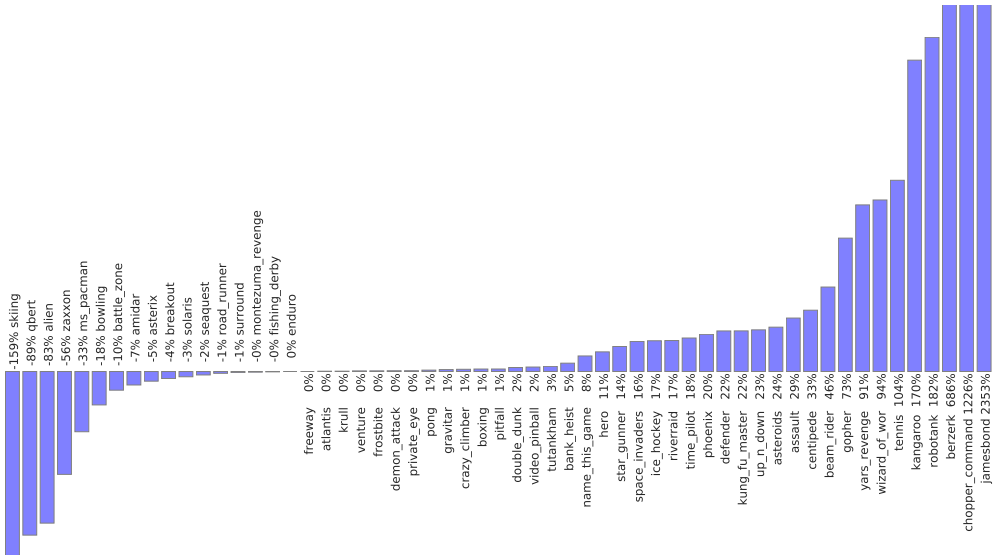


Figure 2: The relative performance improvement of the meta-gradient algorithm, adapting discount factor, i.e., $\eta = \{\gamma\}$, compared to the baseline IMPALA ($\gamma = 0.99$) in individual Atari 2600 games, where the gain is given by $\frac{\text{proposed} - \text{baseline}}{\max(\text{human}, \text{baseline}) - \text{random}}$ [Wang et al., 2016b]. Improvement over 200% is capped into 200% for visualisation.

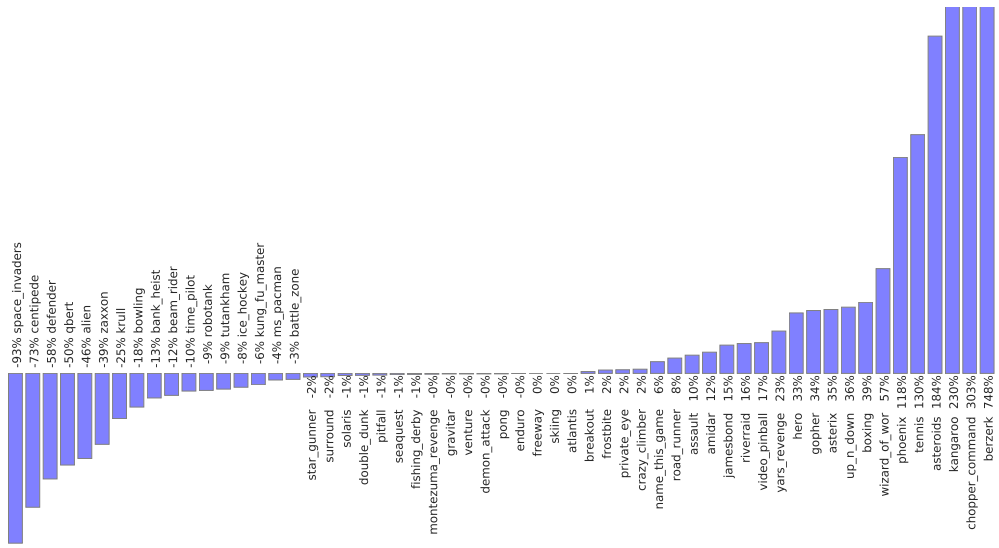


Figure 3: The relative performance improvement of the meta-gradient algorithm, adapting both discount factor and bootstrapping parameter, $\eta = \{\gamma, \lambda\}$, compared to the baseline IMPALA ($\gamma = 0.995$) in individual Atari 2600 games, where the gain is given by $\frac{\text{proposed}-\text{baseline}}{\max(\text{human},\text{baseline})-\text{random}}$ [Wang et al., 2016b]. Improvement over 200% is capped into 200% for visualisation.

D.2 Training Curves

In this section, we provide the training curves for two representative experiments: adapting $\eta = \{\gamma\}$ with $\gamma' = 0.99$ (Figure 4) and adapting $\eta = \{\gamma, \lambda\}$ with $\gamma' = 0.995$ (Figure 5).

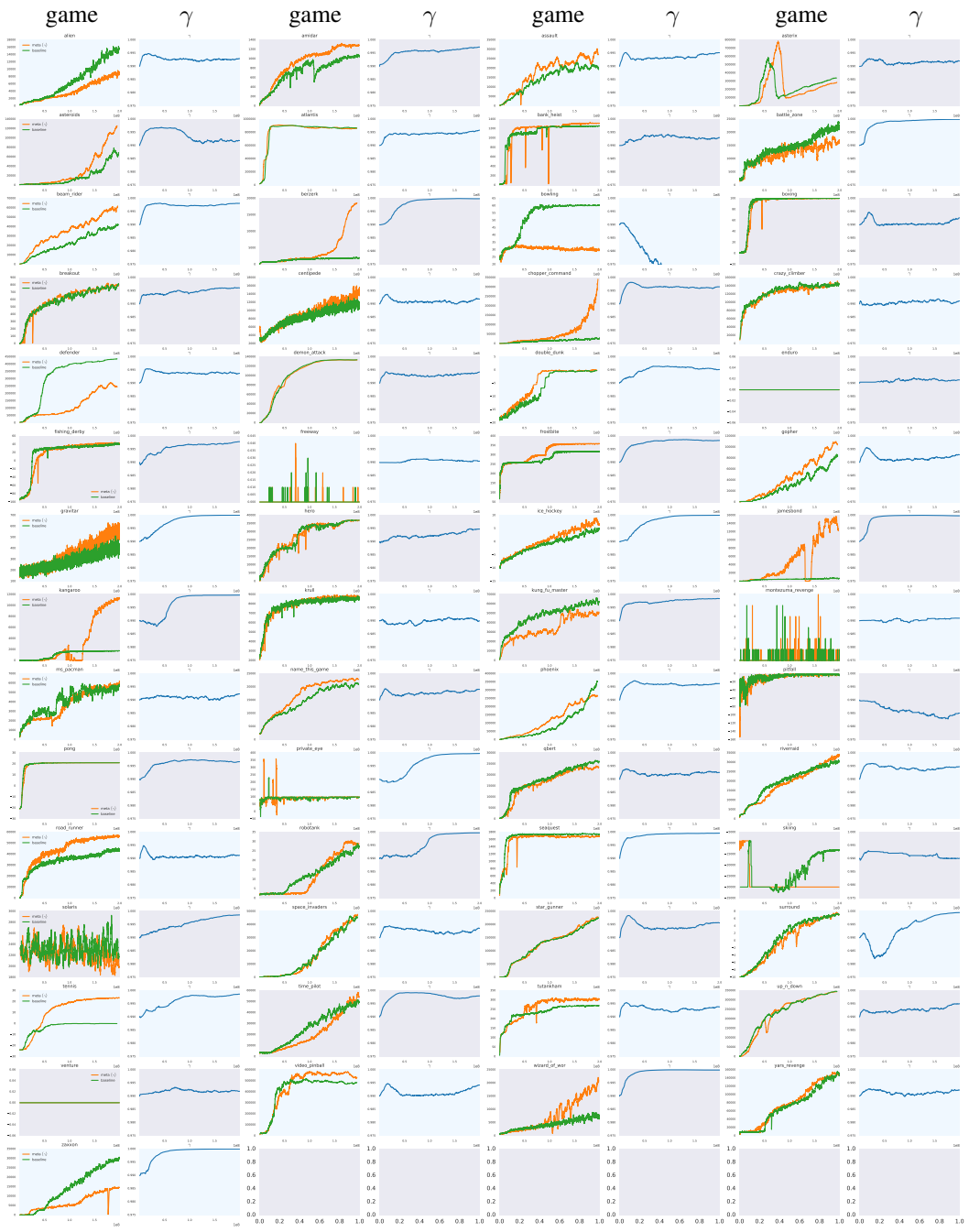


Figure 4: Training curves for meta learning $\eta = \{\gamma\}$ on $\gamma' = 0.99$. We provide the comparison of scores against baseline, and the change of γ for each game. Best viewed in electronic version.

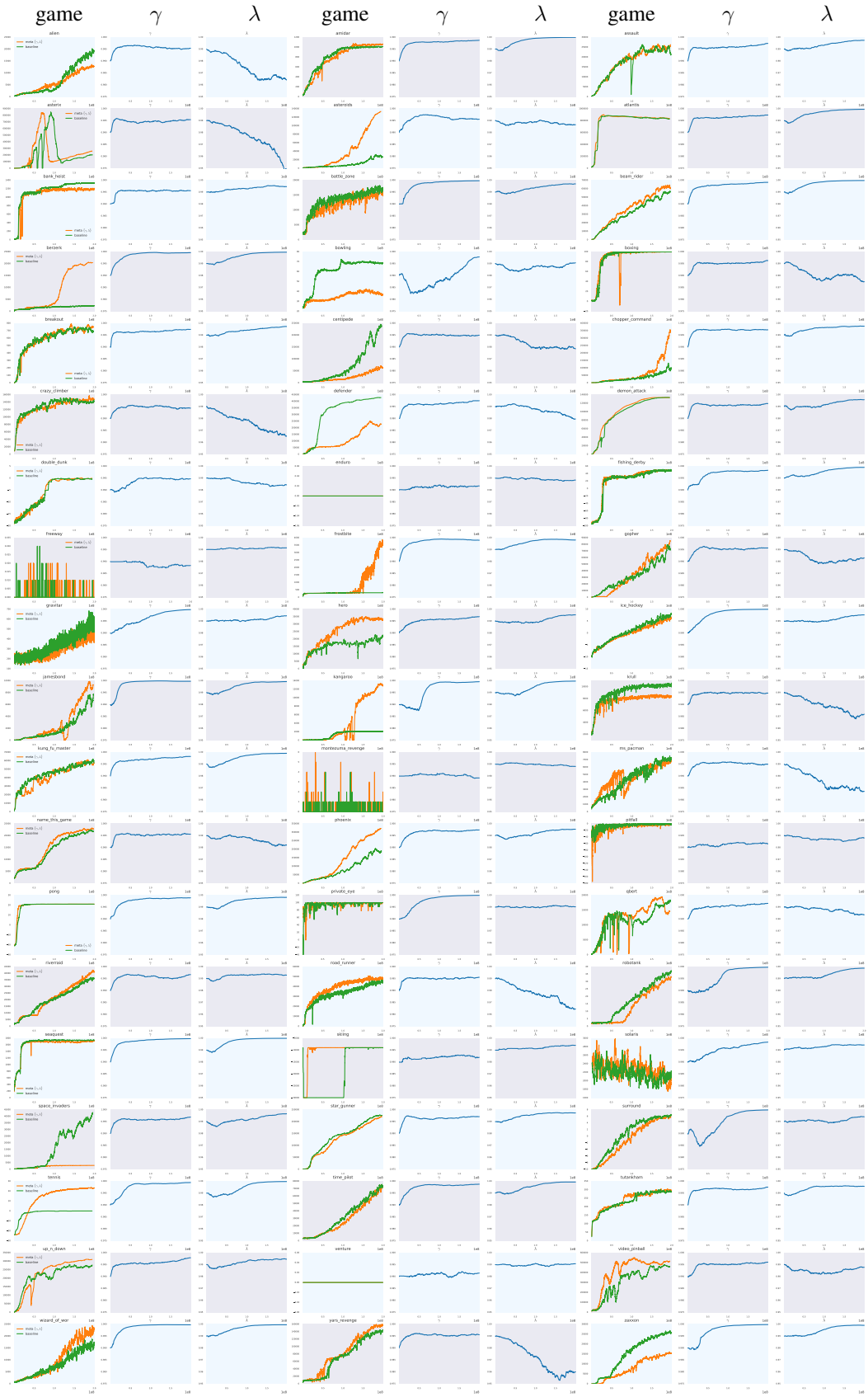


Figure 5: Training curves for meta learning $\eta = \{\gamma, \lambda\}$. We provide the comparison of scores against baseline, the change of γ, and the change of λ for each game. Best viewed in electronic version.