Posterior Value Functions: Hindsight Baselines for Policy Gradient Methods

Chris Nota [1] **Bruno Castro da Silva** [1] **Philip S. Thomas** [1]

Abstract

Hindsight allows reinforcement learning agents to leverage new observations to make inferences about earlier states and transitions. In this paper, we exploit the idea of hindsight and introduce posterior value functions. Posterior value functions are computed by inferring the posterior distribution over hidden components of the state in previous timesteps and can be used to construct novel unbiased baselines for policy gradient methods. Importantly, we prove that these baselines reduce (and never increase) the variance of policy gradient estimators compared to traditional state value functions. While the posterior value function is motivated by partial observability, we extend these results to arbitrary stochastic MDPs by showing that hindsight-capable agents can model stochasticity in the environment as a special case of partial observability. Finally, we introduce a pair of methods for learning posterior value functions and prove their convergence.

1. Introduction

Reinforcement learning (RL) is an area of artificial intelligence in which computational agents learn to act in complex environments through reward and punishment. Many RL agents work by choosing a sequence of actions and comparing the resulting outcome (in terms of rewards) to some prior expected outcome (called a *baseline*). Such agents then alter their behavior to make better-than-expected outcomes more likely and worse-than-expected ones less so. To avoid bias, baselines are usually computed in a way that carefully avoids incorporating any information about the actual outcome, including anything the agent learns about the environment after choosing an action. However, in many cases, such information can be useful for assessing which outcomes were likely to have occurred, and failing to use it can mislead the agent.

For example, consider a morning commute. When you get into your car, you may expect the drive to take a certain length of time. This morning, however, you run into traffic on the freeway and discover you will be substantially delayed in getting to work. If you were an RL agent using traditional methods, you would respond to this worse-than-expected outcome by reducing the likelihood of all of the "actions" that you took prior to encountering the traffic. For example, you would decrease the probability of using your turn indicator when approaching the on-ramp. As a human, you know this is foolish—the traffic was not caused by using the turn indicator or any other actions you took while driving (presuming there was no alternate route available!). Unless you chose to leave earlier, arriving to work late was inevitable, and your past driving should be evaluated *relative* to this new information (the presence of traffic), not according to the assumptions you made while sitting in your driveway. The ability to use such information to better understand earlier circumstances is known as *hindsight*.

We present a new class of baselines, called *posterior value functions*, which exploit the idea of hindsight. We prove that posterior value functions can reduce the variance of updates to an agent's behavior (i.e., *policy gradient estimators*) compared to standard approaches, while never increasing variance or introducing bias. We analyze the resulting variance in detail and show using several examples that the improvement can be large. We introduce efficient methods for learning posterior value functions through interactions with the environment and prove that they converge almost surely. Finally, we illustrate these results empirically.

2. Background and Notation

In RL, the environment is typically represented as a *Markov decision process* (MDP). An MDP is a tuple, $(\mathcal{S}, \mathcal{A}, P, R, d_0)$, where $\mathcal{S}$ is the set of possible states, $\mathcal{A}$ is the set of possible actions, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition function, and $d_0 : \mathcal{S} \to [0, 1]$ is the initial state distribution.

[1]College of Information and Computer Science, University of Massachusetts, Amherst, MA. Correspondence to: Chris Nota <cnota@cs.umass.edu>, Philip S. Thomas <pthomas@cs.umass.edu>.

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

The MDP framework can be extended to encompass the problem of *partial observability*, in which not all parts of $\mathcal{S}$ are visible to the agent. An MDP with partial observability is called a *partially observable Markov decision process* (POMDP). We break the state set into two components, $\mathcal{S} = \mathcal{O} \times \mathcal{Z}$, where $\mathcal{O}$ is the set of *observations* and $\mathcal{Z}$ is the set of *hidden components* of the state. The difference between a POMDP and an MDP is that in a POMDP, the agent may only consider the observable component. This formulation of POMDPs differs slightly from the standard formulation (Puterman, 2014), but is of equivalent descriptive power.[1]

Interactions with the environment are broken into discrete sequences called *episodes*. Each episode is further divided into integer timesteps in the range $[0, \infty)$. The state at each time t is denoted by the random variable S_t, with observation O_t and hidden component Z_t; that is, $S_t = (O_t, Z_t)$. Similarly, the action at time t is denoted by A_t and the reward by R_t. Each episode begins at timestep $t = 0$. An initial state is sampled, $S_0 \sim d_0(\cdot)$. The agent then observes O_t and selects some A_t. Finally, the next state, $S_{t+1} \sim P(S_t, A_t, \cdot)$, and reward, R_t, is generated by the environment such that $\mathbb{E}[R_t | S_t, A_t] = R(S_t, A_t)$. The episode terminates when the agent enters a special state, s_∞, called the *terminal absorbing state*. The *history*, H, for a given episode describes the complete sequence of interactions *as observed by the agent*, i.e., $H := (O_0, A_0, R_0, O_1, A_1, R_1, \ldots, s_\infty)$. A *partial history*, H_t, describes the sequence of interactions until a given timestep, including the current observation. That is, for $t > 0$, $H_t := (O_0, A_0, R_0, \ldots, O_t)$, and $H_0 := (O_0)$.

The *return*, $G_t = \sum_{k=0}^{\infty} R_{t+k}$, is the sum of rewards starting at a given timestep. We note that if for all timesteps, $R_t \in [-R_{\max}, R_{\max}]$ for some finite constant $R_{\max}$, and the probability of the episode ending satisfies $\sum_{t=0}^{\infty} \Pr(S_t \neq s_\infty) < \infty$, then $\Pr(G_t < \infty) = 1$. In this paper, we always assume that these conditions hold. One special case of the above is the finite horizon setting, in which there exists some T such that for all $t > T$, $\Pr(S_t = s_\infty) = 1$.

2.1. The Policy Gradient Theorem

One way of selecting A_t is by sampling from a stochastic *policy*, $\pi_\theta : \mathcal{O} \times \mathcal{A} \to [0, 1]$, where θ is a parameter vector, e.g., the weights of a neural network, such that $\pi_\theta(O_t, A_t) = \Pr(A_t | O_t, \theta)$. From here on, it should be assumed that all random variables, expectations, and variance expressions are conditioned on θ unless otherwise specified. The goal of the agent is to find the parameter vector θ that maximizes the *objective function*, $J(\theta) := \mathbb{E}[G_0]$. *Policy gradient methods* are a class of methods for maximizing $J(\theta)$. The *policy gradient theorem* (Sutton et al., 1999) gives an ascent direction for θ:

$$\nabla J(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} G_t \frac{\partial \ln \pi_\theta(O_t, A_t)}{\partial \theta}\right]. \tag{1}$$

$\nabla J(\theta)$ can be estimated by executing π_θ for an entire episode and computing the inner expression of (1). We can then update θ using stochastic gradient descent:

$$\theta \leftarrow \theta + \alpha_k \sum_{t=0}^{\infty} G_t \frac{\partial \ln \pi_\theta(O_t, A_t)}{\partial \theta}, \tag{2}$$

where α_k is the k^{th} element in a sequence of non-negative step sizes. For a properly decaying sequence, this update provides the standard convergence guarantees of stochastic gradient descent (Bertsekas & Tsitsiklis, 2000). These results can be extended to the case where O_t is replaced with a learned representation, e.g., the output of a recurrent neural network (Wierstra et al., 2010).

2.2. Variance of the Sample Policy Gradient

The quantity $G_t \frac{\partial \ln \pi_\theta(O_t, A_t)}{\partial \theta}$ is known as the *sample gradient*. Notice that the sample gradient is a random vector. We consider the following scalar notion of variance:

$$\mathrm{Var}(\vec{X}) := \mathbb{E}\big[(\vec{X} - \mathbb{E}[\vec{X}])^\top (\vec{X} - \mathbb{E}[\vec{X}])\big]. \tag{3}$$

This quantity is recognizable as the trace of the covariance matrix, and may also be understood as the expected value of the square of the Euclidean distance between $\vec{X}$ and $\mathbb{E}[\vec{X}]$. Similarly, we define the conditional variance of $\vec{X}$ given some other random variable, Y, to be:

$$\mathrm{Var}(\vec{X}|Y) := \mathbb{E}\big[(\vec{X} - \mathbb{E}[\vec{X}|Y])^\top (\vec{X} - \mathbb{E}[\vec{X}|Y]) \big| Y\big]. \tag{4}$$

Finally, the covariance of $\vec{X}$ and $\vec{Y}$ is:

$$\mathrm{Cov}(\vec{X}, \vec{Y}) := \mathbb{E}\big[(\vec{X} - \mathbb{E}[\vec{X}])^\top (\vec{Y} - \mathbb{E}[\vec{Y}])\big]. \tag{5}$$

2.3. Baselines in Policy Gradient Methods

Due to stochasticity in both the environment and the policy, the variance of (2) can be large, resulting in slow convergence. One way of reducing that variance is by introducing a *baseline* (also known as a *control variate*), which we write as B_t, resulting in the update:

$$\theta \leftarrow \theta + \alpha \sum_{t=0}^{\infty} (G_t - B_t) \frac{\partial \ln \pi_\theta(O_t, A_t)}{\partial \theta}. \tag{6}$$

The quantity $G_t - B_t$ is sometimes known as an *advantage estimator*, as it estimates the "advantage" of choosing action A_t and receiving G_t compared to the baseline. If B_t is sufficiently correlated with G_t, then the variance of the

[1] See the supplemental materials for more details.

resulting update is decreased compared to (2). Specifically, the total variance is reduced if and only if:

$$2\text{Cov}\left(G_t\frac{\partial \ln \pi_\theta(O_t,A_t)}{\partial\theta}, B_t\frac{\partial \ln \pi_\theta(O_t,A_t)}{\partial\theta}\right) > \text{Var}\left(G_t\frac{\partial \ln \pi_\theta(O_t,A_t)}{\partial\theta}\right) + \text{Var}\left(B_t\frac{\partial \ln \pi_\theta(O_t,A_t)}{\partial\theta}\right).$$

However, B_t must be chosen carefully to avoid introducing bias. The most common choice of B_t is the *state value function*, $v^\theta(S_t) \coloneqq \mathbb{E}[G_t|S_t]$, which often provides a significant amount of variance reduction without introducing bias. In general, it can be shown that using *any* function of S_t (and any other information available to the agent prior to time t) as a baseline will never introduce bias (Williams, 1992; Baxter & Bartlett, 2001).

In the partially observable setting, S_t is not known by the agent, so the "true" state value function is often replaced with what we refer to as the *prior value function*, $v^\theta(H_t) \coloneqq \mathbb{E}[G_t|H_t]$, which is an agent's best estimate of the future return based on the partial history H_t. In many cases, this is approximated by the *observation value function*, $v^\theta(O_t) \coloneqq \mathbb{E}[G_t|O_t]$, which is likewise an unbiased baseline. If O_t is a representation learned by a recurrent neural network, it may be the case that $v^\theta(O_t) \approx v^\theta(H_t)$. However, note that H_t contains O_t, so $v^\theta(H_t)$ is strictly more informative. A glossary of value functions can be found in Table 1.

2.4. Future-Dependent Baselines

Variance can in some cases be further reduced using a *future-dependent* baseline, which is any baseline incorporating A_t, R_t, S_{t+1}, or any other information available after time t. However, future-dependent baselines must be constructed carefully to avoid introducing bias. For example, consider the baseline $B_t = G_t$. The variance of the resulting update is zero, but only because the update (i.e., (6)) is always the zero vector. Because in most cases $\nabla J(\theta) \neq 0$, this baseline is almost always extremely biased and is never useful, in spite of its low (zero) variance.

Action-dependent baselines are a class of unbiased future-dependent baselines that depend on both S_t and A_t; various forms have been proposed (Gu et al., 2016; Thomas & Brunskill, 2017; Liu et al., 2017; Wu et al., 2018). While these methods showed initial promise, they were later argued to be a "mirage" (Tucker et al., 2018) offering little benefit over state-dependent baselines. One reason cited by Tucker et al. (2018) is that, empirically, A_t usually only accounts for a small portion of $\mathbb{E}[\text{Var}(G_t|S_t)]$; that is, in most cases:

$$\mathbb{E}[\text{Var}(G_t|S_t)] \gg \mathbb{E}[\text{Var}(\mathbb{E}[G_t|S_t,A_t]|S_t)]. \quad (7)$$

The new future-dependent baselines that we introduce next depend on the entire history, H, allowing them to achieve large variance reduction compared to existing baselines.

Function	Value	Name
$v^\theta(O_t)$	$\mathbb{E}[G_t\|O_t]$	Observation Value Function
$v^\theta(H_t)$	$\mathbb{E}[G_t\|H_t]$	Prior Value Function
$v^\theta(S_t)$	$\mathbb{E}[G_t\|S_t]$	State Value Function
$u_t^\theta(H)$	$\mathbb{E}[v^\theta(S_t)\|H]$	Posterior Value Function

Table 1. Glossary of value functions.

3. The Posterior Value Function

We begin by considering the partially observable setting. Notice that in this setting, the state value function cannot be used as a baseline in practice because the state contains hidden components. However, an agent may be able to make inferences about these hidden components as an episode progresses. At the end of an episode, suppose we consider an agent's "best guess" of the state value function at an earlier timestep, t, based on everything it has seen. We refer to the resulting quantity as the *posterior value function*:

$$u_t^\theta(H) \coloneqq \mathbb{E}\big[v^\theta(S_t)|H\big]. \quad (8)$$

Recall that S_t comprises both the observable and hidden components, (O_t, Z_t), whereas H contains only the observable components, actions, and rewards. Therefore, when conditioning on H, O_t is known but Z_t is uncertain. Therefore, another way of writing the posterior value function is:

$$u_t^\theta(H) = \sum_{z\in\mathcal{Z}} \Pr(Z_t = z|H)v^\theta(O_t, z). \quad (9)$$

We call this quantity the "posterior value function" due its dependence on the posterior distribution over Z_t given H. We can immediately prove the resulting advantage estimator has generally lower variance than the standard estimators:

Theorem 1. *For all POMDPs and all timesteps, t:*

$$\text{Var}(G_t - u_t^\theta(H)) \leq \text{Var}(G_t - v^\theta(S_t)) \leq \text{Var}(G_t - v^\theta(H_t)) \leq \text{Var}(G_t - v^\theta(O_t))$$

Proof. See the supplemental material. □

To ground the posterior value function and its variance reduction properties in a concrete example, we elaborate on the traffic example discussed earlier. The traffic problem can be modeled as the five-state POMDP shown in Figure 1. Each state is labeled with an observable component, o, and a hidden component, z. The agent always starts at `Home`, however, there is a $1/2$ probability that, unbeknownst to the agent, there is traffic on the road. The traffic is not directly observable at `Home`, so it is part of the hidden state at `Home`, but part of the observed state while on the `Road`. If there is traffic, the agent arrives to work late and receives a reward

Baseline	Variance	Reduction
0	$\mathrm{Var}(G_t)$	–
$v^\theta(O_t)$	$\mathbb{E}[\mathrm{Var}(G_t\|O_t)]$	$\mathrm{Var}(v^\theta(O_t))$
$v^\theta(H_t)$	$\mathbb{E}[\mathrm{Var}(G_t\|H_t)]$	$\mathbb{E}[\mathrm{Var}(v^\theta(H_t)\|O_t)]$
$v^\theta(S_t)$	$\mathbb{E}[\mathrm{Var}(G_t\|S_t)]$	$\mathbb{E}[\mathrm{Var}(v^\theta(S_t)\|H_t)]$
$u_t^\theta(H)$	$\mathrm{Var}(\mathbb{E}[G_t - v^\theta(S_t)\|H])$	$\mathbb{E}[\mathrm{Var}(v^\theta(S_t)\|H)]$

Table 2. The variance and variance reduction in the advantage estimator, $G_t - B_t$, achieved by each baseline. The third column indicates the difference between the variance of the estimator on the previous row and the current row, from top to bottom. Proof of these results can be found in the supplemental material.

of -1. If the roads are clear, the agent arrives on time and receives a reward of $+1$. The agent does not have any choice of action; it must go to work and only one route is available. We are sure many readers can sympathize.

First, consider the variance of the advantage estimator derived from the prior value function, $v^\theta(H_t)$. The partial history in the initial state, H_0, will contain only the initial observation, $O_0 = \texttt{Home}$, so $v^\theta(H_0) = v^\theta(\texttt{Home}) = 0$, always. Therefore, $\mathrm{Var}(G_0 - v^\theta(H_0)) = \mathrm{Var}(G_0) = 1$.

Next, consider the advantage estimator derived from the posterior value function, which we call the *posterior advantage estimator*. There are only two possible values of H, one which contains traffic, which we call h_{tr}, and one in which the road is clear, which we call h_{cl}. Given h_{tr}, we know that $Z_0 = \texttt{Traffic}$ with probability 1, and given h_{cl}, we know that $Z_0 = \texttt{Clear}$, therefore:

$$u_0^\theta(h_{\text{tr}}) = v^\theta\big(\texttt{Home, Traffic}\big) = -1$$
$$u_0^\theta(h_{\text{cl}}) = v^\theta\big(\texttt{Home, Clear}\big) = 1.$$

$\mathbb{E}[G_0 - u_0^\theta(H)] = 0$, so the variance of the posterior advantage estimator at time 0 is:

$$\begin{aligned}\mathrm{Var}(G_0 - u_0^\theta) =& \mathbb{E}\Big[\big(G_0 - u_0^\theta(H)\big)^2\Big] \\ =& \Pr(H = h_{\text{tr}})(-1 - v_0^\theta(h_{\text{tr}})) \\ &+ \Pr(H = h_{\text{cl}})(1 - v_0^\theta(h_{\text{cl}})) \\ =& \frac{1}{2}((-1) - (-1))^2 + \frac{1}{2}(1-1)^2 = 0.\end{aligned}$$

The posterior advantage estimator achieves a variance of 0. Additional details can be found in Figure 3. While it is easy to show that the variance of the posterior advantage estimator is 0 for this particular POMDP, the variance is not zero for all POMDPs. Nevertheless, Theorem 1 shows that variance is always less than or equal to that of standard advantage estimators. Further comparisons of the variance of these estimators are given in Table 2.

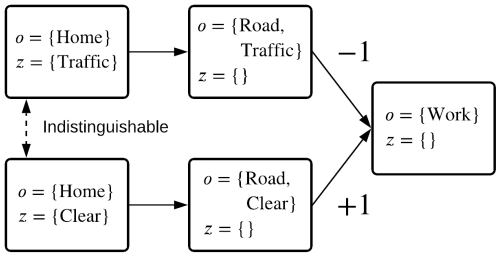

Figure 1. POMDP representation of the traffic example.

Trajectory	G_0	$u_0^\theta(H)$	$v^\theta(H_0)$	G_1	$u^\theta(H)$	$v_1^\theta(H_1)$
Traffic	-1	-1	0	-1	-1	-1
Clear	1	1	0	1	1	1

Table 3. Comparison between the return G_t, the posterior value function $u_t^\theta(H)$, and the prior value function $v^\theta(H_t)$ at each time step and for each possible trajectory in the traffic POMDP presented in Figure 1.

3.1. Comparison to the State Value Function

In the traffic example, the posterior advantage estimator achieved lower variance than the advantage estimator derived from the prior value function (the *prior advantage estimator*). This is an important comparison because in POMDPs the state value function is not computable as the agent does not know the state's hidden components, making the prior value function the next best option. However, in this section, we go further and discuss how the posterior value function can produce a lower variance estimator than even the state value function. From Table 2, we know:

$$\mathrm{Var}\big(G_t - u_t^\theta(H)\big) - \mathrm{Var}\big(G_t - v^\theta(S_t)\big) = \mathbb{E}[\mathrm{Var}(v^\theta(S_t)|H)].$$

This expression tells us that if there are multiple possible values of Z_t given H, and these values correspond to different values of $v^\theta(S_t)$, then this uncertainty in Z_t will contribute to $\mathrm{Var}(G_t - v^\theta(S_t))$. In comparison, the posterior value function averages over these different values of Z_t, eliminating this source of variance.

This particular source of variance did not appear in the traffic example. However, in Figure 2, we provide an example of a POMDP where it figures prominently. In this example, the agent again has no choice of action. The agent begins in a state with observation o_0 and no hidden component, and then transitions to a state with observation o_1 with a hidden component of either z_1 or z_2, each with 50% probability. Finally, the agent transitions to one of three terminal states and receives a reward of -100, 0, or $+100$. If the hidden component is z_1, it receives a reward of -100 or 0 with a 50% chance each, and if the component is z_2, it receives a reward of 0 or 100 with a 50% chance each. The value

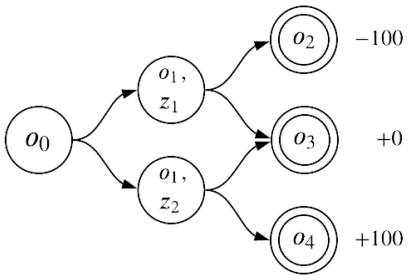

Figure 2. A POMDP wherein the posterior value function decreases the variance of the advantage estimator relative to the state value function.

Trajectory	G_0	$u_0^\theta(H)$	$v^\theta(S_0)$	$v_0^\theta(H_0)$
o_0, o_1, z_1, o_2	-100	0	0	0
o_0, o_1, z_1, o_3	0	0	0	0
o_0, o_1, z_2, o_3	0	0	0	0
o_0, o_1, z_2, o_4	100	0	0	0
Trajectory	G_1	$u_1^\theta(H)$	$v^\theta(S_1)$	$v_1^\theta(H_1)$
o_0, o_1, z_1, o_2	-100	-50	-50	0
o_0, o_1, z_1, o_3	0	0	-50	0
o_0, o_1, z_2, o_3	0	0	50	0
o_0, o_1, z_2, o_4	100	50	50	0

Table 4. Comparison of baselines for POMDP in Figure 2.

estimates for the posterior, state, and prior value functions for each possible trajectory are given by Table 4.

The key in this example is that the agent is only able to uniquely identify the state entered at $t = 1$ when a reward of -100 or $+100$ is received. In these cases, the posterior value function is equivalent to the state value function and produces the same conditional variance, with an absolute error of 50. However, consider the history when a reward of 0 is received: the agent is not able to determine whether the hidden component is z_1 or z_2. For time $t = 1$, the posterior value function averages over these two possibilities and produces an estimate of 0, which in this case is exactly the return G_1, resulting in a conditional variance of 0. The state value function, however, will produce an estimate of either -50 or 50, depending on the value of the hidden component, significantly deviating from the true return in either case. In Section 4.1, we will show how it is possible to lower the variance even further, resulting in a total variance of 0.

4. The Posterior Policy Gradient Estimator

If we use u_t^θ as a baseline in the policy gradient theorem, we are given what we refer to as the *posterior policy gradient* estimator. We begin by proving that the expected value of this expression is the policy gradient, ∇J, i.e., that the $u_t^\theta(H)$ baseline does not introduce bias.

Theorem 2. ***(Posterior policy gradient):*** *For any POMDP:*

$$\nabla J(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \big(G_t - u_t^\theta(H)\big) \frac{\partial \ln \pi_\theta(O_t, A_t)}{\partial \theta}\right].$$

Proof. See the supplemental material. □

Consider that the posterior value function is an expectation over the state value function, i.e., $\mathbb{E}[v^\theta(S_t)|H]$. Therefore, the only information the agent uses when constructing it is regarding S_t. It is well known that any function of S_t can be used as a baseline without introducing bias (Williams, 1992). In fact, the proof of Theorem 2, after a few transformations, reduces to the proof that the $v^\theta(S_t)$ baseline does not introduce bias. We encourage the reader to study the full proof of Theorem 2 in the supplemental material. Recall that the goal of incorporating hindsight into our baseline is to reduce the variance of the policy gradient estimator. Theorem 3 shows the posterior value function achieves this:

Theorem 3. *For all POMDPs and all timesteps, t:*

$$\begin{aligned}&\mathrm{Var}\left(\big(G_t - u_t^\theta(H)\big)\frac{\partial \ln \pi_\theta(O_t, A_t)}{\partial \theta}\right)\\ \leq &\mathrm{Var}\left(\big(G_t - v^\theta(S_t)\big)\frac{\partial \ln \pi_\theta(O_t, A_t)}{\partial \theta}\right).\end{aligned}$$

Proof. See the supplemental material. □

Thus, we have achieved the main results. In Section 6, we demonstrate the above variance reduction empirically. However, first we ask, can we take it even further?

4.1. The Off-POMDP Estimator

Are different ways of looking at the world equally valid? If they are equally valid, are they equally "good" from the perspective of an RL agent? In Section 3, we introduced the posterior value function, and in Section 4, we showed how it could be used as a baseline to reduce the variance of the policy gradient estimator. In this section, we show in Theorem 4 that there is not one but *many* posterior value functions that serve as unbiased baselines for any POMDP. Further, we find that *any* such posterior value function reduces the variance of the resulting advantage estimator compared to the prior value function.

This has the important implication that, in practice, an agent need not learn the posterior of the "true" POMDP, but need only learn the posterior of some POMDP that satisfies the conditions laid out in Theorem 4. This greatly improves the practicality of learning a suitable posterior value function,

because the agent does not need to model the true posterior, which may require prior knowledge, but rather need only learn some sufficient model of the posterior. We have:

Theorem 4. *Consider two POMDPs, $M = (\mathcal{S}, \mathcal{A}, P, R, d_0)$ and $\tilde{M} = (\tilde{\mathcal{S}}, \mathcal{A}, \tilde{P}, \tilde{R}, \tilde{d}_0)$, such that $\mathcal{S} = \mathcal{O} \times \mathcal{Z}$ and $\tilde{\mathcal{S}} = \mathcal{O} \times \tilde{\mathcal{Z}}$, and where the observations, actions, rewards, and histories in $\tilde{M}$ are given by $\tilde{O}_t, \tilde{A}_t, \tilde{R}_t$, and $\tilde{H}_t$ respectively. If for all observations o in $\mathcal{O}$:*

$$\Pr(O_0 = o) = \Pr(\tilde{O}_0 = o) \tag{10}$$

and for all partial histories $h = (o_0, a_0, r_0, \ldots, o_t)$, $h' = (o_0, a_0, r_0, \ldots, o_{t+1})$, and actions a in $\mathcal{A}$:

$$\begin{aligned} &\Pr(H_{t+1} = h' | H_t = h, A_t = a) \\ &= \Pr(\tilde{H}_{t+1} = h' | \tilde{H}_t = h, \tilde{A}_t = a), \end{aligned} \tag{11}$$

then:

$$\nabla J(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} (G_t - \tilde{u}_t^\theta(H)) \frac{\partial \ln \pi_\theta(O_t, A_t)}{\partial \theta} \middle| \theta\right], \tag{12}$$

where J is the objective for M and $\tilde{u}_t^\theta$ is the posterior value function for $\tilde{M}$.

Proof. See the supplemental material. □

The critical insight in Theorem 4 is that there is a class of POMDPs satisfying the given conditions, and that from the perspective of an agent, these POMDPs are indistinguishable from the true POMDP. Informally, the proof is as follows: Let $\tilde{J}$ be the objective for $\tilde{M}$. Because any two POMDPs satisfying 10 and 11 are indistinguishable, $J(\theta)$ and $\tilde{J}(\theta)$ are equivalent. Therefore, because $\tilde{u}_t^\theta(H)$ is an unbiased baseline for $\nabla \tilde{J}(\theta)$, it is also an unbiased baseline for $\nabla J(\theta)$.

Similarly, $v^\theta(H_t) = \tilde{v}^\theta(H_t)$, so the variance reduction property from Theorem 3 holds transitively. That is:

Theorem 5. *Let M be a fully observable MDP, and let $\tilde{M} = (\tilde{\mathcal{S}}, \mathcal{A}, \tilde{P}, \tilde{R}, \tilde{d}_0)$ be any POMDP satisfying (10) and (11). Then, for all t:*

$$\mathrm{Var}(G_t - \tilde{u}_t^\theta(H)) \leq \mathrm{Var}(G_t - v^\theta(H_t)).$$

Proof. See the supplemental material. □

4.2. The Fully Observable Setting

To this point, we focused primarily on how posterior value functions could be used to reduce variance resulting from partial observability. How could it be useful in fully observable settings, where the state has no hidden component? The key observation is that Theorem 4 can be used to draw a relationship between partial observability and stochasticity:

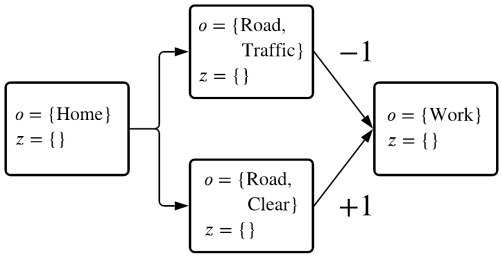

Figure 3. A fully observable version of the traffic problem from Figure 1 satisfying Theorem 4. The hidden component of the state is replaced with a stochastic transition.

Trajectory	G_0	$u_0^\theta(H)$	$v^\theta(H_0)$	G_1	$u^\theta(H)$	$v_1^\theta(H_1)$
Traffic	−1	0	0	−1	−1	−1
Clear	1	0	0	1	1	1

Table 5. Comparison between the return G_t, the posterior value function $u_t^\theta(H)$, and the prior value function $v^\theta(H_t)$ at each time step and for each possible trajectory in the traffic POMDP presented in Figure 3. Note that unlike in Figure 3, the posterior value function offers no improvement over the prior value function.

from the perspective of the agent, partial observability may be modeled as stochasticity and vice versa without introducing bias. By choosing to model stochastic events as a special case of partial observability, the agent may apply hindsight to produce a variance-lowering posterior value function.

Consider again Figure 1. In order to reduce the variance of the advantage estimator, we modeled the traffic as a hidden component of the state that was present even when the agent was sitting at `Home`. Consider how the traffic problem can be alternately modeled as a fully observable problem while satisfying Theorem 4: Instead of treating the traffic as a hidden component of the state, we can treat the transition from `Home` to `Road` as a stochastic transition, as shown in Figure 3. First, we remove the hidden component from the initial states, leaving just one initial state, `Home`. Then, we add transitions from `Home` to `(Road, Traffic)` and `(Road, Clear)`, each with a transition probability of 50%. It is easily seen that the conditions given by (10) and (11) are satisfied by the resulting MDP. However, notice that because the hidden state at time 0 is empty, we now have $u_0^\theta(H) = 0$ in all cases. Because our reward is either 1 or -1, the variance of the posterior advantage estimator at time 0 is now 1, whereas previously it was 0. The agent was better served by modeling the "random" appearance of traffic as partially observable.

Many types of stochasticity can be modeled this way: the agent may treat the outcome of a seemingly random dice roll or coin flip as preordained and will be rewarded with a lower variance advantage estimator. We show in Corollary

1 that, from this perspective, the fully observable stochastic formulation is always the worst possible formulation, and any valid hindsight model is at least as good.

Corollary 1. *Let M be a fully observable MDP, and let $\tilde{M} = (\tilde{\mathcal{S}}, \mathcal{A}, \tilde{P}, \tilde{R}, \tilde{d}_0)$ be any POMDP satisfying (10) and (11) for M. Then, for all t:*

$$\mathrm{Var}(G_t - \tilde{u}_t^\theta(H)) \leq \mathrm{Var}(G_t - u_t^\theta(H)).$$

Proof. In the fully observable setting, we have:

$$u_t^\theta(H) = v^\theta(S_t) = v^\theta(H_t),$$

by the Markov property. Therefore, the statement holds by Theorem 5. □

In general, we suggest that when learning a hindsight model, an agent should attempt to minimize the stochasticity of the resulting model and account for as much of it as possible using partial observability in order to maximize the benefits of posterior value functions.

5. Learning Posterior Value Functions

In this section, we present methods for learning posterior value functions from data. We assume that the agent possesses a hindsight model for each timestep, $q_t(z|H) := \Pr(Z_t = z|H)$, that satisfies (10) and (11). One way of estimating the posterior value function is to learn an estimate, $\hat{v}(O_t, Z_t)$, of the state value function.[2] We can then form an estimate, $\hat{u}_t(H)$, of $u_t^\theta(H)$ as

$$\hat{u}_t(H) = \sum_{z \in \mathcal{Z}} q_t(z|H)\hat{v}(O_t, z). \tag{13}$$

The problem is then reduced to learning $\hat{v}$. However, we cannot use standard methods for this as we still cannot observe Z_t directly. Instead, we need to leverage our hindsight model to perform weighted updates for all values of Z_t with non-zero probability conditioned on H. Consider a sequence, $(\hat{v}_0, \hat{v}_1, \dots)$, such that $\hat{v}_i$ is the estimate of the value function at the beginning of the i^{th} training episode. At the end of each episode, we update for all $z \in \mathcal{Z}$:

$$\hat{v}_{i+1}\big(O_t, z\big) = \hat{v}_i\big(O_t, z\big) + \alpha_i q_t\big(z|H\big)\big(G_t - \hat{v}_i(O_t, z)\big), \tag{14}$$

where α_i is a positive learning rate satisfying $\sum_{i=0}^\infty \alpha_i = \infty$ and $\sum_{i=0}^\infty \alpha_i^2 < \infty$, and all random variables are sampled from the i^{th} episode. We then have:

Theorem 6. *For any POMDP, let $\hat{u}_{t,i}(H) := \hat{\sum}_z q_t(z|H)\hat{v}_i(O_t, z)$, where v_i is the i^{th} element in the sequence defined by (14). Then, for any history h such that $\Pr(H = h|\theta) > 0$:*

$$\Pr\Big(\lim_{i\to\infty} \hat{u}_{t,i}(h) = u_t^\theta(h)\Big) = 1.$$

[2] Recall that the state space is $\mathcal{S} = \mathcal{O} \times \mathcal{Z}$.

Proof. See the supplemental material. □

Computing the above expression is tractable where $|\mathcal{Z}|$ is small. However, as the update scales with $O(|\mathcal{Z}|)$, if $|\mathcal{Z}|$ is very large, it may quickly become infeasible. In these cases, we instead consider computing some sufficient statistic, ϕ_t, for Z_t given H. For such a statistic, trivially, $\mathbb{E}[v^\theta(S_t)|\phi_t] = E[v^\theta(S_t)|H]$. Further, we assume that we are able to generate samples, $\tilde{Z}$, such that for all z, $\Pr(\tilde{Z} = z|\phi_t) = \Pr(Z = z|\phi_t)$.

We may then perform the following update once per episode:

$$\hat{v}_{i+1}\big(O_t, \tilde{Z}_t\big) = \hat{v}_i\big(O_t, \tilde{Z}_t\big) + \alpha_i\big(G_t - \hat{v}_i(O_t, \tilde{Z}_t)\big) \tag{15}$$

$$\hat{u}_{i+1}\big(O_t, \phi_t\big) = \hat{u}_i\big(O_t, \phi_t\big) + \beta_i\big(\hat{v}_i(O_t, \tilde{Z}_t) - \hat{u}_i(O_t, \phi_t)\big), \tag{16}$$

where the above conditions for learning rates α_i and β_i are satisfied and additionally, $\lim_{i\to\infty} \frac{\alpha_i}{\beta_i} = 0$. The complexity of this more efficient update is $O(1)$, making it tractable even for large $|\mathcal{Z}|$. Finally, we have:

Theorem 7. *For any POMDP, let $\hat{u}_i$ be the i^{th} element in the sequence defined by (16). Then:*

$$\Pr\Big(\lim_{i\to\infty} \hat{u}_i(O_t, \phi_t) = u_t^\theta(H)\Big) = 1.$$

Proof. See the supplemental material. □

Both of the above techniques, with small modifications, can be implemented using function approximation. However, as with other techniques, unless the network is overparameterized (Allen-Zhu et al., 2019), they may not converge to exact solutions.

6. Experiments

We further illustrate the variance reduction properties of posterior value functions on a tabular gridworld domain with partial observability. We compared agents using learned estimates of the posterior, prior, and observation value functions as baselines for the policy gradient theorem. The agents used either a "reactive" policy, which considered only the current position on the grid, or a "belief" policy, which conditioned the action on an analytically computed belief distribution based on the partial history. The policies were trained using the standard REINFORCE with baselines algorithms (Williams, 1992). The posterior value function was learned using Equation 14. The full experimental details are in the supplemental material.

The environment is a composition of the four 5×5 gridworlds shown in Figure 5. The agent always starts in the top left square, $\{0, 0\}$, and the episode terminates upon reaching the bottom right square, $\{4, 4\}$. At the beginning of

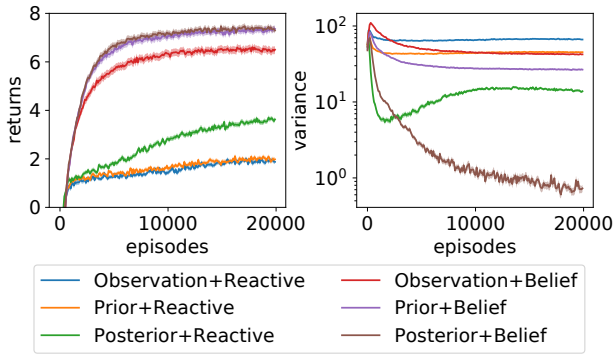

Figure 4. The performance of six different REINFORCE agents on a 5×5 tabular gridworld with partial observability, in terms of the mean episode returns and the estimated mean squared error of the advantage estimator, $\mathbb{E}[(G_t - B_t)^2]$. The results were averaged over approximately 300 trials per agent. The shaded regions indicate the standard error.

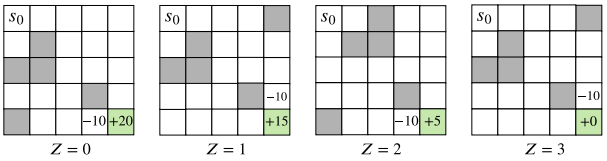

Figure 5. Depiction of the four gridworlds in which the agent could find itself. White tiles are passable. Gray tiles are impassable walls. The agent begins in the tile marked by s_0 and terminates in the green tile. Rewards are zero except where noted.

each episode, the agent is placed randomly in one of the four gridworlds with equal probability. The agent can move up, down, left, or right, moving one square at at time and transitions are fully deterministic. The agent must explore in order to deduce which gridworld it is in during a given episode. The maximum reward varies, so the value of early states can only be determined using hindsight.

We ran each of the six agents for twenty thousand episodes on the partially observable gridworld. We compared the returns and estimated the variance of the advantage function for each agent. The variance was estimated by the sample mean squared error for the advantage estimator, $\mathbb{E}[(G_t - B_t)^2]$. Results during individual trials were averaged over 100 episode chunks. The results were then averaged over 300 trials and are shown in Figure 4.

The agents using the posterior value function baselines produced drastically lower variance advantage estimates than the agents using the prior or observation value functions, illustrating that the variance reduction achieved due to Theorem 1 can be non-trivial. However, the agents using the belief policy trained with the prior and posterior value function baselines achieved essentially equal returns.

7. Discussion and Future Work

The results derived in this paper suggest that posterior methods possess fundamental advantages over existing methods in a broad range of settings. We showed that in many settings, it is not possible to generate low-variance advantage estimates without the benefit of hindsight. We introduced an approach for doing so which is both intuitive and well-principled. However, we refrained from proposing a specific end-to-end RL algorithm in order to focus on the more fundamental insights discussed herein. Translating these theoretical insights to empirical benefits is left as future work.

One of the most interesting connections uncovered by these results is the deep connection between RL and inference. In many practical settings, estimating and learning the posterior value function requires learning a hindsight inference model, $q(z|H)$. The study of methods for finding approximate solutions for such problems is a subfield of machine learning called *variational inference* (VI). While application of VI to hindsight inference is a compelling problem in and of itself, the methods introduced in this paper offer a practical reason to study agents that blend VI and RL.

The results presented in Section 4.2 may provide new insight into "hindsight bias" or "creeping determinism" (Fischhoff, 1975), the observation that humans tend to overestimate the predictability of past events. Our results suggest that modeling stochastic events as deterministic events dependent on hidden information has beneficial variance-reducing properties for agents. This result is consistent with the *causal model theory* (CMT) of hindsight bias (Nestler et al., 2008), which argues that hindsight bias emerges as a consequence of an internal motivation to explain the outcome of given events, as opposed to memory-related distortions or errors. Variance reduction in learning processes provides a plausible evolutionary advantage that could support the CMT. Further exploration of this connection could be an interesting avenue for future research.

8. Related Work

There has been a great deal of work on reducing the variance of policy gradient methods through various means; however, the posterior value functions introduced here appear to be novel. The related literature on RL, POMDPs, and value function learning is too extensive to cover exhaustively, but we will try discuss the work we feel is most closely related.

The approach most similar to ours is the concurrent work

on *counterfactual credit assignment* (CCA) (Mesnard et al., 2020). Like our approach, CCA attempts to extract future information from a trajectory using a future-dependent value function, in our notation, $v^\theta(O_t, \Phi_t) \coloneqq \mathbb{E}[G_t|O_t, \Phi_t]$, where Φ_t is some sufficient statistic over future events that is independent of A_t. $v^\theta(O_t, \Phi_t)$, like posterior value functions, is shown to reduce (and never increase) the variance of the policy gradient estimator compared to $v^\theta(O_t)$. However, posterior value functions differ in that the statistics over Z_t need not be independent of A_t. Further, the independence maximization algorithm the authors introduce contrasts with the inference-based approach discussed herein.

There have been several other recent works involving the idea of hindsight. *Hindsight credit assignment* (Harutyunyan et al., 2019) is a method that attempts to determine the impact of a particular action on the likelihood of the agent finding itself in a particular state, and credits the action with the resulting returns in proportion to this likelihood. *Hindsight value modeling* (Guez et al., 2020) is superficially similar to our approach, but rather than using the features of the future trajectory (which are similar conceptually to our Z_t) to construct a control variate directly, an estimate of these features is made using only the information available at t and the control variate is constructed using this estimate.

Another related area is *Bayesian RL*, which was surveyed by Ghavamzadeh et al. (2016). Bayesian RL is a broad area, but methods typically involve maintaining a posterior distribution over some entity, such as a model of the MDP or the value function, with the goal of identifying the MDP (Osband et al., 2013) or value function (Eriksson et al., 2020) associated with the environment after many episodes.

9. Conclusions

We derived posterior value functions, a new class of value functions which use hindsight in order make better inferences about the value of previous states. We showed how posterior value functions can be used to reduce the variance of policy gradient estimators without introducing bias. We analyzed the variance of several advantage estimators and showed how posterior value functions are able to eliminate several common sources of variance. Further, we showed that by modeling stochasticity in the environment as a special case of partial observability, posterior value functions are able to reduce variance even in settings which are generally considered fully observable. We introduced a pair of methods for learning posterior value functions, and demonstrated their properties on a simple tabular problem. Finally, we suggested several future directions for research, including connections with other areas such as variational inference and human psychology.

Acknowledgements

Special thanks to Wes Cowley for assisting with the preparation of this manuscript and correcting dozens of typographical and grammatical errors.

References

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 242–252, 2019.

Baxter, J. and Bartlett, P. L. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

Bertsekas, D. P. and Tsitsiklis, J. N. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

Eriksson, H., Jorge, E., Dimitrakakis, C., Basu, D., and Grover, D. Inferential induction: A novel framework for Bayesian reinforcement learning. *arXiv preprint arXiv:2002.03098*, 2020.

Fischhoff, B. Hindsight≠ foresight: the effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, 1(3):288–299, 1975.

Ghavamzadeh, M., Mannor, S., Pineau, J., and Tamar, A. Bayesian reinforcement learning: A survey. *arXiv preprint arXiv:1609.04436*, 2016.

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.

Guez, A., Viola, F., Weber, T., Buesing, L., Kapturowski, S., Precup, D., Silver, D., and Heess, N. Value-driven hindsight modelling. *arXiv preprint arXiv:2002.08329*, 2020.

Harutyunyan, A., Dabney, W., Mesnard, T., Azar, M. G., Piot, B., Heess, N., van Hasselt, H. P., Wayne, G., Singh, S., Precup, D., and Munos, R. Hindsight credit assignment. In *Advances in Neural Information Processing Systems*, pp. 12488–12497, 2019.

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

Mesnard, T., Weber, T., Viola, F., Thakoor, S., Saade, A., Harutyunyan, A., Dabney, W., Stepleton, T., Heess, N.,

Guez, A., Hutter, M., Buesing, L., and Munos, R. Counterfactual credit assignment in model-free reinforcement learning. *arXiv preprint arXiv:2011.09464*, 2020.

Nestler, S., Blank, H., and von Collani, G. Hindsight bias and causal attribution: A causal model theory of creeping determinism. *Social Psychology*, 39(3):182–188, 2008.

Osband, I., Russo, D., and Van Roy, B. (More) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pp. 3003–3011, 2013.

Puterman, M. L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Hoboken, NJ, 2014.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, pp. 1057–1063, 1999.

Thomas, P. S. and Brunskill, E. Policy gradient methods for reinforcement learning with function approximation and action-dependent baselines. *arXiv preprint arXiv:1706.06643*, 2017.

Tucker, G., Bhupatiraju, S., Gu, S., Turner, R. E., Ghahramani, Z., and Levine, S. The mirage of action-dependent baselines in reinforcement learning. *arXiv preprint arXiv:1802.10031*, 2018.

Wierstra, D., Förster, A., Peters, J., and Schmidhuber, J. Recurrent policy gradients. *Logic Journal of the IGPL*, 18(5):620–634, 2010.

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.