# Beyond Variance Reduction: Understanding the True Impact of Baselines on Policy Optimization

**Wesley Chung** [*1] **Valentin Thomas** [*2] **Marlos C. Machado** [3 4 5] **Nicolas Le Roux** [6 1 2]

## Abstract

Bandit and reinforcement learning (RL) problems can often be framed as optimization problems where the goal is to maximize average performance while having access only to stochastic estimates of the true gradient. Traditionally, stochastic optimization theory predicts that learning dynamics are governed by the curvature of the loss function and the noise of the gradient estimates. In this paper we demonstrate that the standard view is too limited for bandit and RL problems. To allow our analysis to be interpreted in light of multi-step MDPs, we focus on techniques derived from stochastic optimization principles (e.g., natural policy gradient and EXP3) and we show that some standard assumptions from optimization theory are violated in these problems. We present theoretical results showing that, at least for bandit problems, curvature and noise are not sufficient to explain the learning dynamics and that seemingly innocuous choices like the baseline can determine whether an algorithm converges. These theoretical findings match our empirical evaluation, which we extend to multi-state MDPs.

## 1. Introduction

In the standard multi-arm bandit setting (Robbins, 1952), an agent needs to choose, at each timestep $t$, an arm $a_t \in \{1, ..., n\}$ to play, receiving a potentially stochastic reward $r_t$ with mean $\mu_{a_t}$. The goal of the agent is usually to maximize the total sum of rewards, $\sum_{i=1}^{T} r_i$, or to maximize the average performance at time $T$, $\mathbb{E}_{i\sim\pi}\mu_i$ with $\pi$ being the probability of the agent of drawing each arm (Bubeck & Cesa-Bianchi, 2012). While the former measure is often used in the context of bandits,[1] $\mathbb{E}_{i\sim\pi}\mu_i$ is more common in the context of Markov Decision Processes (MDPs), which have multi-arm bandits as a special case.

In this paper we focus on techniques derived from stochastic optimization principles, such as EXP3 (Auer et al., 2002; Seldin et al., 2013). In particular, we study *policy gradient* methods, a family of algorithms useful in the more general MDP setting which have seen empirical success in recent times (Schulman et al., 2017).

We analyze the problem of learning to maximize the average reward, $J$, by gradient ascent:

$$\theta^* = \arg\max_\theta J(\theta) = \arg\max_\theta \sum_a \pi_\theta(a)\mu_a \ , \quad (1)$$

with $\mu_a$ being the average reward of arm $a$. In this case, we are mainly interested in outputting an effective policy at the end of the optimization process, without explicitly considering the performance of intermediary policies.

Optimization theory predicts that the convergence speed of stochastic gradient methods will be affected by the variance of the gradient estimates and by the geometry of the function $J$, represented by its curvature. Roughly speaking, the geometry dictates how effective true gradient ascent is at optimizing $J(\theta)$ while the variance can be viewed as a penalty, capturing how much slower the optimization process is by using noisy versions of this true gradient. More concretely, doing one gradient step with stepsize $\alpha$, using a stochastic estimate $g_t$ of the gradient, leads to (Bottou et al., 2018):

$$\mathbb{E}[J(\theta_{t+1})] - J(\theta_t) \geq (\alpha - \tfrac{L\alpha^2}{2})\|\mathbb{E}[g_t]\|_2^2 - \tfrac{L\alpha^2}{2}\text{Var}[g_t],$$

when $J$ is $L$-smooth, i.e. its gradients are $L$-Lipschitz.

As large variance has been identified as an issue for policy gradient (PG) methods, many works have focused on reducing the noise of the updates. One common technique is the use of control variates (Greensmith et al., 2004; Hofmann et al., 2015), referred to as *baselines* in the context of RL. These baselines $b$ are subtracted from the observed returns to obtain shifted returns, $r(a_i) - b$, and do not change the

[*]Equal contribution [1]Mila, McGill University [2]Mila, University of Montreal [3]DeepMind [4]Amii, University of Alberta [5]Work partially done at Google Research [6]Google Research, Brain Team. Correspondence to: <wesley.chung2@gmail.com, vltn.thomas@gmail.com, marlosm@deepmind.com, nicolas@leroux.name>.

*Proceedings of the $38^{th}$ International Conference on Machine Learning*, PMLR 139, 2021. 

[1]The objective is usually presented as regret minimization.

expectation of the gradient. In MDPs, they are typically state-dependent. While the value function is a common choice, previous work showed that the minimum-variance baseline for the REINFORCE (Williams, 1992) estimator is different and involves the norm of the gradient (Peters & Schaal, 2008). Reducing variance has been the main motivation for many previous works on baselines (e.g., Gu et al., 2016; Liu et al., 2017; Grathwohl et al., 2017; Wu et al., 2018; Cheng et al., 2020), but the influence of baselines on other aspects of the optimization process has hardly been studied. We take a deeper look at baselines and their effects on optimization.

CONTRIBUTIONS

We show that baselines can impact the optimization process beyond variance reduction and lead to qualitatively different learning curves, even when the variance of the gradients is the same. For instance, given two baselines with the same variance, the more negative baseline promotes *committal* behaviour where a policy quickly tends towards a deterministic one, while the more positive baseline leads to *non-committal* behaviour, where the policy retains higher entropy for a longer period.

Furthermore, we show that **the choice of baseline can even impact the convergence of natural policy gradient** (NPG), something variance cannot explain. In particular, we construct a three-armed bandit where using the baseline minimizing the variance can lead to convergence to a deterministic, sub-optimal policy for any positive stepsize, while another baseline, with larger variance, guarantees convergence to the optimal policy. As such a behaviour is impossible under the standard assumptions in optimization, this result shows how these assumptions may be violated in practice. It also provides a counterexample to the convergence of NPG algorithms in general, a popular variant with much faster convergence rates than vanilla PG when using the true gradient in tabular MDPs (Agarwal et al., 2019).

Further, **we identify on-policy sampling as a key factor to these convergence issues** as it induces a vicious cycle where making bad updates can lead to worse policies, in turn leading to worse updates. A natural solution is to break the dependency between the sampling distribution and the updates through off-policy sampling. We show that ensuring all actions are sampled with sufficiently large probability at each step is enough to guarantee convergence in probability. Note that this form of convergence is stronger than convergence of the expected iterates, a more common type of result (e.g., Mei et al., 2020b; Agarwal et al., 2019).

We also perform an empirical evaluation on multi-step MDPs, showing that baselines have a similar impact in that setting. We observe **a significant impact on the empirical performance** of agents when using two different sets of baselines yielding the same variance, once again suggesting that learning dynamics in MDPs are governed by more than the curvature of the loss and the variance of the gradients.

## 2. Baselines, learning dynamics & exploration

The problem defined in Eq. 1 can be solved by gradient ascent. Given access only to samples, the true gradient cannot generally be computed and the true update is replaced with a stochastic one, resulting in the following update:

$$\theta_{t+1} = \theta_t + \frac{\alpha}{N} \sum_i r(a_i) \nabla_\theta \log \pi_\theta(a_i) \ , \qquad (2)$$

where $a_i$ are actions drawn according to the agent's current policy $\pi_\theta$, $\alpha$ is the stepsize, and $N$, which can be 1, is the number of samples used to compute the update. To reduce the variance of this estimate without introducing bias, we can introduce a baseline $b$, resulting in the gradient estimate $(r(a_i) - b)\nabla_\theta \log \pi_\theta(a_i)$.

While the choice of baseline is known to affect the variance, we show that baselines can also lead to qualitatively different behaviour of the optimization process, even when the variance is the same. This difference cannot be explained by the expectation or variance, quantities which govern the usual bounds for convergence rates (Bottou et al., 2018).

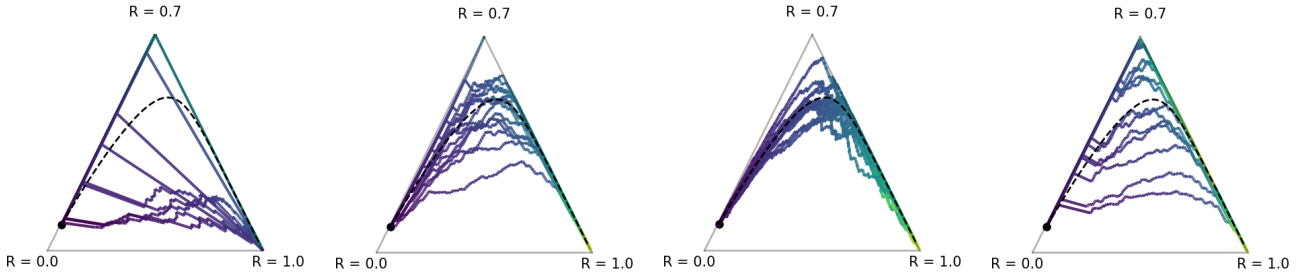
### 2.1. Committal and non-committal behaviours

To provide a complete picture of the optimization process, we analyze the evolution of the policy during optimization. We start in a simple setting, a deterministic three-armed bandit, where it is easier to produce informative visualizations.

To eliminate variance as a potential confounding factor, we consider different baselines with the same variance. We start by computing the baseline leading to the minimum-variance of the gradients for the algorithm we use. For vanilla policy gradient, we have $b^*_\theta = \frac{\mathbb{E}[r(a_i)\|\nabla \log \pi_\theta(a_i)\|_2^2]}{\mathbb{E}[\|\nabla \log \pi_\theta(a_i)\|_2^2]}$ (Peters & Schaal, 2008; Greensmith et al., 2004) (see Appendix D.1 for details and the NPG version). Note that this baseline depends on the current policy and changes throughout the optimization. As the variance is a quadratic function of the baseline, the two baselines $b^+_\theta = b^*_\theta + \epsilon$ and $b^-_\theta = b^*_\theta - \epsilon$ result in gradients with the same variance (see Appendix D.4 for details). Thus, we use these two perturbed baselines to demonstrate that there are phenomena in the optimization process that variance cannot explain.

Fig. 1 presents fifteen learning curves on the probability simplex representing the space of possible policies for the three-arm bandit, when using NPG and a softmax parameterization. We choose $\epsilon = 1/2$ to obtain two baselines with the same variance: $b^+_\theta = b^*_\theta + 1/2$ and $b^-_\theta = b^*_\theta - 1/2$.

Inspecting the plots, the learning curves for $\epsilon = -1/2$ and




(a) $b_\theta^- = b_\theta^* - 1/2$ (b) $b_\theta = b_\theta^*$ (c) $b_\theta^+ = b_\theta^* + 1/2$ (d) $b_\theta = V^{\pi_\theta}$

Figure 1: We plot 15 different trajectories of natural policy gradient with softmax parameterization, when using various baselines, on a 3-arm bandit problem with rewards $(1, 0.7, 0)$ and stepsize $\alpha = 0.025$ and $\theta_0 = (0, 3, 5)$. The black dot is the initial policy and colors represent time, from purple to yellow. The dashed black line is the trajectory when following the true gradient (which is unaffected by the baseline). Different values of $\epsilon$ denote different perturbations to the minimum-variance baseline. We see some cases of convergence to a suboptimal policy for both $\epsilon = -1/2$ and $\epsilon = 0$. This does not happen for the larger baseline $\epsilon = 1/2$ or the value function as baseline. Figure made with Ternary (Harper & Weinstein, 2015).

$\epsilon = 1/2$ are qualitatively different, even though the gradient estimates have the same variance. For $\epsilon = -1/2$, the policies quickly reach a deterministic policy (i.e., a neighborhood of a corner of the probability simplex), which can be suboptimal, as indicated by the curves ending up at the policy choosing action 2. On the other hand, for $\epsilon = 1/2$, every learning curve ends up at the optimal policy, although the convergence might be slower. The learning curves also do not deviate much from the curve for the true gradient. Again, these differences cannot be explained by the variance since the baselines result in identical variances.

Additionally, for $b_\theta = b_\theta^*$, the learning curves spread out further. Compared to $\epsilon = 1/2$, some get closer to the top corner of the simplex, leading to convergence to a suboptimal solution, suggesting that the minimum-variance baseline may be worse than other, larger baselines. In the next section, we theoretically substantiate this and show that, for NPG, it is possible to converge to a suboptimal policy with the minimum-variance baseline; but there are larger baselines that guarantee convergence to an optimal policy.

We look at the update rules to explain these different behaviours. When using a baseline $b$ with NPG, sampling $a_i$ results in the update

$$\begin{aligned}\theta_{t+1} &= \theta_t + \alpha[r(a_i) - b]F_\theta^{-1}\nabla_\theta \log \pi_\theta(a_i) \\ &= \theta_t + \alpha\frac{r(a_i) - b}{\pi_\theta(a_i)}\mathbf{1}_{a_i} + \alpha\lambda e\end{aligned}$$

where $F_\theta^{-1} = \mathbb{E}_{a\sim\pi}[\nabla \log \pi_\theta(a) \nabla \log \pi_\theta(a)^\top]$, $\mathbf{1}_{a_i}$ is a one-hot vector with 1 at index $i$, and $\lambda e$ is a vector containing $\lambda$ in each entry. The second line follows for the softmax policy (see Appendix D.2) and $\lambda$ is arbitrary since shifting $\theta$ by a constant does not change the policy.

Thus, supposing we sample action $a_i$, if $r(a_i) - b$ is positive, which happens more often when the baseline $b$ is small (more negative), the update rule will increase the probability $\pi_\theta(a_i)$. This leads to an increase in the probability of taking the actions the agent took before, regardless of their quality (see Fig.1a for $\epsilon = -1/2$). Because the agent is likely to choose the same actions again, we call this *committal* behaviour.

While a smaller baseline leads to committal behaviour, a larger (more positive) baseline makes the agent second-guess itself. If $r(a_i) - b$ is negative, which happens more often when $b$ is large, the parameter update decreases the probability $\pi_\theta(a_i)$ of the sampled action $a_i$, reducing the probability the agent will re-take the actions it just took, while increasing the probability of other actions. This might slow down convergence but it also makes it harder for the agent to get stuck. This is reflected in the $\epsilon = 1/2$ case (Fig.1c), as all the learning curves end up at the optimal policy. We call this *non-committal* behaviour.

While the previous experiments used perturbed variants of the minimum-variance baseline to control for the variance, this baseline would usually be infeasible to compute in more complex MDPs. Instead, a more typical choice of baseline would be the value function (Sutton & Barto, 2018, Ch. 13), which we evaluate in Fig. 1d. Choosing the value function as a baseline generated trajectories converging to the optimal policy, even though their convergence may be slow, despite it not being the minimum variance baseline. The reason becomes clearer when we write the value function as $V^\pi = b_\theta^* - \frac{\mathrm{Cov}(r, \|\nabla \log \pi\|^2)}{\mathbb{E}[\|\nabla \log \pi\|^2]}$ (see Appendix D.3). The term $\mathrm{Cov}(r, \|\nabla \log \pi\|^2)$ typically becomes negative as the gradient becomes smaller on actions with high rewards during the optimization process, leading to the value function being a noncommittal baseline, justifying a choice often made by practitioners.

Additional empirical results can be found in Appendix A.1 for natural policy gradient and vanilla policy gradient for the softmax parameterization. Furthermore, we explore the use of different parameterizations: First, we testprojected stochastic gradient ascent and directly optimizing the policy probabilities $\pi_\theta(a)$. Next, we try the escort transform (Mei et al., 2020a), which was designed to improve the curvature of the objective. We find qualitatively similar results in all cases; baselines can induce *committal* and *non-committal* behaviour.

## 3. Convergence to suboptimal policies with natural policy gradient (NPG)

We empirically showed that PG algorithms can reach suboptimal policies and that the choice of baseline can affect the likelihood of this occurring. In this section, we provide theoretical results proving that it is indeed possible to converge to a suboptimal policy when using NPG. We discuss how this finding fits with existing convergence results and why standard assumptions are not satisfied in this setting.

### 3.1. A simple example

Standard convergence results assume access to the true gradient (e.g., Agarwal et al., 2019) or, in the stochastic case, assume that the variance of the updates is uniformly bounded for all parameter values (e.g., Bottou et al., 2018). These assumptions are in fact quite strong and are violated in a simple two-arm bandit problem with fixed rewards. Pulling the optimal arm gives a reward of $r_1 = +1$, while pulling the suboptimal arm leads to a reward of $r_0 = 0$. We use the sigmoid parameterization and call $p_t = \sigma(\theta_t)$ the probability of sampling the optimal arm at time $t$.

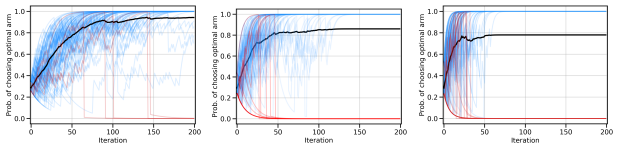
Our stochastic estimator of the natural gradient is

$$g_t = \begin{cases} \frac{1-b}{p_t}, \text{with probability } p_t \\ \frac{b}{1-p_t}, \text{with probability } 1-p_t, \end{cases}$$

where $b$ is a baseline that does not depend on the action sampled at time $t$ but may depend on $\theta_t$. By computing the variance of the updates, $\text{Var}[g_t] = \frac{(1-p_t-b)^2}{p_t(1-p_t)}$, we notice it is unbounded when the policy becomes deterministic, i.e. $p_t \to 0$ or $p_t \to 1$, violating the assumption of uniformly bounded variance, unless $b = 1 - p_t$, which is the optimal baseline. Note that using vanilla (non-natural) PG would, on the contrary, yield a bounded variance. In fact, we prove a convergence result in its favour in Appendix B (Prop. 4).

For NPG, the proposition below establishes potential convergence to a suboptimal arm and we demonstrate this empirically in Fig. 2.

**Proposition 1.** *Consider a two-arm bandit with rewards* $1$ *and* $0$ *for the optimal and suboptimal arms, respectively.*




(a) $\alpha = 0.05$ (b) $\alpha = 0.1$ (c) $\alpha = 0.15$

Figure 2: Learning curves for 100 runs of 200 steps, on the two-arm bandit, with baseline $b = -1$ for three different stepsizes $\alpha$. *Blue:* Curves converging to the optimal policy. *Red:* Curves converging to a suboptimal policy. *Black:* Avg. performance. The number of runs that converged to the suboptimal solution are 5%, 14% and 22% for the three $\alpha$'s. Larger $\alpha$'s are more prone to getting stuck at a suboptimal solution but settle on a deterministic policy more quickly.

*Suppose we use natural policy gradient starting from $\theta_0$, with a fixed baseline $b < 0$, and fixed stepsize $\alpha > 0$. If the policy samples the optimal action with probability $\sigma(\theta)$, then the probability of picking the suboptimal action forever and having $\theta_t$ go to $-\infty$ is strictly positive. Additionally, if $\theta_0 \leq 0$, we have*

$$P(\text{suboptimal action forever}) \geq (1-e^{\theta_0})(1-e^{\theta_0+\alpha b})^{-\frac{1}{\alpha b}}.$$

*Proof.* All the proofs may be found in the appendix. □

The updates provide some intuition as to why there is convergence to suboptimal policies. The issue is the *committal* nature of the baseline. Choosing an action leads to an increase of that action's probability, even if it is a poor choice. Choosing the suboptimal arm leads to a decrease in $\theta$ by $\frac{\alpha b}{1-p_t}$, thus increasing the probability the same arm is drawn again and further decreasing $\theta$. By checking the probability of this occurring forever, $P(\text{suboptimal arm forever}) = \prod_{t=1}^{\infty}(1-p_t)$, we show that $1-p_t$ converges quickly enough to 1 that the infinite product is nonzero, showing it is possible to get trapped choosing the wrong arm forever (Prop. 1), and $\theta_t \to -\infty$ as $t$ grows.

This issue could be solved by picking a baseline with lower variance. For instance, the minimum-variance baseline $b = 1 - p_t$ leads to 0 variance and both possible updates are equal to $+\alpha$, guaranteeing that $\theta \to +\infty$, thus convergence. In fact, any baseline $b \in (0, 1)$ suffices since both updates are positive and greater than $\alpha \min(b, 1-b)$. However, this is not always the case, as we show in the next section.

To decouple the impact of the variance with that of the committal nature of the baseline, Prop. 2 analyzes the learning dynamics in the two-arm bandit case for perturbations of the optimal baseline, i.e. we study baselines of the form $b = b^* + \epsilon$ and show how $\epsilon$, and particularly its sign, affects learning. Note that, because the variance is a quadratic func-

tion with its minimum in $b^*$, both $+\epsilon$ and $-\epsilon$ have the same variance. Our findings can be summarized as follows:

**Proposition 2.** *For the two-armed bandit defined in Prop. 1, when using a perturbed min-variance baseline $b = b^* + \epsilon$, the value of $\epsilon$ determines the learning dynamics as follows:*

- *For $\epsilon < -1$, there is a positive probability of converging to the suboptimal arm.*
- *For $\epsilon \in (-1, 1)$, we have convergence in probability to the optimal policy.*
- *For $\epsilon \geq 1$, the supremum of the iterates goes to $+\infty$ in probability.*

While the proofs can be found in Appendix B.2, we provide here some intuition behind these results.

For $\epsilon < -1$, we reuse the same argument as for $b < 0$ in Prop. 1. The probability of drawing the correct arm can decrease quickly enough to lead to convergence to the suboptimal arm.

For $\epsilon \in (-1, 1)$, the probability of drawing the correct arm cannot decrease too fast. Hence, although the updates, as well as the variance of the gradient estimate, are potentially unbounded, we still have convergence to the optimal solution in probability.

Finally, for $\epsilon \geq 1$, we can reuse an intermediate argument from the $\epsilon \in (0, 1)$ case to argue that for any threshold $C$, the parameter will eventually exceed that threshold. For $\epsilon \in (0, 1)$, once a certain threshold is crossed, the policy is guaranteed to improve at each step. However, with a large positive perturbation, updates are larger and we lose this additional guarantee, leading to the weaker result.

We want to emphasize that not only we get provably different dynamics for $\epsilon < -1$ and $\epsilon \geq 1$, showing the importance of the sign of the perturbation, but that there also is a sharp transition around $|\epsilon| = 1$, which cannot be captured solely by the variance.

The above analysis was specific to these updates. To predict committal vs. non-committal behaviour more generally, it may be possible to utilize higher order moments or other distributional properties, even when the mean and variance is the same. Unfortunately, it is difficult to utilize higher-moment information in theoretical bounds in a general manner as Markov-type inequalities do not take into account the sign of the higher moment, which we think is where the committal vs. non-committal distinction would appear.

### 3.2. Reducing variance with baselines can be detrimental

As we saw with the two-armed bandit, the direction of the updates is important in assessing convergence. More specifically, problems can arise when the choice of baseline induces committal behaviour. We now show a different bandit setting where committal behaviour happens even when using the minimum-variance baseline, thus leading to convergence to a suboptimal policy. Furthermore, we design a better baseline which ensures all updates move the parameters towards the optimal policy. This cements the idea that the quality of parameter updates must not be analyzed in terms of variance but rather in terms of the probability of going in a bad direction, since a baseline that induces higher variance leads to convergence while the minimum-variance baseline does not. The following theorem summarizes this.

**Theorem 1.** *There exists a three-arm bandit where using the stochastic natural gradient on a softmax-parameterized policy with the minimum-variance baseline can lead to convergence to a suboptimal policy with probability $\rho > 0$, and there is a different baseline (with larger variance) which results in convergence to the optimal policy with probability 1.*

The bandit used in this theorem is the one we used for the experiments depicted in Fig. 1. The key is that the minimum-variance baseline can be lower than the second best reward; so pulling the second arm will increase its probability and induce committal behaviour. This can cause the agent to prematurely commit to the second arm and converge to the wrong policy. On the other hand, using any baseline whose value is between the optimal reward and the second best reward, which we term a *gap* baseline, will always increase the probability of the optimal action at every step, no matter which arm is drawn. Since the updates are sufficiently large at every step, this is enough to ensure convergence with probability 1, despite the higher variance compared to the minimum variance baseline. The key is that whether a baseline underestimates or overestimates the second best reward can affect the algorithm convergence and this is more critical than the resulting variance of the gradient estimates.

As such, more than lower variance, good baselines are those that can assign positive effective returns to the good trajectories and negative effective returns to the others. These results cast doubt on whether finding baselines which minimize variance is a meaningful goal to pursue. The baseline can affect optimization in subtle ways, beyond variance, and further study is needed to identify the true causes of some improved empirical results observed in previous works. This importance of the sign of the returns, rather than their exact value, echoes with the cross-entropy method (De Boer et al., 2005), which maximizes the probability of the trajectories with the largest returns, regardless of their actual value.

## 4. Off-policy sampling

So far, we have seen that *committal* behaviour can be problematic as it can cause convergence to a suboptimal policy.

This can be especially problematic when the agent follows a near-deterministic policy as it is unlikely to receive different samples which would move the policy away from the closest deterministic one, regardless of the quality of that policy.

Up to this point, we assumed that actions were sampled according to the current policy, a setting known as *on-policy*. This setting couples the updates and the policy and is a root cause of the *committal* behaviour: the update at the current step changes the policy, which affects the distribution of rewards obtained and hence the next updates. However, we know from the optimization literature that bounding the variance of the updates will lead to convergence (Bottou et al., 2018). As the variance becomes unbounded when the probability of drawing some actions goes to 0, a natural solution to avoid these issues is to sample actions from a behaviour policy that selects every action with sufficiently high probability. Such a policy would make it impossible to choose the same, suboptimal action forever.

### 4.1. Convergence guarantees with IS

Because the behaviour policy changed, we introduce importance sampling (IS) corrections to preserve the unbiased updates (Kahn & Harris, 1951; Precup, 2000). These changes are sufficient to guarantee convergence for any baseline:

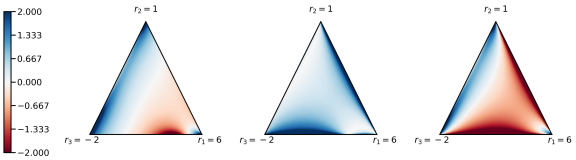
**Proposition 3.** *Consider a $n$-armed bandit with stochastic rewards with bounded support and a unique optimal action. The behaviour policy $\mu_t$ selects action $i$ with probability $\mu_t(i)$ and let $\epsilon_t = \min_i \mu_t(i)$. When using NPG with importance sampling and a bounded baseline $b$, if $\lim_{t\to\infty} t\,\epsilon_t^2 = +\infty$ , then the target policy $\pi_t$ converges to the optimal policy in probability.*

*Proof. (Sketch)* Using Azuma-Hoeffding's inequality, we can show that for well chosen constants $\Delta_i, \delta$ and $C > 0$ ,

$$\mathbb{P}\left(\theta_t^1 \geq \theta_0^1 + \alpha\delta\Delta_1 t\right) \quad \geq \quad 1 - \exp\left(-\frac{\delta^2\Delta_1^2}{2C^2} t\epsilon_t^2\right)$$

where $\theta^1$ is the parameter associated to the optimal arm. Thus if $\lim_{t\to\infty} t\epsilon_t^2 = +\infty$, the RHS goes to 1. In a similar manner, we can upper bound $\mathbb{P}\left(\theta_t^i \geq \theta_0^i + \alpha\delta\Delta_i t\right)$ for all suboptimal arms, and applying an union bound, we get the desired result. □

The condition on $\mu_t$ imposes a cap on how fast the behaviour policy can become deterministic: no faster than $t^{-1/2}$. Intuitively, this ensures each action is sampled sufficiently often and prevents premature convergence to a suboptimal policy. The condition is satisfied for any sequence of behaviour policies which assign at least $\epsilon_t$ probability to each action at each step, such as $\epsilon$-greedy policies. It also holds if $\epsilon_t$ decreases over time at a sufficiently slow rate. By choosing as behaviour policy $\mu$ a linear interpolation between $\pi$ and the uniform policy, $\mu(a) = (1-\gamma)\pi(a) + \frac{\gamma}{K}, \gamma \in (0, 1]$, where $K$ is the number of arms, we recover the classic EXP3 algorithm (Auer et al., 2002; Seldin et al., 2012).




(a) $b = 0$ / IS. (b) $b = 0$ / $b^*$. (c) $b^*$ / IS.

Figure 3: Comparison between the variance of different methods on a 3-arm bandit. Each plot depicts the log of the ratio between the variance of two approaches. For example, Fig. (a) depicts $\log \frac{\mathrm{Var}[g_{b=0}]}{\mathrm{Var}[g_{\mathrm{IS}}]}$, the log of the ratio between the variance of the gradients of PG without a baseline and PG with IS. The triangle represents the probability simplex with each corner representing a deterministic policy on a specific arm. The method written in blue (resp. red) in each figure has lower variance in blue (resp. red) regions of the simplex. The sampling policy $\mu$, used in the PG method with IS, is a linear interpolation between $\pi$ and the uniform distribution, $\mu(a) = \frac{1}{2}\pi(a) + \frac{1}{6}$. Note that this is not the min. variance sampling distribution and it leads to higher variance than PG without a baseline in some parts of the simplex.

We can also confirm that this condition is not satisfied for the simple example we presented when discussing convergence to suboptimal policies. There, $p_t$ could decrease exponentially fast since the tails of the sigmoid function decay exponentially and the parameters move by at least a constant at every step. In this case, $\epsilon_t = \Omega(e^{-t})$, resulting in $\lim_{t\to\infty} te^{-2t} = 0$, so Proposition 3 does not apply.

### 4.2. Importance sampling, baselines & variance

As we have seen, using a separate behaviour policy that samples all actions sufficiently often may lead to stronger convergence guarantees, even if it increases the variance of the gradient estimates in most of the space, as what matters is what happens in the high variance regions, which are usually close to the boundaries. Fig. 3 shows the ratios of gradient variances between on-policy PG without baseline, on-policy PG with the minimum variance baseline, and off-policy PG using importance sampling (IS) where the sampling distribution is $\mu(a) = \frac{1}{2}\pi(a) + \frac{1}{6}$, i.e. a mixture of the current policy $\pi$ and the uniform distribution. While using the minimum variance baseline decreases the variance on the entire space compared to not using a baseline, IS actually *increases* the variance when the current policy is close to uniform. However, IS does a much better job at reducing the variance close to the boundaries of the simplex, where it actually matters to guarantee convergence.

This suggests that convergence of PG methods is not so much governed by the variance of the gradient estimates in general, but by the variance in the worst regions, usually near the boundary. While baselines can reduce the variance, they generally cannot prevent the variance in those regions from exploding, leading to the policy getting stuck. Thus, good baselines are not the ones reducing the variance across the space but rather those that can prevent the learning from reaching these regions altogether. Large values of $b$, such that $r(a_i) - b$ is negative for most actions, achieve precisely that. On the other hand, due to the increased flexibility of sampling distributions, IS can limit the nefariousness of these critical regions, offering better convergence guarantees despite not reducing variance everywhere.

Importantly, although IS is usually used in RL to correct for the distribution of past samples (e.g., Munos et al., 2016), we advocate here for expanding the research on designing appropriate sampling distributions as done by Hanna et al. (2017; 2018) and Parmas & Sugiyama (2019). This line of work has a long history in statistics (c.f., Liu, 2008).

### 4.3. Other mitigating strategies

We conclude this section by discussing alternative strategies to mitigate the convergence issues. While they might be effective, and some are indeed used in practice, they are not without pitfalls.

First, one could consider reducing the stepsizes, with the hope that the policy would not converge as quickly towards a suboptimal deterministic policy and would eventually leave that bad region. Indeed, if we are to use vanilla PG in the two-arm bandit example, instead of NPG, this effectively reduces the stepsize by a factor of $\sigma(\theta)(1-\sigma(\theta))$ (the Fisher information). In this case, we are able to show convergence in probability to the optimal policy. See Proposition 4 in Appendix B.

Empirically, we find that, when using vanilla PG, the policy may still remain stuck near a suboptimal policy when using a negative baseline, similar to Fig. 2. While the previous proposition guarantees convergence eventually, the rate may be very slow, which remains problematic in practice. There is theoretical evidence that following even the true vanilla PG may result in slow convergence (Schaul et al., 2019), suggesting that the problem is not necessarily due to noise.

An alternative solution would be to add entropy regularization to the objective. By doing so, the policy would be prevented from getting too close to deterministic policies. While this might prevent convergence to a suboptimal policy, it would also exclude the possibility of fully converging to the optimal policy, though the policy may remain near it.

In bandits, EXP3 has been found not to enjoy high-probability guarantees on its regret so variants have been developed to address this deficiency (c.f. Lattimore & Szepesvári, 2020). For example, by introducing bias in the updates, their variance can be reduced significantly (Auer et al., 2002; Neu, 2015). Finally, other works have also developed provably convergent policy gradient algorithms using different mechanisms, such as exploration bonuses or ensembles of policies (Cai et al., 2019; Efroni et al., 2020; Agarwal et al., 2020).

## 5. Extension to multi-step MDPs

We focused our theoretical analyses on multi-arm bandits so far. However, we are also interested in more general environments where gradient-based methods are commonplace. We now turn our attention to the Markov Decision Process (MDP) framework (Puterman, 2014). An MDP is a set $\{\mathcal{S}, \mathcal{A}, P, r, \gamma, \rho\}$ where $\mathcal{S}$ and $\mathcal{A}$ are the set of states and actions, $P$ is the environment transition function, $r$ is the reward function, $\gamma \in [0, 1)$ the discount factor, and $\rho$ is the initial state distribution. The goal of RL algorithms is to find a policy $\pi_\theta$, parameterized by $\theta$, which maximizes the (discounted) expected return; i.e. Eq. 1 becomes

$$\arg\max_\theta J(\theta) = \arg\max_\theta \sum_s d_\gamma^{\pi_\theta}(s) \sum_a \pi_\theta(a|s) r(s,a),$$

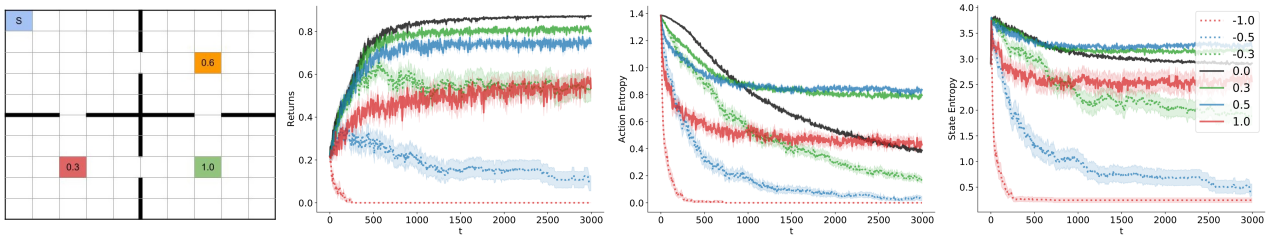
where there is now a discounted distribution over states induced by $\pi_\theta$. Although that distribution depends on $\pi_\theta$ in a potentially complex way, the parameter updates are similar to Eq. 2:

$$\theta_{t+1} = \theta_t + \frac{\alpha}{N} \sum_i [Q(s_i, a_i) - b(s_i)] \nabla_\theta \log \pi_\theta(a_i|s_i) \ ,$$

where $(a_i, s_i)$ pairs are drawn according to the discounted state-visitation distribution induced by $\pi_\theta$ and $Q$ is the state-action value function induced by $\pi_\theta$ (c.f. Sutton & Barto, 2018). To match the bandit setting and common practice, we made the baseline state dependent.

Although our theoretical analyses do not easily extend to multi-step MDPs, we empirically investigated if the similarity between these formulations leads to similar differences in learning dynamics when changing the baseline. We consider a 10x10 gridworld consisting of 4 rooms as depicted on Fig. 4a. We use a discount factor $\gamma = 0.99$. The agent starts in the upper left room and two adjacent rooms contain a goal state of value $0.6$ or $0.3$. The best goal (even discounted), with a value of $1$, lies in the furthest room, so that the agent must learn to cross the sub-optimal rooms and reach the furthest one.

Similar to the bandit setting, for a state $s$, we can derive the minimum-variance baseline $b^*(s)$ assuming access to state-action values $Q(s, a)$ for $\pi_\theta$ and consider perturbations to it. Again, we use baselines $b(s) = b^*(s)+\epsilon$ and $b(s) = b^*(s)-$




(a) MDP (b) Returns (c) Entropy (A) (d) Entropy (S)

Figure 4: We plot the discounted returns, the entropy of the policy over the states visited in each trajectory, and the entropy of the state visitation distribution, averaged over 50 runs, for multiple baselines. The baselines are of the form $b(s) = b^*(s) + \epsilon$, perturbations of the minimum-variance baseline, with $\epsilon$ indicated in the legend. The shaded regions denote one standard error. Note that the policy entropy of lower baselines tends to decay faster than for larger baselines. Also, smaller baselines tend to get stuck on suboptimal policies, as indicated by the returns plot. See text for additional details.

$\epsilon$, since they result in identical variances (this would not be the case if we used standard REINFORCE). We use a natural policy gradient estimate, which substitutes $\nabla \log \pi(a_i|s_i)$ by $F_{s_i}^{-1} \nabla \log \pi(a_i|s_i)$ in the update rule, where $F_{s_i}$ is the Fisher information matrix for state $s_i$ and solve for the exact $Q(s, a)$ values using dynamic programming for all updates (see Appendix D.6 for details).

In order to identify the committal vs. non-committal behaviour of the agent depending on the baseline, we monitor the entropy of the policy and the entropy of the stationary state distribution over time. Fig.4b shows the average returns over time and Fig.4c and 4d show the entropy of the policy in two ways. The first is the average entropy of the action distribution along the states visited in each trajectory, and the second is the entropy of the distribution of the number of times each state is visited up to that point in training.

The action entropy for smaller baselines tends to decay faster compared to larger ones, indicating convergence to a deterministic policy. This quick convergence is premature in some cases since the returns are not as high for the lower baselines. In fact for $\epsilon = -1$, we see that the agent gets stuck on a policy that is unable to reach any goal within the time limit, as indicated by the returns of $0$. On the other hand, the larger baselines tend to achieve larger returns with larger entropy policies, but do not fully converge to the optimal policy as evidenced by the gap in the returns plot.

Since committal and non-committal behaviour can be directly inferred from the PG and the sign of the effective rewards $R(\tau) - b$, we posit that these effects extend to all MDPs. In particular, in complex MDPs, the first trajectories explored are likely to be suboptimal and a low baseline will increase their probability of being sampled again, requiring the use of techniques such as entropy regularization to prevent the policy from getting stuck too quickly. In some preliminary experiments with a deep RL policy gradient algorithm, PPO (Schulman et al., 2017), where we perturb the baseline by a fixed constant, seem to indicate that negative perturbations perform slightly worse than positive perturbations. The results are not conclusive though and there are many confounding factors in this setting which could affect the outcome, including clipping due to PPO, neural network generalization, and adaptive optimizers. It is likely that a more careful strategy to perturb the baseline is needed to gain benefits, similar to using exploration bonuses.

## 6. Conclusion

We presented results that dispute common beliefs about baselines, variance, and policy gradient methods in general. As opposed to the common belief that baselines only provide benefits through variance reduction, we showed that they can significantly affect the optimization process in ways that cannot be explained by the variance and that lower variance can even sometimes be detrimental.

Different baselines can give rise to very different learning dynamics, even when they reduce the variance of the gradients equally. They do that by either making a policy quickly tend towards a deterministic one (*committal* behaviour) or by maintaining high-entropy for a longer period of time (*non-committal* behaviour). We showed that *committal* behaviour can be problematic and lead to convergence to a suboptimal policy. Specifically, we showed that stochastic natural policy gradient does not always converge to the optimal solution due to the unusual situation in which the iterates converge to the optimal policy in expectation but not almost surely. Moreover, we showed that baselines that lead to lower-variance can sometimes be detrimental to optimization, highlighting the limitations of using variance to analyze the convergence properties of these methods. We also showed that standard convergence guarantees for PG methods do not apply to some settings because the assumption of bounded variance of the updates is violated.

The aforementioned convergence issues are also caused by the problematic coupling between the algorithm's updates and its sampling distribution since one directly impacts the other. As a potential solution, we showed that off-policy sampling can sidestep these difficulties by ensuring we use a sampling distribution that is different than the one induced by the agent's current policy. This supports the hypothesis that on-policy learning can be problematic, as observed in previous work (Schaul et al., 2019; Hennes et al., 2020). Nevertheless, importance sampling in RL is generally seen as problematic (van Hasselt et al., 2018) due to instabilities it introduces to the learning process. Moving from an imposed policy, using past trajectories, to a chosen sampling policy reduces the variance of the gradients for near-deterministic policies and can lead to much better behaviour. In general, other variance-reduction strategies may also be more effective (Xu et al., 2019).

More broadly, this work suggests that treating bandit and reinforcement learning problems as a black-box optimization of a function $J(\theta)$ may be insufficient to perform well. As we have seen, the current parameter value can affect all future parameter values by influencing the data collection process and thus the updates performed. Theoretically, relying on immediately available quantities such as the gradient variance and ignoring the sequential nature of the optimization problem is not enough to discriminate between certain optimization algorithms. In essence, to design highly-effective policy optimization algorithms, it may be necessary to develop a better understanding of how the optimization process evolves over many steps.

## Acknowledgements

We would like to thank Kris de Asis, Alan Chan, Ofir Nachum, Doina Precup, Dale Schuurmans, and Ahmed Touati for helpful discussions. We also thank Courtney Paquette, Vincent Liu, Scott Fujimoto and Csaba Szepesvári for reviewing an earlier version of this paper. Marlos C. Machado and Nicolas Le Roux are supported by a Canada CIFAR AI Chair.

## References


Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. Optimality and approximation with policy gradient methods in markov decision processes. *arXiv preprint arXiv:1908.00261*, 2019.

Agarwal, A., Henaff, M., Kakade, S., and Sun, W. Pc-pg: Policy cover directed exploration for provable policy gradient learning. *arXiv preprint arXiv:2007.08459*, 2020.

Asadi, K. and Littman, M. L. An alternative softmax operator for reinforcement learning. In *International Conference on Machine Learning*, pp. 243–252. PMLR, 2017.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*, 2019.

Cheng, C.-A., Yan, X., and Boots, B. Trajectory-wise control variates for variance reduction in policy gradient methods. In *Conference on Robot Learning*, pp. 1379–1394, 2020.

De Boer, P.-T., Kroese, D. P., Mannor, S., and Rubinstein, R. Y. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67, 2005.

Efroni, Y., Shani, L., Rosenberg, A., and Mannor, S. Optimistic policy optimization with bandit feedback. *arXiv preprint arXiv:2002.08243*, 2020.

Grathwohl, W., Choi, D., Wu, Y., Roeder, G., and Duvenaud, D. Backpropagation through the void: Optimizing control variates for black-box gradient estimation. *arXiv preprint arXiv:1711.00123*, 2017.

Greensmith, E., Bartlett, P. L., and Baxter, J. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5(Nov): 1471–1530, 2004.

Gu, S., Lillicrap, T., Ghahramani, Z., Turner, R. E., and Levine, S. Q-prop: Sample-efficient policy gradient with an off-policy critic. *arXiv preprint arXiv:1611.02247*, 2016.

Hanna, J. P., Thomas, P. S., Stone, P., and Niekum, S. Data-efficient policy evaluation through behavior policy search. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1394–1403. JMLR. org, 2017.

Hanna, J. P., Niekum, S., and Stone, P. Importance sampling policy evaluation with an estimated behavior policy. *arXiv preprint arXiv:1806.01347*, 2018.

Harper, M. and Weinstein, B. python-ternary: Ternary plots in python. *Zenodo 10.5281/zenodo.594435*, 2015. doi: 10.5281/zenodo.594435. URL https://github.com/marcharper/python-ternary.

Hennes, D., Morrill, D., Omidshafiei, S., Munos, R., Perolat, J., Lanctot, M., Gruslys, A., Lespiau, J.-B., Parmas, P., Duéñez-Guzmán, E., et al. Neural replicator dynamics: Multiagent learning via hedging policy gradients. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 492–501, 2020.

Hofmann, T., Lucchi, A., Lacoste-Julien, S., and McWilliams, B. Variance reduced stochastic gradient descent with neighbors. In *Advances in Neural Information Processing Systems*, pp. 2305–2313, 2015.

Kahn, H. and Harris, T. E. Estimation of particle transmission by random sampling. *National Bureau of Standards applied mathematics series*, 12:27–30, 1951.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

Liu, H., Feng, Y., Mao, Y., Zhou, D., Peng, J., and Liu, Q. Action-depedent control variates for policy optimization via stein's identity. *arXiv preprint arXiv:1710.11198*, 2017.

Liu, J. S. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

Mei, J., Xiao, C., Dai, B., Li, L., Szepesvári, C., and Schuurmans, D. Escaping the gravitational pull of softmax. *Advances in Neural Information Processing Systems*, 33, 2020a.

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pp. 6820–6829. PMLR, 2020b.

Munos, R., Stepleton, T., Harutyunyan, A., and Bellemare, M. G. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 1046–1054, 2016.

Neu, G. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. *arXiv preprint arXiv:1506.03271*, 2015.

Parmas, P. and Sugiyama, M. A unified view of likelihood ratio and reparameterization gradients and an optimal importance sampling scheme. *arXiv preprint arXiv:1910.06419*, 2019.

Peters, J. and Schaal, S. Reinforcement learning of motor skills with policy gradients. *Neural networks*, 21(4):682–697, 2008.

Precup, D. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, pp. 80, 2000.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Robbins, H. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

Schaul, T., Borsa, D., Modayil, J., and Pascanu, R. Ray interference: a source of plateaus in deep reinforcement learning. *arXiv preprint arXiv:1904.11455*, 2019.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Seldin, Y., Szepesvári, C., Auer, P., and Abbasi-Yadkori, Y. Evaluation and analysis of the performance of the exp3 algorithm in stochastic environments. In *EWRL*, pp. 103–116, 2012.

Seldin, Y., Szepesvári, C., Auer, P., and Abbasi-Yadkori, Y. Evaluation and analysis of the performance of the exp3 algorithm in stochastic environments. In *European Workshop on Reinforcement Learning*, pp. 103–116. PMLR, 2013.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018.

van Hasselt, H., Doron, Y., Strub, F., Hessel, M., Sonnerat, N., and Modayil, J. Deep reinforcement learning and the deadly triad. *CoRR*, abs/1812.02648, 2018.

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

Wu, C., Rajeswaran, A., Duan, Y., Kumar, V., Bayen, A. M., Kakade, S., Mordatch, I., and Abbeel, P. Variance reduction for policy gradient with action-dependent factorized baselines. *arXiv preprint arXiv:1803.07246*, 2018.

Xu, P., Gao, F., and Gu, Q. Sample efficient policy gradient methods with recursive variance reduction. *arXiv preprint arXiv:1909.08610*, 2019.