Journal of Machine Learning Research 5 (2004) 1471–1530 Submitted 10/03; Revised 10/04; Published 11/04

Variance Reduction Techniques for Gradient Estimates in Reinforcement Learning

Evan Greensmith EVAN@CSL.ANU.EDU.AU
Research School of Information Sciences and Engineering
Australian National University
Canberra 0200, Australia

Peter L. Bartlett BARTLETT@STAT.BERKELEY.EDU
Computer Science Division & Department of Statistics
UC Berkeley
Berkeley, CA 94720, USA

Jonathan Baxter JBAXTER@PANSCIENT.COM
Panscient Pty. Ltd.
10 Gawler Terrace
Walkerville, SA 5081, Australia

Editor: Michael Littman

Abstract

Policy gradient methods for reinforcement learning avoid some of the undesirable properties of the value function approaches, such as policy degradation (Baxter and Bartlett, 2001). However, the variance of the performance gradient estimates obtained from the simulation is sometimes excessive. In this paper, we consider variance reduction methods that were developed for Monte Carlo estimates of integrals. We study two commonly used policy gradient techniques, the baseline and actor-critic methods, from this perspective. Both can be interpreted as additive control variate variance reduction methods. We consider the expected average reward performance measure, and we focus on the GPOMDP algorithm for estimating performance gradients in partially observable Markov decision processes controlled by stochastic reactive policies. We give bounds for the estimation error of the gradient estimates for both baseline and actor-critic algorithms, in terms of the sample size and mixing properties of the controlled system. For the baseline technique, we compute the optimal baseline, and show that the popular approach of using the average reward to define the baseline can be suboptimal. For actor-critic algorithms, we show that using the true value function as the critic can be suboptimal. We also discuss algorithms for estimating the optimal baseline and approximate value function.

Keywords: reinforcement learning, policy gradient, baseline, actor-critic, GPOMDP

1. Introduction

The task in reinforcement learning problems is to select a controller that will perform well in some given environment. This environment is often modelled as a partially observable Markov decision process (POMDP); see, for example, Kaelbling et al. (1998); Aberdeen (2002); Lovejoy (1991). At any step in time this process sits in some state, and that state is updated when the POMDP is supplied with an action. An observation is generated from the current state and given as information to a controller. A reward is also generated, as an indication of how good that state is to be in.

The controller can use the observations to determine which action to produce, thereby altering the POMDP state. The expectation of the average reward over possible future sequences of states given a particular controller (the expected average reward) can be used as a measure of how well a controller performs. This performance measure can then be used to select a controller that will perform well.

Given a parameterized space of controllers, one method to select a controller is by gradient ascent (see, for example, Glynn, 1990; Glynn and L'Ecuyer, 1995; Reiman and Weiss, 1989; Rubinstein, 1991; Williams, 1992). An initial controller is selected, then the gradient direction in the controller space of the expected average reward is calculated. The gradient information can then be used to find the locally optimal controller for the problem. The benefit of using a gradient approach, as opposed to directly comparing the expected average reward at different points, is that it can be less susceptible to error in the presence of noise. The noise arises from the fact that we estimate, rather than calculate, properties of the controlled POMDP.

Determining the gradient requires the calculation of an integral. We can produce an estimate of this integral through Monte Carlo techniques. This changes the integration problem into one of calculating a weighted average of samples. It turns out that these samples can be generated purely by watching the controller act in the environment (see Section 3.3). However, this estimation tends to have a high variance associated with it, which means a large number of steps is needed to obtain a good estimate.

GPOMDP (Baxter and Bartlett, 2001) is an algorithm for generating an estimate of the gradient in this way. Compared with other approaches (such as the algorithms described in Glynn, 1990; Rubinstein, 1991; Williams, 1992, for example), it is especially suitable for systems with large state spaces, when the time between visits to a recurrent state is large but the mixing time of the controlled POMDP is short. However, it can suffer from the problem of high variance in its estimates. We seek to alter GPOMDP so that the estimation variance is reduced, and thereby reduce the number of steps required to train a controller.

One generic approach to reducing the variance of Monte Carlo estimates of integrals is to use an additive control variate (see, for example, Hammersley and Handscomb, 1965; Fishman, 1996; Evans and Swartz, 2000). Suppose we wish to estimate the integral of the function $\mathrm{f} : \mathcal{X} \to \mathbb{R}$, and we happen to know the value of the integral of another function on the same space $\varphi : \mathcal{X} \to \mathbb{R}$. As we have

$$\int_{\mathcal{X}} \mathrm{f}(x) = \int_{\mathcal{X}} (\mathrm{f}(x) - \varphi(x)) + \int_{\mathcal{X}} \varphi(x) \tag{1}$$

the integral of $\mathrm{f}(x) - \varphi(x)$ can be estimated instead. Obviously if $\varphi(x) = \mathrm{f}(x)$ then we have managed to reduce our variance to zero. More generally,

$$\mathrm{Var}(f - \varphi) = \mathrm{Var}(f) - 2\mathrm{Cov}(f, \varphi) + \mathrm{Var}(\varphi).$$

If φ and f are strongly correlated, so that the covariance term on the right hand side is greater than the variance of φ, then a variance improvement has been made over the original estimation problem.

In this paper, we consider two applications of the control variate approach to the problem of gradient estimation in reinforcement learning. The first is the technique of adding a baseline, which is often used as a way to affect estimation variance whilst adding no bias. We show that adding a baseline can be viewed as a control variate method, and we find the optimal choice of baseline to use. We show that the additional variance of a suboptimal baseline can be expressed as a certain weighted squared distance between the baseline and the optimal one. A constant baseline, which

does not depend on the state, has been commonly suggested (Sutton and Barto, 1998; Williams, 1992; Kimura et al., 1995, 1997; Kimura and Kobayashi, 1998b; Marbach and Tsitsiklis, 2001). The expectation over all states of the discounted value of the state has been proposed, and widely used, as a constant baseline, by replacing the reward at each step by the difference between the reward and the average reward. We give bounds on the estimation variance that show that, perhaps surprisingly, this may not be the best choice. Our results are consistent with the experimental observations of Dayan (1990).

The second application of the control variate approach is the use of a value function. The discounted value function is usually not known, and needs to be estimated. Using some fixed, or learnt, value function in place of this estimate can reduce the overall estimation variance. Such *actor-critic methods* have been investigated extensively (Barto et al., 1983; Kimura and Kobayashi, 1998a; Baird, 1999; Sutton et al., 2000; Konda and Tsitsiklis, 2000, 2003). Generally the idea is to minimize some notion of distance between the value function and the true discounted value function, using, for example, TD (Sutton, 1988) or Least-Squares TD (Bradtke and Barto, 1996). In this paper we show that this may not be the best approach: selecting a value function to be equal to the true discounted value function is not always the best choice. Even more surprisingly, we give examples for which the use of a value function that is different from the true discounted value function reduces the variance to zero, for no increase in bias. We consider a value function to be forming part of a control variate, and find a corresponding bound on the expected squared error (that is, including the estimation variance) of the gradient estimate produced in this way.

While the main contribution of this paper is in understanding a variety of ideas in gradient estimation as variance reduction techniques, our results suggest a number of algorithms that could be used to augment the GPOMDP algorithm. We present new algorithms to learn the optimum baseline, and to learn a value function that minimizes the bound on the expected squared error of a gradient estimate, and we describe the results of preliminary experiments, which show that these algorithms give performance improvements.

2. Overview of Paper

Section 3 gives some background information. The POMDP setting and controller are defined, and the measure of performance and its gradient are described. Monte Carlo estimation of integrals, and how these integrals can be estimated, is covered, followed by a discussion of the GPOMDP algorithm, and how it relates to the Monte Carlo estimations. Finally, we outline the control variates that we use.

The samples used in the Monte Carlo estimations are taken from a single sequence of observations. Little can be said about the correlations between these samples. However, Section 4 shows that we can bound the effect they have on the variance in terms of the variance of the iid case (that is, when samples are generated iid according to the stationary distribution of the Markov chain).

Section 5 derives results for a baseline control variate in the iid setting, using results in Section 4 to interpret these as bounds in the more general case. In particular, we give an expression for the minimum variance that may be obtained, and the baseline that achieves this minimum variance. The section also compares the minimum variance against the common technique of using the expectation over states of the discounted value function, and it looks at a restricted class of baselines that use only observation information.

Section 6 looks at the technique of replacing the estimate of the discounted value function with some value function, in a control variate context. It shows that using the true discounted value function may not be the best choice, and that additional gains may be made. It also gives bounds on the expected squared error introduced by a value function.

Section 7 presents an algorithm to learn the optimal baseline. It also presents an algorithm to learn a value function by minimizing an estimate of the resulting expected squared error. Section 8 describes the results of experiments investigating the performance of these algorithms.

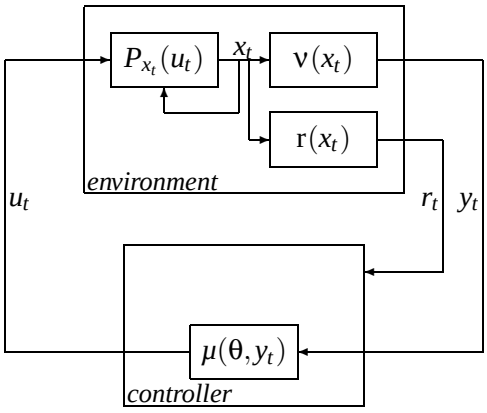
3. Background

Here we formally define the learning setting, including the performance and its gradient. We then give an intuitive discussion of the GPOMDP algorithm, starting with its approximation to the true gradient, and how it may be estimated by Monte Carlo techniques. Finally, we introduce the two variance reduction techniques studied in this paper.

3.1 System Model

A partially observable Markov decision process (POMDP) can be modelled by a system consisting of a state space, $\mathcal{S}$, an action space, $\mathcal{U}$, and an observation space, $\mathcal{Y}$, all of which will be considered finite here. State transitions are governed by a set of probability transition matrices $P(u)$, where $u \in \mathcal{U}$, components of which will be denoted $\mathrm{p}_{ij}(u)$, where $i, j \in \mathcal{S}$. There is also an observation process $\nu : \mathcal{S} \to \mathcal{P}_{\mathcal{Y}}$, where $\mathcal{P}_{\mathcal{Y}}$ is the space of probability distributions over $\mathcal{Y}$, and a reward function $\mathrm{r} : \mathcal{S} \to \mathbb{R}$. Together these define the POMDP $(\mathcal{S}, \mathcal{U}, \mathcal{Y}, P, \nu, \mathrm{r})$.

A policy for this POMDP is a mapping $\mu : \mathcal{Y}^* \to \mathcal{P}_{\mathcal{U}}$, where $\mathcal{Y}^*$ denotes the space of all finite sequences of observations $y_1, \ldots, y_t \in \mathcal{Y}$ and $\mathcal{P}_{\mathcal{U}}$ is the space of probability distributions over $\mathcal{U}$. If only the set of reactive policies $\mu : \mathcal{Y} \to \mathcal{P}_{\mathcal{U}}$ is considered then the joint process of state, observation and action, denoted $\{X_t, Y_t, U_t\}$, is Markov. This paper considers reactive parameterized policies $\mu(y, \theta)$, where $\theta \in \mathbb{R}^K$ and $y \in \mathcal{Y}$. A reactive parameterized policy together with a POMDP defines a *controlled POMDP* $(\mathcal{S}, \mathcal{U}, \mathcal{Y}, P, \nu, \mathrm{r}, \mu)$. See Figure 1.



Average Reward: $\eta = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} \mathrm{r}(X_t)\right]$

Figure 1: POMDP with reactive parameterized policy

Given a controlled POMDP the subprocess of states, $\{X_t\}$, is also Markov. A parameterized transition matrix $P(\theta)$, with entries $\mathrm{p}_{ij}(\theta)$, can be constructed, with

$$\mathrm{p}_{ij}(\theta) = \mathbb{E}_{y\sim\nu(i)}\left[\mathbb{E}_{u\sim\mu(y,\theta)}\left[\mathrm{p}_{ij}(u)\right]\right] = \sum_{y\in\mathcal{Y},u\in\mathcal{U}} \nu_y(i)\mu_u(y,\theta)\mathrm{p}_{ij}(u),$$

where $\nu_y(i)$ denotes the probability of observation y given the state i, and $\mu_u(y,\theta)$ denotes the probability of action u given the parameters θ and an observation y. The Markov chain $\mathrm{M}(\theta) = (\mathcal{S}, P(\theta))$ then describes the behavior of the process $\{X_t\}$.

We will also be interested in the special case where the state is fully observable.

Definition 1. *A* controlled Markov decision process *is a controlled POMDP* $(\mathcal{S},\mathcal{U},\mathcal{Y},P,\nu,\mathrm{r},\mu)$ *with* $\mathcal{Y}=\mathcal{S}$ *and* $\nu_y(i)=\delta_{yi}$, *where*

$$\delta_{yi} = \begin{cases} 1 & y = i \\ 0 & \textit{otherwise}, \end{cases}$$

and is defined by the tuple $(\mathcal{S},\mathcal{U},P,\mathrm{r},\mu)$.

In this case the set of reactive policies contains the optimal policy, that is, for our performance measure there is a reactive policy that will perform at least as well as any history dependent policy. Indeed, we need only consider mappings to point distributions over actions. Of course, this is not necessarily true of the parameterized class of reactive policies. In the partially observable setting the optimal policy may be history dependent; although a reactive policy may still perform well. For a study of using reactive policies for POMDPs see Singh et al. (1994); Jaakkola et al. (1995); Baird (1999). For a recent survey of POMDP techniques see Aberdeen (2002).

We operate under a number of assumptions for the controlled POMDP $(\mathcal{S},\mathcal{U},\mathcal{Y},P,\nu,\mathrm{r},\mu)$. Note that any arbitrary vector v is considered to be a column vector, and that we write v' to denote its transpose, a row vector. Also, the operator ∇ takes a function $\mathrm{f}(\theta)$ to a vector of its partial derivatives, that is

$$\nabla \mathrm{f}(\theta) = \left(\frac{\partial \mathrm{f}(\theta)}{\partial\theta_1},\ldots,\frac{\partial \mathrm{f}(\theta)}{\partial\theta_K}\right)',$$

where θ_k denotes the k^{th} element of θ.

Assumption 1. *For all* $\theta\in\mathbb{R}^K$ *the Markov chain* $\mathrm{M}(\theta)=(\mathcal{S},P(\theta))$ *is irreducible and aperiodic (ergodic), and hence has a unique stationary distribution* $\pi(\theta)$ *satisfying*

$$\pi(\theta)'P(\theta) = \pi(\theta)'$$

The terms *irreducible* and *aperiodic* are defined in Appendix A. Appendix A also contains a discussion of Assumption 1 and how both the irreducibility and aperiodicity conditions may be relaxed.

Assumption 2. *There is a* $\mathbf{R}<\infty$ *such that for all* $i\in\mathcal{S}$, $|\mathrm{r}(i)|\leq\mathbf{R}$.

Assumption 3. *For all* $u\in\mathcal{U}$, $y\in\mathcal{Y}$ *and* $\theta\in\mathbb{R}^K$ *the partial derivatives*

$$\frac{\partial\mu_u(y,\theta)}{\partial\theta_k}, \qquad \forall k\in\{1,\ldots,K\}$$

exist and there is a $\mathbf{B} < \infty$ *such that the Euclidean norms*

$$\left\| \frac{\nabla \mu_u(y,\theta)}{\mu_u(y,\theta)} \right\|$$

are uniformly bounded by $\mathbf{B}$*. We interpret* $0/0$ *to be* 0 *here, that is, we may have* $\mu_u(y,\theta) = 0$ *provided* $\|\nabla \mu_u(y,\theta)\| = 0$*. The Euclidean norm of a vector* v *is given by* $\sqrt{\sum_k v_k^2}$*.*

Note that Assumption 3 implies that

$$\left\| \frac{\nabla \mathrm{p}_{ij}(\theta)}{\mathrm{p}_{ij}(\theta)} \right\| \leq \mathbf{B},$$

where, as in Assumption 3, we interpret $0/0$ to be 0, and so we may have $\mathrm{p}_{ij}(\theta) = 0$ provided $\|\nabla \mathrm{p}_{ij}(\theta)\| = 0$. This bound can be seen from

$$\begin{aligned}
\left\|\nabla \mathrm{p}_{ij}(\theta)\right\| &= \left\| \nabla \sum_{y \in \mathcal{Y}, u \in \mathcal{U}} \nu_y(i) \mu_u(y,\theta) \mathrm{p}_{ij}(u) \right\| \\
&= \left\| \sum_{y \in \mathcal{Y}, u \in \mathcal{U}} \nu_y(i) \nabla \mu_u(y,\theta) \mathrm{p}_{ij}(u) \right\| \\
&\leq \mathbf{B} \sum_{y \in \mathcal{Y}, u \in \mathcal{U}} \nu_y(i) \mu_u(y,\theta) \mathrm{p}_{ij}(u) \\
&= \mathbf{B} \mathrm{p}_{ij}(\theta).
\end{aligned}$$

A useful measure of the system's performance is the expected average reward,

$$\eta(\theta) \stackrel{\text{def}}{=} \lim_{T \to \infty} \mathbb{E}\left[\frac{1}{T} \sum_{t=0}^{T-1} \mathrm{r}(X_t) \right]. \tag{2}$$

From Equation (24) in Appendix A we see that $\eta(\theta) = \mathbb{E}[\mathrm{r}(X) | X \sim \pi(\theta)]$, and hence is independent of the starting state. In this paper we analyze certain training algorithms that aim to select a policy such that this quantity is (locally) maximized.

It is also useful to consider the discounted value function,

$$\mathrm{J}_\beta(i,\theta) \stackrel{\text{def}}{=} \lim_{T \to \infty} \mathbb{E}\left[\sum_{t=0}^{T-1} \beta^t \mathrm{r}(X_t) \middle| X_0 = i \right].$$

Throughout the rest of the paper the dependence upon θ is assumed, and dropped in the notation.

3.2 Gradient Calculation

It is shown in Baxter and Bartlett (2001) that we can calculate an approximation to the gradient of the expected average reward by

$$\nabla_\beta \eta = \sum_{i,j \in \mathcal{S}} \pi_i \nabla p_{ij} \mathrm{J}_\beta(j),$$

and that the limit of $\nabla_\beta \eta$ as β approaches 1 is the true gradient $\nabla\eta$. Note that $\nabla_\beta\eta$ is a parameterized vector in $\mathbb{R}^K$ approximating the gradient of η, and there need not exist any function $f(\theta)$ with $\nabla f(\theta) = \nabla_\beta \eta$.

The gradient approximation $\nabla_\beta\eta$ can be considered as the integration over the state transition space,

$$\nabla_\beta \eta = \int_{(i,j)\in\mathcal{S}\times\mathcal{S}} \pi_i \nabla p_{ij} J_\beta(j) \mathfrak{C}(di \times dj), \tag{3}$$

where $\mathfrak{C}$ is a counting measure, that is, for a countable space $\mathcal{C}$, and a set $A \subset \mathcal{C}$, we have $\mathfrak{C}(A) = \text{card}(A)$ when A is finite, and $\mathfrak{C}(A) = \infty$ otherwise. Here $\text{card}(A)$ is the cardinality of the set A. It is unlikely that the true value function will be known. The value function can, however, be expressed as the integral over a sample path of the chain, as Assumption 1 implies ergodicity.

$$\nabla_\beta \eta = \int_{(i_0,i_1,\ldots)\in\mathcal{S}\times\mathcal{S}\times\ldots} \pi_{i_0} (\nabla p_{i_0 i_1}) p_{i_1 i_2} p_{i_2 i_3} \ldots \left(r(i_1) + \beta r(i_2) + \beta^2 r(i_3) + \cdots\right) \mathfrak{C}(di_0 \times \ldots).$$

To aid in analysis, the problem will be split into an integral and a sub integral problem.

$$\begin{aligned}
\nabla_\beta \eta &= \int_{(i,j)\in\mathcal{S}\times\mathcal{S}} \int_{(x_1,\ldots)\in\mathcal{S}\times\ldots} \pi_i (\nabla p_{ij}) \delta_{x_1 j} p_{x_1 x_2} \ldots (r(x_1) + \cdots) \mathfrak{C}(dx_1 \times \ldots) \mathfrak{C}(di \times dj) \\
&= \int_{(i,j)\in\mathcal{S}\times\mathcal{S}} \pi_i (\nabla p_{ij}) \int_{(x_1,\ldots)\in\mathcal{S}\times\ldots} \delta_{x_1 j} p_{x_1 x_2} \ldots (r(x_1) + \cdots) \mathfrak{C}(dx_1 \times \ldots) \mathfrak{C}(di \times dj).
\end{aligned}$$

3.3 Monte Carlo Estimation

Integrals can be estimated through the use of Monte Carlo techniques by averaging over samples taken from a particular distribution (see Hammersley and Handscomb, 1965; Fishman, 1996; Evans and Swartz, 2000). Take a function $f : \mathcal{X} \to \mathbb{R}$ and a probability distribution ρ over the space $\mathcal{X}$. An unbiased estimate of $\int_{x\in\mathcal{X}} f(x)$ can be generated from samples $\{x_0, x_1, \ldots, x_{m-1}\}$ taken from ρ by

$$\frac{1}{m} \sum_{n=0}^{m-1} \frac{f(x_n)}{\rho(x_n)}.$$

Consider a finite ergodic Markov chain $M = (\mathcal{S}, P)$ with stationary distribution π. Generate the Markov process $\{X_t\}$ from M starting from the stationary distribution. The integral of the function $f : \mathcal{S} \to \mathbb{R}$ over the space $\mathcal{S}$ can be estimated by

$$\frac{1}{T} \sum_{t=0}^{T-1} \frac{f(X_t)}{\pi_{X_t}}.$$

This can be used to estimate the integral

$$\int_{(i,j)\in\mathcal{S}\times\mathcal{S}} \pi_i \nabla p_{ij} J_\beta(j) \mathfrak{C}(di \times dj).$$

The finite ergodic Markov chain $M = (\mathcal{S}, P)$, with stationary distribution π, can be used to create the extended Markov process $\{X_t, X_{t+1}\}$ and its associated chain. Its stationary distribution has the probability mass function $\rho(i,j) = \pi_i p_{ij}$, allowing the estimation of the above integral by

$$\frac{1}{T} \sum_{t=0}^{T-1} \frac{\nabla p_{X_t X_{t+1}}}{p_{X_t X_{t+1}}} J_{t+1}, \qquad J_t = \sum_{s=t}^{\infty} \beta^{s-t} r(X_s). \tag{4}$$

In addition to the Monte Carlo estimation, the value function has been replaced with an unbiased estimate of the value function. In practice we would need to truncate this sum; a point discussed in the next section. Note, however, that

$$\begin{aligned}\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla p_{X_tX_{t+1}}}{p_{X_tX_{t+1}}}J_{t+1}\right] &= \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\frac{\nabla p_{X_tX_{t+1}}}{p_{X_tX_{t+1}}}\mathbb{E}\left[J_{t+1}\middle|X_{t+1}\right]\right]\\ &= \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla p_{X_tX_{t+1}}}{p_{X_tX_{t+1}}}\mathrm{J}_\beta(X_{t+1})\right].\end{aligned}$$

We will often be looking at estimates produced by larger Markov chains, such as that formed by the process $\{X_t, Y_t, U_t, X_{t+1}\}$. The discussion above also holds for functions on such chains.

3.4 GPOMDP Algorithm

The GPOMDP algorithm uses a single sample path of the Markov process $\{Z_t\} = \{X_t, Y_t, U_t, X_{t+1}\}$ to produce an estimate of $\nabla_\beta \eta$. We denote an estimate produced by GPOMDP with T samples by Δ_T.

$$\Delta_T \stackrel{\text{def}}{=} \frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}J_{t+1}, \qquad J_t \stackrel{\text{def}}{=} \sum_{s=t}^{T}\beta^{s-t}\mathrm{r}(X_s). \tag{5}$$

This differs from the estimate given in (4), but can be obtained similarly by considering the estimation of $\nabla_\beta \eta$ by samples from $\{Z_t\}$, and noting that

$$\nabla p_{ij} = \sum_{y\in\mathcal{Y},u\in\mathcal{U}} \nu_y(i)\nabla\mu_u(y)\mathrm{p}_{ij}(u).$$

GPOMDP can be represented as the two dimensional calculation

$$\begin{array}{rccccccccc}\Delta_T = \frac{1}{T}\Big(& \mathrm{f}(Z_0) & J_1 & + & \mathrm{f}(Z_1) & J_2 & + & \cdots & + & \mathrm{f}(Z_{T-1}) \quad J_T \quad \Big)\\ & & \|^{\text{def}} & & & \|^{\text{def}} & & & & \|^{\text{def}}\\ & & \mathrm{g}(Z_0) & & & \mathrm{g}(Z_1) & & \vdots & & \mathrm{g}(Z_{T-1})\\ & & +\beta\mathrm{g}(Z_1) & & & +\beta\mathrm{g}(Z_2) & & & & \\ & & +\beta^2\mathrm{g}(Z_2) & & & \vdots & & & & \\ & & \vdots & & & +\beta^{T-2}\mathrm{g}(Z_{T-1}) & & & & \\ & & +\beta^{T-1}\mathrm{g}(Z_{T-1}) & & & & & & & \end{array}$$

where $\mathrm{f}(Z_t) = (\nabla\mu_{U_t}(Y_t))/\mu_{U_t}(Y_t)$ and $\mathrm{g}(Z_t) = \mathrm{r}(X_{t+1})$.

One way to understand the behavior of GPOMDP is to assume that the chains being used to calculate each J_t sample are independent. This is reasonable when the chain is rapidly mixing and T is large compared with the mixing time, because then most pairs J_{t_1} and J_{t_2} are approximately independent. Replacing J_t by these independent versions, $J_t^{(\text{ind})}$, the calculation becomes

$$\begin{array}{rccccccccc}\Delta_T^{(\text{ind})} \stackrel{\text{def}}{=} \frac{1}{T}\Big(& \mathrm{f}(Z_0) & J_1^{(\text{ind})} & + & \mathrm{f}(Z_1) & J_2^{(\text{ind})} & + & \cdots & + & \mathrm{f}(Z_{T-1}) \quad J_T^{(\text{ind})} \quad \Big)\\ & & \|^{\text{def}} & & & \|^{\text{def}} & & & & \|^{\text{def}}\\ & & \mathrm{g}(Z_{00}) & & & \mathrm{g}(Z_{10}) & & \vdots & & \mathrm{g}\big(Z_{(T-1)0}\big)\\ & & +\beta\mathrm{g}(Z_{01}) & & & +\beta\mathrm{g}(Z_{11}) & & & & \\ & & +\beta^2\mathrm{g}(Z_{02}) & & & \vdots & & & & \\ & & \vdots & & & +\beta^{T-2}\mathrm{g}\big(Z_{1(T-2)}\big) & & & & \\ & & +\beta^{T-1}\mathrm{g}\big(Z_{0(T-1)}\big) & & & & & & & \end{array}$$

where the truncated process $\{Z_{tn}\}$ is an independent sample path generated from the Markov chain of the associated POMDP starting from the state $Z_t = Z_{t0}$.

The truncation of the discounted sum of future rewards would cause a bias from $\nabla_\beta \eta$. By considering T to be large compared to $1/(1-\beta)$ then this bias becomes small for a large proportion of the samples. Replacing each $J_t^{(\mathrm{ind})}$ by an untruncated version, $J_t^{(\mathrm{est})}$, shows how GPOMDP can be thought of as similar to the calculation

$$\begin{array}{ccccccccc}
\Delta_T^{(\mathrm{est})} \stackrel{\mathrm{def}}{=} \frac{1}{T}\Big(& \mathrm{f}(Z_0) & J_1^{(\mathrm{est})} & + & \mathrm{f}(Z_1) & J_2^{(\mathrm{est})} & + \cdots + & \mathrm{f}(Z_{T-1}) & J_T^{(\mathrm{est})} \Big) \\
& & \| & & & \| & & & \| \\
& & \mathrm{g}(Z_{00}) & & & \mathrm{g}(Z_{10}) & \vdots & & \mathrm{g}\big(Z_{(T-1)0}\big) \\
& & +\beta \mathrm{g}(Z_{01}) & & & +\beta \mathrm{g}(Z_{11}) & & & +\beta \mathrm{g}\big(Z_{(T-1)1}\big) \\
& & +\beta^2 \mathrm{g}(Z_{02}) & & & +\beta^2 \mathrm{g}(Z_{12}) & & & +\beta^2 \mathrm{g}\big(Z_{(T-1)2}\big) \\
& & \vdots & & & \vdots & & & \vdots
\end{array}$$

The altered Δ_T sum can be written as

$$\Delta_T^{(\mathrm{est})} = \frac{1}{T} \sum_{t=0}^{T-1} \frac{\nabla \mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)} J_{t+1}^{(\mathrm{est})}. \tag{6}$$

3.5 Variance Reduction

Equation (1) shows how a control variate can be used to change an estimation problem. To be of benefit the use of the control variate must lower estimation variance, and the integral of the control variate must have a known value. We look at two classes of control variate for which the value of the integral may be determined (or assumed).

The Monte Carlo estimates performed use correlated samples, making it difficult to analyze the variance gain. Given that we wish to deal with quite unrestricted environments, little is known about this sample correlation. We therefore consider the case of iid samples and show how this case gives a bound on the case using correlated samples.

The first form of control variate considered is the baseline control variate. With this, the integral shown in Equation (3) is altered by a control variate of the form $\pi_i \nabla p_{ij} \mathrm{b}(i)$.

$$\begin{aligned}
\int_{(i,j)\in\mathcal{S}\times\mathcal{S}} \pi_i \nabla p_{ij} \mathrm{J}_\beta(j) \mathfrak{C}(di\times dj) &= \int_{(i,j)\in\mathcal{S}\times\mathcal{S}} \pi_i \nabla p_{ij} \big(\mathrm{J}_\beta(j) - \mathrm{b}(i)\big) \mathfrak{C}(di\times dj) \\
&\quad + \int_{(i,j)\in\mathcal{S}\times\mathcal{S}} \pi_i \nabla p_{ij} \mathrm{b}(i) \mathfrak{C}(di\times dj)
\end{aligned}$$

The integral of the control variate term is zero, since

$$\begin{aligned}
\int_{(i,j)\in\mathcal{S}\times\mathcal{S}} \pi_i \nabla p_{ij} \mathrm{b}(i) \mathfrak{C}(di\times dj) &= \sum_{i\in\mathcal{S}} \pi_i \mathrm{b}(i) \nabla \sum_{j\in\mathcal{S}} p_{ij} \\
&= \sum_{i\in\mathcal{S}} \pi_i \mathrm{b}(i) \nabla(1) \\
&= 0.
\end{aligned} \tag{7}$$

Thus, we are free to select an arbitrary $\mathrm{b}(i)$ with consideration for the variance minimization alone.

The second form of control variate considered is constructed from a value function, $V(j)$, a mapping $\mathcal{S} \to \mathbb{R}$.

$$\begin{aligned}\int_{(i,j)\in\mathcal{S}\times\mathcal{S}} \pi_i \nabla p_{ij} J_\beta(j) \mathfrak{C}(di \times dj) &= \int_{(i,j)\in\mathcal{S}\times\mathcal{S}} \pi_i \nabla p_{ij} \left(J_\beta(j) - \left(J_\beta(j) - V(j)\right)\right) \mathfrak{C}(di \times dj) \\ &\quad + \int_{(i,j)\in\mathcal{S}\times\mathcal{S}} \pi_i \nabla p_{ij} \left(J_\beta(j) - V(j)\right) \mathfrak{C}(di \times dj)\end{aligned}$$

The integral of this control variate (the last term in the equation above) is the error associated with using a value function in place of the true discounted value function. The task is then to find a value function such that the integral of the control variate is small, and yet it still provides good variance minimization of the estimated integral.

Note that the integrals being estimated here are vector quantities. We consider the trace of the covariance matrix of these quantities, that is, the sum of the variance of the components of the vector. Given the random vector $A = (A_1, A_2, \ldots, A_k)'$, we write

$$\mathrm{Var}(A) = \sum_{m=1}^{k} \mathrm{Var}(A_m) = \mathbb{E}\left[(A - \mathbb{E}[A])'(A - \mathbb{E}[A])\right] = \mathbb{E}\left[(A - \mathbb{E}[A])^2\right],$$

where, for a vector a, a^2 denotes $a'a$.

4. Dependent Samples

In Sections 5 and 6 we study the variance of quantities that, like $\Delta_T^{(\mathrm{est})}$ (Equation (6)), are formed from the sample average of a process generated by a controlled (PO)MDP. From Section 3 we know this process is Markov, is ergodic, and has a stationary distribution, and so the sample average is an estimate of the expectation of a sample drawn from the stationary distribution, π (note that, as in Section 3.3, we can also look at samples formed from an extended space, and its associated stationary distributions). In this section we investigate how the variance of the sample average relates to the variance of a sample drawn from π. This allows us to derive results for the variance of a sample drawn from π and relate them to the variance of the sample average. In the iid case, that is, when the process generates a sequence of samples $X_0, \ldots, X_{T-1}$ drawn independently from the distribution π, we have the relationship

$$\mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1} f(X_t)\right) = \frac{1}{T}\mathrm{Var}(f(X)),$$

where X is a random variable also distributed according to π. More generally, however, correlation between the samples makes finding an exact relationship difficult. Instead we look to find a bound of the form

$$\mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1} f(X_t)\right) \leq h\left(\frac{1}{T}\mathrm{Var}(f(X))\right),$$

where h is some "well behaved" function.

We first define a notion of mixing time for a Markov chain. The mixing time is a measure of the forgetfulness of a Markov chain. More specifically, it is a measure of how long it takes for the distance between the distributions of two sequences, starting in distinct states, to become small. The distance measure we use is the total variation distance.

Definition 2. *The* total variation distance *between two distributions p,q on the finite set $\mathcal{S}$ is given by*

$$\mathrm{d}_{TV}(p,q) \stackrel{\text{def}}{=} \frac{1}{2}\sum_{i\in\mathcal{S}} |p_i - q_i|.$$

Definition 3. *The* mixing time *of a finite ergodic Markov chain $M=(\mathcal{S},P)$ is defined as*

$$\tau \stackrel{\text{def}}{=} \min\left\{ t>0 : \max_{i,j} \mathrm{d}_{TV}\left(P_i^t, P_j^t\right) \le e^{-1} \right\},$$

where P_i^t denotes the i^{th} row of the t-step transition matrix P^t.

The results in this section are given for a Markov chain with mixing time τ. In later sections we will use τ as a measure of the mixing time of the resultant Markov chain of states of a controlled (PO)MDP, but will look at sample averages over larger spaces. The following lemma, due to Bartlett and Baxter (2002), shows that the mixing time does not grow too fast when looking at the Markov chain on sequences of states.

Lemma 1. *(Bartlett and Baxter, 2002, Lemma 4.3) If the Markov chain $M=(\mathcal{S},P)$ has mixing time τ, then the Markov chain formed by the process $\{X_t, X_{t+1}, \ldots, X_{t+k}\}$ has mixing time $\tilde{\tau}$, where*

$$\tilde{\tau} \le \tau \ln\left(e(k+1)\right).$$

Note 1. *For a controlled POMDP, the Markov chain formed by the process $\{X_t, X_{t+1}, \ldots, X_{t+k}\}$ has the same mixing time as the Markov chain formed by the process $\{X_t, Y_t, U_t, X_{t+1}, \ldots, Y_{t+k-1}, U_{t+k-1}, X_{t+k}\}$.*

We now look at showing the relationship between the covariance between two samples in a sequence and the variance of an individual sample. We show that the gain of the covariance of two samples X_t, X_{t+s} over the variance of an individual sample decreases exponentially in s.

Theorem 2. *Let $M=(\mathcal{S},P)$ be a finite ergodic Markov chain, and let π be its stationary distribution. Let f be some mapping $\mathrm{f}:\mathcal{S}\to\mathbb{R}$. The tuple (M,f) has associated positive constants α and $\mathbf{L}$ (called* mixing constants *$(\alpha,\mathbf{L})$) such that, for all $t\ge 0$,*

$$|\mathrm{Cov}_\pi(t;\mathrm{f})| \le \mathbf{L}\alpha^t \mathrm{Var}(\mathrm{f}(X))$$

where $X\sim\pi$, and $\mathrm{Cov}_\pi(t;\mathrm{f})$ is the auto-covariance of the process $\{\mathrm{f}(X_s)\}$, i.e. $\mathrm{Cov}_\pi(t;\mathrm{f}) = \mathbb{E}_\pi\left[(\mathrm{f}(X_s)-\mathbb{E}_\pi\mathrm{f}(X_s))\,(\mathrm{f}(X_{s+t})-\mathbb{E}_\pi\mathrm{f}(X_{s+t}))\right]$, where $\mathbb{E}_\pi[\cdot]$ denotes the expectation over the chain with initial distribution π. Furthermore, if M has mixing time τ, we have:

1. *for reversible M, and any f, we may choose $\mathbf{L}=2e$ and $\alpha=\exp(-1/\tau)$; and*
2. *for any M (that is, any finite ergodic M), and any f, we may choose $\mathbf{L}=\sqrt{2|\mathcal{S}|e}$ and $\alpha=\exp(-1/(2\tau))$.*

The proof is shown in Appendix B, along with proofs for the rest of this section.
Using this result, the variance of the sample average can be bounded as follows.

Theorem 3. *Let $M = (\mathcal{S}, P)$ be a finite ergodic Markov chain, with mixing time τ, and let π be its stationary distribution. Let f be some mapping $\mathrm{f} : \mathcal{S} \to \mathbb{R}$. Let $\{X_t\}$ be a sample path generated by M, with initial distribution π, and let $X \sim \pi$. With (M, f) mixing constants $(\alpha, \mathbf{L})$ chosen such that $\alpha \leq \exp(-1/(2\tau))$, there is an $\Omega^* \leq 6\mathbf{L}\tau$ such that*

$$\mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathrm{f}(X_t)\right) \leq \frac{\Omega^*}{T}\mathrm{Var}(\mathrm{f}(X)).$$

Provided acceptable mixing constants can be chosen, Theorem 3 gives the same rate as in the case of independent random variables, that is, the variance decreases as $O(1/T)$. The most that can be done to improve the bound of Theorem 3 is to reduce the constant Ω^*. It was seen, in Theorem 2, that good mixing constants can be chosen for functions on reversible Markov chains. We would like to deal with more general chains also, and the mixing constants given in Theorem 2 for functions on ergodic Markov chains lead to Ω^* increasing with the size of the state space. However, for bounded functions on ergodic Markov chains we have the following result:

Theorem 4. *Let $M = (\mathcal{S}, P)$ be a finite ergodic Markov chain, and let π be its stationary distribution. If M has mixing time τ, then for any function $\mathrm{f} : \mathcal{S} \to [-c, c]$ and any $0 < \varepsilon < e^{-1}$, we have*

$$\mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathrm{f}(X_t)\right) \leq \varepsilon + \left(1 + 25\tau(1+c)\varepsilon + 4\tau\ln\frac{1}{\varepsilon}\right)\frac{1}{T}\mathrm{Var}\left(\mathrm{f}(X)\right),$$

where $\{X_t\}$ is a process generated by M with initial distribution $X_0 \sim \pi$, and $X \sim \pi$.

Here we have an additional error ε, which we may decrease at the cost of a $\ln\varepsilon^{-1}$ penalty in the constant multiplying the variance term.

Consider the following corollary of Theorem 4.

Corollary 5. *Let $M = (\mathcal{S}, P)$ be a finite ergodic Markov chain, and let π be its stationary distribution. If M has mixing time τ, then for any function $\mathrm{f} : \mathcal{S} \to [-c, c]$, we have*

$$\begin{aligned}\mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathrm{f}(X_t)\right) \quad \leq \quad & 4\tau\ln\left(7(1+c) + \frac{1}{4\tau}\left(\frac{1}{T}\mathrm{Var}\left(\mathrm{f}(X)\right)\right)^{-1}\right)\frac{1}{T}\mathrm{Var}\left(\mathrm{f}(X)\right) \\ & + (1+8\tau)\frac{1}{T}\mathrm{Var}\left(\mathrm{f}(X)\right)\end{aligned}$$

where $\{X_t\}$ is a process generated by M with initial distribution $X_0 \sim \pi$, and $X \sim \pi$.

Here, again, our bound approaches zero as $\mathrm{Var}(\mathrm{f}(X))/T \to 0$, but at the slightly slower rate of

$$O\left(\frac{1}{T}\mathrm{Var}(\mathrm{f}(X))\ln\left(e + \left(\frac{1}{T}\mathrm{Var}(\mathrm{f}(X))\right)^{-1}\right)\right),$$

where we have ignored the dependence on τ and c. For a fixed variance the rate of decrease in T is $O(\ln(T)/T)$, slightly worse than the $O(1/T)$ rate for independent random variables.

5. Baseline Control Variate

As stated previously, a baseline may be selected with regard given only to the estimation variance. In this section we consider how the baseline affects the variance of our gradient estimates when the samples are iid, and the discounted value function is known. We show that, when using Theorem 3 or Theorem 4 to bound covariance terms, this is reasonable, and in fact the error in analysis (that is, from not analyzing the variance of Δ_T with baseline directly) associated with the choice of baseline is negligible. This statement will be made more precise later.

Section 5.2 looks at the Markov chain of states generated by the controlled POMDP and is concerned with producing a baseline $\mathrm{b}_{\mathcal{S}} : \mathcal{S} \to \mathbb{R}$ to minimize the variance

$$\sigma^2_{\mathcal{S}}(\mathrm{b}_{\mathcal{S}}) = \mathrm{Var}_\pi\left(\frac{\nabla p_{ij}}{p_{ij}}\left(\mathrm{J}_\beta(j) - \mathrm{b}_{\mathcal{S}}(i)\right)\right), \tag{8}$$

where, for some $f : \mathcal{S}\times\mathcal{S} \to \mathbb{R}^K$, $\mathrm{Var}_\pi(\mathrm{f}(i,j)) = \mathbb{E}_\pi\left(\mathrm{f}(i,j) - \mathbb{E}_\pi \mathrm{f}(i,j)\right)^2$ with $\mathbb{E}_\pi[\cdot]$ denoting the expectation over the random variables i, j with $i \sim \pi$ and $j \sim P_i$. Equation (8) serves as a definition of $\sigma^2_{\mathcal{S}}(\mathrm{b}_{\mathcal{S}})$. The section gives the minimal value of this variance, and the minimizing baseline. Additionally, the minimum variance and corresponding baseline is given for the case where the baseline is a constant, $b \in \mathbb{R}$. In both cases, we give expressions for the excess variance of a suboptimal baseline, in terms of a weighted squared distance between the baseline and the optimal one. We can thus show the difference between the variance for the optimal constant baseline and the variance obtained when $b = \mathbb{E}_\pi \mathrm{J}_\beta(i)$.

Section 5.3 considers a baseline $\mathrm{b}_{\mathcal{Y}} : \mathcal{Y} \to \mathbb{R}$ for the GPOMDP estimates. It shows how to minimize the variance of the estimate

$$\sigma^2_{\mathcal{Y}}(\mathrm{b}_{\mathcal{Y}}) = \mathrm{Var}_\pi\left(\frac{\nabla \mu_u(y)}{\mu_u(y)}\left(\mathrm{J}_\beta(j) - \mathrm{b}_{\mathcal{Y}}(y)\right)\right), \tag{9}$$

where, for some $f : \mathcal{S}\times\mathcal{Y}\times\mathcal{U}\times\mathcal{S} \to \mathbb{R}^K$, $\mathrm{Var}_\pi(\mathrm{f}(i,y,u,j)) = \mathbb{E}_\pi\left(\mathrm{f}(i,y,u,j) - \mathbb{E}_\pi \mathrm{f}(i,y,u,j)\right)^2$ with, in this case, $\mathbb{E}_\pi[\cdot]$ denoting the expectation over the random variables i, y, u, j with $i \sim \pi$, $y \sim \nu(i)$, $u \sim \mu(y)$, and $j \sim P_i(u)$. Equation (9) serves as a definition of $\sigma^2_{\mathcal{Y}}(\mathrm{b}_{\mathcal{Y}})$. The case where the state space is fully observed is shown as a consequence.

5.1 Matching Analysis and Algorithm

The analysis in following sections will look at Equation (8) and Equation (9). Here we will show that the results of that analysis can be applied to the variance of a realizable algorithm for generating $\nabla_\beta \eta$ estimates. Specifically, we compare the variance quantity of Equation (9) to a slight variation of the Δ_T estimate produced by GPOMDP, where the chain is run for an extra S steps. We consider the estimate

$$\Delta_T^{(+S)} \stackrel{\text{def}}{=} \frac{1}{T}\sum_{t=0}^{T-1} \frac{\nabla \mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)} J_{t+1}^{(+S)}, \qquad J_t^{(+S)} \stackrel{\text{def}}{=} \sum_{s=t}^{T+S} \beta^{s-t}\mathrm{r}(X_s), \tag{10}$$

and are interested in improving the variance by use of a baseline, that is, by using the estimate

$$\Delta_T^{(+S)}(\mathrm{b}_{\mathcal{Y}}) \stackrel{\text{def}}{=} \frac{1}{T}\sum_{t=0}^{T-1} \frac{\nabla \mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(J_{t+1}^{(+S)} - \mathrm{b}_{\mathcal{Y}}(Y_t)\right).$$

We delay the main result of the section, Theorem 7, to gain an insight into the ideas behind it. In Section 3.4 we saw how GPOMDP can be thought of as similar to the estimate $\Delta_T^{(\text{est})}$, Equation (6). Using a baseline gives us the new estimate

$$\Delta_T^{(\text{est})}(\mathrm{b}_{\mathcal{Y}}) \stackrel{\text{def}}{=} \frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(J_{t+1}^{(\text{est})} - \mathrm{b}_{\mathcal{Y}}(Y_t)\right). \tag{11}$$

The term $J_t^{(\text{est})}$ in Equation (11) is an unbiased estimate of the discounted value function. The following lemma shows that, in analysis of the baseline, we can consider the discounted value function to be known, not estimated.

Lemma 6. *Let $\{X_t\}$ be a random process over the space $\mathcal{X}$. Define arbitrary functions on the space $\mathcal{X}$: $\mathrm{f}:\mathcal{X}\to\mathbb{R}$, $\mathrm{J}:\mathcal{X}\to\mathbb{R}$, and $\mathrm{a}:\mathcal{X}\to\mathbb{R}$. For all t let J_t be a random variable such that $\mathbb{E}[J_t|X_t=i]=\mathrm{J}(i)$. Then*

$$\begin{aligned}\mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathrm{f}(X_t)\left(J_t-\mathrm{a}(X_t)\right)\right) - \mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathrm{f}(X_t)\left(\mathrm{J}(X_t)-\mathrm{a}(X_t)\right)\right)\\ = \mathbb{E}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathrm{f}(X_t)\left(J_t-\mathrm{J}(X_t)\right)\right)^2\end{aligned}$$

The proof of Lemma 6 is given in Appendix C, along with the proof of Theorem 7 below.

Direct application of Lemma 6 gives,

$$\begin{aligned}\mathrm{Var}\left(\Delta_T^{(\text{est})}(\mathrm{b}_{\mathcal{Y}})\right) &= \mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(\mathrm{J}_\beta(X_{t+1})-\mathrm{b}_{\mathcal{Y}}(Y_t)\right)\right)\\ &\quad+\mathbb{E}\left(\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(J_{t+1}^{(\text{est})}-\mathrm{J}_\beta(X_{t+1})\right)\right)^2.\end{aligned}$$

Thus, we see that we can split the variance of this estimate into two components: the first is the variance of this estimate with $J_t^{(\text{est})}$ replaced by the true discounted value function; and the second is a component independent of our choice of baseline. We can now use Theorem 3 or Corollary 5 to bound the covariance terms, leaving us to analyze Equation (9).

We can obtain the same sort of result, using the same reasoning, for the estimate we are interested in studying in practice: $\Delta_T^{(+S)}(\mathrm{b}_{\mathcal{Y}})$ (see Equation (12) below).

Theorem 7. *Let $D=(\mathcal{S},\mathcal{U},\mathcal{Y},P,\nu,\mathrm{r},\mu)$ be a controlled POMDP satisfying Assumptions 1, 2 and 3. Let $M=(\mathcal{S},P)$ be the resultant Markov chain of states, and let π be its stationary distribution; M has a mixing time τ; $\{Z_t\}=\{X_t,Y_t,U_t,X_{t+1}\}$ is a process generated by D, starting $X_0\sim\pi$. Suppose that $\mathrm{a}(\cdot)$ is a function uniformly bounded by $\mathbf{M}$, and $\mathcal{J}(j)$ is the random variable $\sum_{s=0}^{\infty}\beta^s\mathrm{r}(W_s)$ where the states W_s are generated by D starting in $W_0=j$. There are constants $C_1\leq 7+7\mathbf{B}(\mathbf{R}+\mathbf{M})$ and*

$C_2 = 20\tau\mathbf{B}^2\mathbf{R}(\mathbf{R}+\mathbf{M})$ *such that for all* $T, S \geq 1$ *we have*

$$\begin{aligned}
&\mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(J_{t+1}^{(+S)}-\mathrm{a}(Z_t)\right)\right)\\
&\leq \mathrm{h}\left(\frac{\tau\ln(e(S+1))}{T}\mathrm{Var}_\pi\left(\frac{\nabla\mu_u(y)}{\mu_u(y)}\left(\mathrm{J}_\beta(j)-\mathrm{a}(i,y,u,j)\right)\right)\right)\\
&\qquad+\mathrm{h}\left(\frac{\tau\ln(e(S+1))}{T}\mathbb{E}_\pi\left(\frac{\nabla\mu_u(y)}{\mu_u(y)}\left(\mathcal{J}(j)-\mathrm{J}_\beta(j)\right)\right)^2\right)\\
&\qquad+\frac{2C_2}{(1-\beta)^2}\left[\ln\frac{1}{\beta}+\ln\left(\frac{C_1}{1-\beta}+\frac{K(1-\beta)^2}{C_2}\right)\right]\frac{(T+S)\ln(e(S+1))}{T}\beta^S,
\end{aligned}$$

where $\mathrm{h}:\mathbb{R}^+\to\mathbb{R}^+$ *is continuous and increasing with* $\mathrm{h}(0)=0$, *and is given by*

$$\mathrm{h}(x) = 9x + 4x\ln\left(\frac{C_1}{1-\beta}+\frac{K}{4}x^{-1}\right).$$

By selecting $S=T$ in Theorem 7, and applying to $\Delta_T^{(+S)}(\mathrm{b}_\mathcal{Y})$ with absolutely bounded $\mathrm{b}_\mathcal{Y}$, we obtain the desired result:

$$\mathrm{Var}\left(\Delta_T^{(+T)}(\mathrm{b}_\mathcal{Y})\right) \leq \mathrm{h}\left(\frac{\tau\ln(e(T+1))}{T}\sigma^2_\mathcal{Y}(\mathrm{b}_\mathcal{Y})\right)+\mathrm{N}(D,T)+O\left(\ln(T)\beta^T\right). \tag{12}$$

Here $\mathrm{N}(D,T)$ is the noise term due to using an estimate in place of the discounted value function, and does not depend on the choice of baseline. The remaining term is of the order $\ln(T)\beta^T$; it is almost exponentially decreasing in T, and hence negligible. The function h is due to the application of Theorem 4, and consequently the discussion in Section 4 on the rate of decrease applies here, that is, a log penalty is paid. In this case, for $\sigma^2_\mathcal{Y}(\mathrm{b}_\mathcal{Y})$ fixed, the rate of decrease is $O(\ln^2(T)/T)$.

Note that we may replace $(\nabla\mu_u(y))/\mu_u(y)$ with $(\nabla p_{ij})/p_{ij}$ in Theorem 7. So if the $(\nabla p_{ij})/p_{ij}$ can be calculated, then Theorem 7 also relates the analysis of Equation 8 with a realizable algorithm for generating $\nabla_\beta\eta$ estimates; in this case an estimate produced by watching the Markov process of states.

5.2 Markov Chains

Here we look at baselines for $\nabla_\beta\eta$ estimates for a parameterized Markov chain and associated reward function (a Markov reward process). The Markov chain of states generated by a controlled POMDP (together with the POMDPs reward function) is an example of such a process. However, the baselines discussed in this section require knowledge of the state to use, and knowledge of $(\nabla\mathrm{p}_{ij}(\theta))/\mathrm{p}_{ij}(\theta)$ to estimate. More practical results for POMDPs are given in the next section.

Consider the following assumption.

Assumption 4. *The parameterized Markov chain* $M(\theta)=(\mathcal{S},P(\theta))$ *and associated reward function* $r:\mathcal{S}\to\mathbb{R}$ *satisfy:* $M(\theta)$ *is irreducible and aperiodic, with stationary distribution* π*; there is a* $\mathbf{R}<\infty$ *such that for all* $i\in\mathcal{S}$ *we have* $|r(i)|\leq\mathbf{R}$*; and for all* $i,j\in\mathcal{S}$, *and all* $\theta\in\mathbb{R}^K$, *the partial derivatives* $\nabla\mathrm{p}_{ij}(\theta)$ *exist, and there is a* $\mathbf{B}<\infty$ *such that* $\|(\nabla\mathrm{p}_{ij}(\theta))/\mathrm{p}_{ij}(\theta)\|\leq\mathbf{B}$.

For any controlled POMDP satisfying Assumptions 1, 2 and 3, Assumption 4 is satisfied for the Markov chain formed by the subprocess $\{X_t\}$ together with the reward function for the controlled POMDP.

Now consider a control variate of the form

$$\varphi_{\mathcal{S}}(i,j) \stackrel{\text{def}}{=} \pi_i \nabla p_{ij} \mathrm{b}_{\mathcal{S}}(i)$$

for estimation of the integral in Equation (3). We refer to the function $\mathrm{b}_{\mathcal{S}} : \mathcal{S} \to \mathbb{R}$ as a baseline.

As shown in Section 3.5, the integral of the baseline control variate $\varphi_{\mathcal{S}}(i,j)$ over $\mathcal{S} \times \mathcal{S}$ can be calculated analytically and is equal to zero. Thus an estimate of the integral

$$\int_{(i,j)\in\mathcal{S}\times\mathcal{S}} \left(\pi_i \nabla p_{ij} \mathrm{J}_\beta(j) - \varphi_{\mathcal{S}}(i,j)\right) \mathfrak{C}(di \times dj)$$

forms an unbiased estimate of $\nabla_\beta \eta$.

The following theorem gives the minimum variance, and the baseline to achieve the minimum variance. We use $\sigma^2_{\mathcal{S}}$ to denote the variance of the estimate without a baseline,

$$\sigma^2_{\mathcal{S}} = \mathrm{Var}_\pi\left(\frac{\nabla p_{ij}}{p_{ij}} \mathrm{J}_\beta(j)\right),$$

and we recall, from Equation (8), that $\sigma^2_{\mathcal{S}}(\mathrm{b}_{\mathcal{S}})$ denotes the variance with a baseline,

$$\sigma^2_{\mathcal{S}}(\mathrm{b}_{\mathcal{S}}) = \mathrm{Var}_\pi\left(\frac{\nabla p_{ij}}{p_{ij}} \left(\mathrm{J}_\beta(j) - \mathrm{b}_{\mathcal{S}}(i)\right)\right).$$

Theorem 8. *Let $M(\theta) = (\mathcal{S}, P(\theta))$ and $r : \mathcal{S} \to \mathbb{R}$ be a parameterized Markov chain and reward function satisfying Assumption 4. Then*

$$\sigma^2_{\mathcal{S}}(\mathrm{b}^*_{\mathcal{S}}) \stackrel{\text{def}}{=} \inf_{\mathrm{b}_{\mathcal{S}} \in \mathbb{R}^{\mathcal{S}}} \sigma^2_{\mathcal{S}}(\mathrm{b}_{\mathcal{S}}) = \sigma^2_{\mathcal{S}} - \mathbb{E}_{i\sim\pi}\left[\frac{\left(\mathbb{E}\left[(\nabla p_{ij}/p_{ij})^2 \mathrm{J}_\beta(j) \middle| i\right]\right)^2}{\mathbb{E}\left[(\nabla p_{ij}/p_{ij})^2 \middle| i\right]}\right],$$

where $\mathbb{E}[\cdot|i]$ is the expectation over the resultant state j conditioned on being in state i, that is, $j \sim P_i$, and $\mathbb{R}^{\mathcal{S}}$ is the space of functions mapping $\mathcal{S}$ to $\mathbb{R}$. This infimum is attained with the baseline

$$\mathrm{b}^*_{\mathcal{S}}(i) = \frac{\mathbb{E}\left[(\nabla p_{ij}/p_{ij})^2 \mathrm{J}_\beta(j) \middle| i\right]}{\mathbb{E}\left[(\nabla p_{ij}/p_{ij})^2 \middle| i\right]}.$$

The proof uses the following lemma.

Lemma 9. *For any $\mathrm{b}_{\mathcal{S}}$,*

$$\sigma^2_{\mathcal{S}}(\mathrm{b}_{\mathcal{S}}) = \sigma^2_{\mathcal{S}} + \mathbb{E}_\pi\left[\mathrm{b}^2_{\mathcal{S}}(i)\,\mathbb{E}\left[\left(\frac{\nabla p_{ij}}{p_{ij}}\right)^2 \middle| i\right] - 2\mathrm{b}_{\mathcal{S}}(i)\mathbb{E}\left[\left(\frac{\nabla p_{ij}}{p_{ij}}\right)^2 \mathrm{J}_\beta(j) \middle| i\right]\right].$$

Proof.

$$\begin{aligned}
\sigma^2_{\mathcal{S}}(\mathrm{b}_{\mathcal{S}}) &= \mathbb{E}_\pi\left(\frac{\nabla p_{ij}}{p_{ij}}\left(\mathrm{J}_\beta(j)-\mathrm{b}_{\mathcal{S}}(i)\right)-\mathbb{E}_\pi\left[\frac{\nabla p_{ij}}{p_{ij}}\left(\mathrm{J}_\beta(j)-\mathrm{b}_{\mathcal{S}}(i)\right)\right]\right)^2 \\
&= \mathbb{E}_\pi\left(\left(\frac{\nabla p_{ij}}{p_{ij}}\mathrm{J}_\beta(j)-\mathbb{E}_\pi\left[\frac{\nabla p_{ij}}{p_{ij}}\mathrm{J}_\beta(j)\right]\right)-\left(\frac{\nabla p_{ij}}{p_{ij}}\mathrm{b}_{\mathcal{S}}(i)-\mathbb{E}_\pi\left[\frac{\nabla p_{ij}}{p_{ij}}\mathrm{b}_{\mathcal{S}}(i)\right]\right)\right)^2 \\
&= \sigma^2_{\mathcal{S}}+\mathbb{E}_\pi\left[\left(\frac{\nabla p_{ij}}{p_{ij}}\mathrm{b}_{\mathcal{S}}(i)\right)^2-2\left(\frac{\nabla p_{ij}}{p_{ij}}\mathrm{b}_{\mathcal{S}}(i)\right)'\left(\frac{\nabla p_{ij}}{p_{ij}}\mathrm{J}_\beta(j)\right)\right] \qquad (13) \\
&= \sigma^2_{\mathcal{S}}+\mathbb{E}_{i\sim\pi}\left[\mathrm{b}^2_{\mathcal{S}}(i)\,\mathbb{E}\left[\left(\frac{\nabla p_{\tilde{i}\blacksquare}}{p_{\tilde{i}\blacksquare}}\right)^2\middle|\tilde{\imath}=i\right]\right. \\
&\qquad \left.-2\mathrm{b}_{\mathcal{S}}(i)\mathbb{E}\left[\left(\frac{\nabla p_{\tilde{i}\blacksquare}}{p_{\tilde{i}\blacksquare}}\right)^2\mathrm{J}_\beta(\blacksquare)\middle|\tilde{\imath}=i\right]\right],
\end{aligned}$$

where Equation (13) uses

$$\mathbb{E}_\pi\left[\frac{\nabla p_{ij}}{p_{ij}}\mathrm{b}_{\mathcal{S}}(i)\right]=\int_{(i,j)\in\mathcal{S}\times\mathcal{S}}\pi_i\nabla p_{ij}\mathrm{b}_{\mathcal{S}}(i)\mathfrak{C}(di\times dj)=0,$$

from (7). ■

Proof of Theorem 8. We use Lemma 9 and minimize for each $i\in\mathcal{S}$. Differentiating with respect to each $\mathrm{b}_{\mathcal{S}}(i)$ gives

$$\begin{aligned}
2\mathrm{b}_{\mathcal{S}}(i)\mathbb{E}\left[\left(\frac{\nabla p_{ij}}{p_{ij}}\right)^2\middle|i\right]-2\mathbb{E}\left[\left(\frac{\nabla p_{ij}}{p_{ij}}\right)^2\mathrm{J}_\beta(j)\middle|i\right] &= 0 \\
\Rightarrow \mathrm{b}_{\mathcal{S}}(i) &= \frac{\mathbb{E}\left[(\nabla p_{ij}/p_{ij})^2\mathrm{J}_\beta(j)\middle|i\right]}{\mathbb{E}\left[(\nabla p_{ij}/p_{ij})^2\middle|i\right]},
\end{aligned}$$

which implies the result. ■

The following theorem shows that the excess variance due to a suboptimal baseline function can be expressed as a weighted squared distance to the optimal baseline.

Theorem 10. *Let $M(\theta)=(\mathcal{S},P(\theta))$ and $r:\mathcal{S}\to\mathbb{R}$ be a parameterized Markov chain and reward function satisfying Assumption 4. Then*

$$\sigma^2_{\mathcal{S}}(\mathrm{b}_{\mathcal{S}})-\sigma^2_{\mathcal{S}}(\mathrm{b}^*_{\mathcal{S}})=\mathbb{E}_\pi\left[\left(\frac{\nabla p_{ij}}{p_{ij}}\right)^2\left(\mathrm{b}_{\mathcal{S}}(i)-\mathrm{b}^*_{\mathcal{S}}(i)\right)^2\right].$$

Proof. For each $i\in\mathcal{S}$, define S_i and W_i as

$$\begin{aligned}
S_i &= \mathbb{E}\left[\left(\frac{\nabla p_{ij}}{p_{ij}}\right)^2\middle|i\right], \\
W_i &= \mathbb{E}\left[\left(\frac{\nabla p_{ij}}{p_{ij}}\right)^2\mathrm{J}_\beta(j)\middle|i\right].
\end{aligned}$$

Lemma 9 and the definition of $\mathrm{b}^*_{\mathcal{S}}$ in Theorem 8 imply that

$$\begin{aligned}
\sigma^2_{\mathcal{S}}(\mathrm{b}_{\mathcal{S}}) - \sigma^2_{\mathcal{S}}(\mathrm{b}^*_{\mathcal{S}}) &= \mathbb{E}_\pi\left[\mathrm{b}^2_{\mathcal{S}}(i)\,S_i - 2\mathrm{b}_{\mathcal{S}}(i)W_i + \frac{W_i^2}{S_i}\right] \\
&= \mathbb{E}_\pi\left(\mathrm{b}_{\mathcal{S}}(i)\sqrt{S_i} - \frac{W_i}{\sqrt{S_i}}\right)^2 \\
&= \mathbb{E}_\pi\left[\left(\mathrm{b}_{\mathcal{S}}(i) - \mathrm{b}^*_{\mathcal{S}}(i)\right)^2 S_i\right] \\
&= \mathbb{E}_\pi\left[\left(\frac{\nabla p_{ij}}{p_{ij}}\right)^2 \left(\mathrm{b}_{\mathcal{S}}(i) - \mathrm{b}^*_{\mathcal{S}}(i)\right)^2\right].
\end{aligned}$$

■

The following theorem gives the minimum variance, the baseline to achieve the minimum variance, and the additional variance away from this minimum, when restricted to a constant baseline, $b \in \mathbb{R}$. We use $\sigma^2_{\mathcal{S}}(b)$ to denote the variance with constant baseline b,

$$\sigma^2_{\mathcal{S}}(b) = \mathrm{Var}_\pi\left(\frac{\nabla p_{ij}}{p_{ij}}\left(\mathrm{J}_\beta(j) - b\right)\right). \tag{14}$$

The proof uses Lemma 9 in the same way as the proof of Theorem 8. The proof of the last statement follows that of Theorem 10 by replacing S_i with $S = \mathbb{E}_\pi S_i$, and W_i with $W = \mathbb{E}_\pi W_i$.

Theorem 11. *Let $M(\theta) = (\mathcal{S}, P(\theta))$ and $r : \mathcal{S} \to \mathbb{R}$ be a parameterized Markov chain and reward function satisfying Assumption 4. Then*

$$\sigma^2_{\mathcal{S}}(b^*) \stackrel{\text{def}}{=} \inf_{b\in\mathbb{R}} \sigma^2_{\mathcal{S}}(b) = \sigma^2_{\mathcal{S}} - \frac{\left(\mathbb{E}_\pi\left[(\nabla p_{ij}/p_{ij})^2\,\mathrm{J}_\beta(j)\right]\right)^2}{\mathbb{E}_\pi\left(\nabla p_{ij}/p_{ij}\right)^2}.$$

This infimum is attained with

$$b^* = \frac{\mathbb{E}_\pi\left[(\nabla p_{ij}/p_{ij})^2\,\mathrm{J}_\beta(j)\right]}{\mathbb{E}_\pi\left(\nabla p_{ij}/p_{ij}\right)^2}.$$

The excess variance due to a suboptimal constant baseline b is given by,

$$\sigma^2_{\mathcal{S}}(b) - \sigma^2_{\mathcal{S}}(b^*) = \mathbb{E}_\pi\left(\frac{\nabla p_{ij}}{p_{ij}}\right)^2 (b - b^*)^2.$$

A baseline of the form $b = \mathbb{E}_\pi \mathrm{J}_\beta(i)$ is often promoted as a good choice. Theorem 11 gives us a tool to measure how far this choice is from the optimum.

Corollary 12. *Let $M(\theta) = (\mathcal{S}, P(\theta))$ and $r : \mathcal{S} \to \mathbb{R}$ be a Markov chain and reward function satisfying Assumption 4. Then*

$$\sigma^2_{\mathcal{S}}(\mathbb{E}\mathrm{J}_\beta(i)) - \sigma^2_{\mathcal{S}}(b^*) = \frac{\left(\mathbb{E}_\pi\left(\nabla p_{ij}/p_{ij}\right)^2 \mathbb{E}_\pi \mathrm{J}_\beta(j) - \mathbb{E}_\pi\left[(\nabla p_{ij}/p_{ij})^2\,\mathrm{J}_\beta(j)\right]\right)^2}{\mathbb{E}_\pi\left(\nabla p_{ij}/p_{ij}\right)^2}.$$

Notice that the sub-optimality of the choice $b = \mathbb{E}_\pi \mathrm{J}_\beta(i)$ depends on the independence of the random variables $(\nabla p_{ij}/p_{ij})^2$ and $\mathrm{J}_\beta(j)$; if they are nearly independent, $\mathbb{E}_\pi \mathrm{J}_\beta(i)$ is a good choice.

Of course, when considering sample paths of Markov chains, Corollary 12 only shows the difference of the two *bounds* on the variance given by Theorem 7, but it gives an indication of the true distance. In particular, as the ratio of the mixing time to the sample path length becomes small, the difference between the variances in the dependent case approaches that of Corollary 12.

5.3 POMDPs

Consider a control variate over the extended space $\mathcal{S} \times \mathcal{Y} \times \mathcal{U} \times \mathcal{S}$ of the form

$$\varphi(i,y,u,j) = \pi_i \nu_y(i) \nabla \mu_u(y) \mathrm{p}_{ij}(u) \mathrm{b}(i,y).$$

Again, its integral is zero.

$$\begin{aligned}&\int_{(i,y,u,j)\in\mathcal{S}\times\mathcal{Y}\times\mathcal{U}\times\mathcal{S}} \varphi(i,y,u,j)\mathfrak{C}(di\times dy\times du\times dj)\\ &\quad= \sum_{i\in\mathcal{S},y\in\mathcal{Y}} \pi_i \nu_y(i)\mathrm{b}(i,y)\nabla\left(\sum_{u\in\mathcal{U},j\in\mathcal{S}} \mu_u(y)\mathrm{p}_{ij}(u)\right) = 0.\end{aligned}$$

Thus an unbiased estimate of the integral

$$\int_{(i,y,u,j)\in\mathcal{S}\times\mathcal{Y}\times\mathcal{U}\times\mathcal{S}} \left(\pi_i \nu_y(i)\nabla\mu_u(y)\mathrm{p}_{ij}(u)\mathrm{J}_\beta(j) - \varphi(i,y,u,j)\right)\mathfrak{C}(di\times dy\times du\times dj)$$

is an unbiased estimate of $\nabla_\beta \eta$. Here results analogous to those achieved for $\varphi_{\mathcal{S}}(i,j)$ can be obtained. However, we focus on the more interesting (and practical) case of the restricted control variate

$$\varphi_{\mathcal{Y}}(i,y,u,j) \stackrel{\text{def}}{=} \pi_i \nu_y(i)\nabla\mu_u(y)\mathrm{p}_{ij}(u)\mathrm{b}_{\mathcal{Y}}(y).$$

Here, only information that can be observed by the controller (the observations y) may be used to minimize the variance. Recall, from Equation (9), we use $\sigma^2_{\mathcal{Y}}(\mathrm{b}_{\mathcal{Y}})$ to denote the variance with such a restricted baseline control variate,

$$\sigma^2_{\mathcal{Y}}(\mathrm{b}_{\mathcal{Y}}) = \mathrm{Var}_\pi\left(\frac{\nabla\mu_u(y)}{\mu_u(y)}\left(\mathrm{J}_\beta(j) - \mathrm{b}_{\mathcal{Y}}(y)\right)\right).$$

We use $\sigma^2_{\mathcal{Y}}$ to denote the variance without a baseline, that is

$$\sigma^2_{\mathcal{Y}} = \mathrm{Var}_\pi\left(\frac{\nabla\mu_u(y)}{\mu_u(y)}\mathrm{J}_\beta(j)\right).$$

We have the following theorem.

Theorem 13. *Let $D = (\mathcal{S},\mathcal{U},\mathcal{Y},P,\nu,\mathrm{r},\mu)$ be a controlled POMDP satisfying Assumptions 1, 2 and 3, with stationary distribution π. Then*

$$\sigma^2_{\mathcal{Y}}(\mathrm{b}^*_{\mathcal{Y}}) \stackrel{\text{def}}{=} \inf_{\mathrm{b}_{\mathcal{Y}}\in\mathbb{R}^{\mathcal{Y}}} \sigma^2_{\mathcal{Y}}(\mathrm{b}_{\mathcal{Y}}) = \sigma^2_{\mathcal{Y}} - \mathbb{E}_\pi\left[\frac{\left(\mathbb{E}_\pi\left[\left(\nabla\mu_u(y)/\mu_u(y)\right)^2\mathrm{J}_\beta(j)\middle|y\right]\right)^2}{\mathbb{E}_\pi\left[\left(\nabla\mu_u(y)/\mu_u(y)\right)^2\middle|y\right]}\right],$$

where $\mathbb{E}_\pi[\cdot|y]$ is the expectation (of π-distributed random variables, that is, random variables distributed as in $\mathbb{E}_\pi[\cdot]$) conditioned on observing y, and this infimum is attained with the baseline

$$\mathrm{b}^*_{\mathcal{Y}}(y) = \frac{\mathbb{E}_\pi\left[(\nabla\mu_u(y)/\mu_u(y))^2 \mathrm{J}_\beta(j)\middle| y\right]}{\mathbb{E}_\pi\left[(\nabla\mu_u(y)/\mu_u(y))^2\middle| y\right]}.$$

Furthermore, when restricted to the class of constant baselines, $b \in \mathbb{R}$, the minimal variance occurs with

$$b^* = \frac{\mathbb{E}_\pi\left[(\nabla\mu_u(y)/\mu_u(y))^2 \mathrm{J}_\beta(j)\right]}{\mathbb{E}_\pi(\nabla\mu_u(y)/\mu_u(y))^2}.$$

We have again used b^* to denote the optimal constant baseline. Note though that the b^* here differs from that given in Theorem 11. The proof uses the following lemma.

Lemma 14. *For any* $\mathrm{b}_{\mathcal{Y}}$*,*

$$\sigma^2_{\mathcal{Y}}(\mathrm{b}_{\mathcal{Y}}) = \sigma^2_{\mathcal{Y}} + \mathbb{E}_\pi\left[\mathrm{b}^2_{\mathcal{Y}}(y)\,\mathbb{E}_\pi\left[\left(\frac{\nabla\mu_u(y)}{\mu_u(y)}\right)^2\middle| y\right] - 2\mathrm{b}_{\mathcal{Y}}(y)\mathbb{E}_\pi\left[\left(\frac{\nabla\mu_u(y)}{\mu_u(y)}\right)^2 J_\beta(j)\middle| y\right]\right].$$

Proof. Following the same steps as in the proof of Lemma 9,

$$\begin{aligned}
\sigma^2_{\mathcal{Y}}(\mathrm{b}_{\mathcal{Y}}) &= \mathbb{E}_\pi\left(\frac{\nabla\mu_u(y)}{\mu_u(y)}\left(\mathrm{J}_\beta(j) - \mathrm{b}_{\mathcal{Y}}(y)\right) - \mathbb{E}_\pi\left[\frac{\nabla\mu_u(y)}{\mu_u(y)}\left(\mathrm{J}_\beta(j) - \mathrm{b}_{\mathcal{Y}}(y)\right)\right]\right)^2 \\
&= \sigma^2_{\mathcal{Y}} + \mathbb{E}_\pi\left[\left(\frac{\nabla\mu_u(y)}{\mu_u(y)}\mathrm{b}_{\mathcal{Y}}(y)\right)^2 - 2\left(\frac{\nabla\mu_u(y)}{\mu_u(y)}\mathrm{b}_{\mathcal{Y}}(y)\right)'\left(\frac{\nabla\mu_u(y)}{\mu_u(y)}\mathrm{J}_\beta(j)\right)\right] \\
&= \sigma^2_{\mathcal{Y}} + \sum_y\left[\mathrm{b}^2_{\mathcal{Y}}(y)\left(\sum_{i,u,j}\pi_i\nu_y(i)\mu_u(y)\mathrm{p}_{ij}(u)\left(\frac{\nabla\mu_u(y)}{\mu_u(y)}\right)^2\right)\right. \\
&\qquad \left. -2\mathrm{b}_{\mathcal{Y}}(y)\left(\sum_{i,u,j}\pi_i\nu_y(i)\mu_u(y)\mathrm{p}_{ij}(u)\left(\frac{\nabla\mu_u(y)}{\mu_u(y)}\right)^2\mathrm{J}_\beta(j)\right)\right].
\end{aligned}$$

Note that for functions $\mathrm{a}:\mathcal{Y}\to\mathbb{R}$ and $\mathrm{f}:\mathcal{S}\times\mathcal{Y}\times\mathcal{U}\times\mathcal{S}\to\mathbb{R}$

$$\begin{aligned}
&\sum_y \mathrm{a}(y)\sum_{\tilde{\imath},\tilde{u},\tilde{\jmath}}\pi_{\tilde{\imath}}\nu_y(\tilde{\imath})\mu_{\tilde{u}}(y)\mathrm{p}_{\tilde{\imath}\tilde{\jmath}}(\tilde{u})\mathrm{f}(\tilde{\imath},y,\tilde{u},\tilde{\jmath}) \\
&\quad= \sum_y \mathrm{a}(y)\sum_i \pi_i\nu_y(i)\sum_{\tilde{\imath},\tilde{y},\tilde{u},\tilde{\jmath}}\frac{\delta_{y\tilde{y}}\pi_{\tilde{\imath}}\nu_y(\tilde{\imath})\mu_{\tilde{u}}(y)\mathrm{p}_{\tilde{\imath}\tilde{\jmath}}(\tilde{u})}{\sum_i\pi_i\nu_y(i)}\mathrm{f}(\tilde{\imath},\tilde{y},\tilde{u},\tilde{\jmath}) \\
&\quad= \sum_{i,y}\pi_i\nu_y(i)\mathrm{a}(y)\sum_{\tilde{\imath},\tilde{y},\tilde{u},\tilde{\jmath}}\Pr\{\tilde{\imath},\tilde{y},\tilde{u},\tilde{\jmath}|\tilde{y}=y\}\,\mathrm{f}(\tilde{\imath},\tilde{y},\tilde{u},\tilde{\jmath}) \\
&\quad= \mathbb{E}_\pi[\mathrm{a}(y)\mathbb{E}_\pi[\mathrm{f}(i,y,u,j)|y]],
\end{aligned}$$

implying the result. ∎

Proof of Theorem 13. We apply Lemma 14 and minimize for each $\mathrm{b}_{\mathcal{Y}}(y)$ independently, to obtain

$$\mathrm{b}_{\mathcal{Y}}(y) = \frac{\mathbb{E}_\pi\left[\left(\nabla\mu_u(y)/\mu_u(y)\right)^2 \mathrm{J}_\beta(j)\middle| y\right]}{\mathbb{E}_\pi\left[\left(\nabla\mu_u(y)/\mu_u(y)\right)^2\middle| y\right]}.$$

Substituting gives the optimal variance. A similar argument gives the optimal constant baseline. ■

Example 1. *Consider the k-armed bandit problem (for example, see Sutton and Barto, 1998). Here each action is taken independently and the resultant state depends only on the action performed; that is $\mu_u(y) = \mu_u$ and $\mathrm{p}_{ij}(u) = \mathrm{p}_j(u)$. So, writing $R_\beta = \mathbb{E}_{U_0\sim\mu}\left[\sum_{t=1}^\infty \beta^t \mathrm{r}(X_t)\right]$, we have*

$$\begin{aligned}
\nabla_\beta\eta &= \mathbb{E}_\pi\left[\frac{\nabla\mu_u(y)}{\mu_u(y)}\mathrm{J}_\beta(j)\right] \\
&= \mathbb{E}_{u\sim\mu}\left[\frac{\nabla\mu_u}{\mu_u}\left(\mathrm{r}(j)+R_\beta\right)\right] \\
&= \mathbb{E}_{u\sim\mu}\left[\frac{\nabla\mu_u}{\mu_u}\mathrm{r}(j)\right].
\end{aligned}$$

Note that this last line is β independent, and it follows from $\lim_{\beta\to 1}\nabla_\beta\eta = \nabla\eta$ that

$$\nabla\eta = \nabla_\beta\eta \qquad \forall\beta\in[0,1]. \tag{15}$$

For $k=2$ (2 actions $\{u_1,u_2\}$) we have $\mu_{u_1}+\mu_{u_2}=1$ and $\nabla\mu_{u_1} = -\nabla\mu_{u_2}$, and so the optimal constant baseline is given by

$$\begin{aligned}
b^* &= \frac{\mathbb{E}_\pi\left[\left(\nabla\mu_u(y)/\mu_u(y)\right)^2\mathrm{J}_\beta(j)\right]}{\mathbb{E}_\pi\left(\nabla\mu_u(y)/\mu_u(y)\right)^2} \\
&= \frac{\mathbb{E}_{u\sim\mu}\left[\left(\nabla\mu_u/\mu_u\right)^2\mathrm{r}(j)\right]}{\mathbb{E}_{u\sim\mu}\left(\nabla\mu_u/\mu_u\right)^2} + R_\beta \\
&= \frac{\mu_{u_1}\left(\nabla\mu_{u_1}/\mu_{u_1}\right)^2\mathbb{E}\left[\mathrm{r}|u_1\right] + \mu_{u_2}\left(\nabla\mu_{u_2}/\mu_{u_2}\right)^2\mathbb{E}\left[\mathrm{r}|u_2\right]}{\mu_{u_1}\left(\nabla\mu_{u_1}/\mu_{u_1}\right)^2 + \mu_{u_2}\left(\nabla\mu_{u_2}/\mu_{u_2}\right)^2} + R_\beta \\
&= \frac{\mu_{u_1}\mu_{u_2}}{\mu_{u_1}+\mu_{u_2}}\left(\frac{1}{\mu_{u_1}}\mathbb{E}\left[\mathrm{r}|u_1\right] + \frac{1}{\mu_{u_2}}\mathbb{E}\left[\mathrm{r}|u_2\right]\right) + R_\beta \\
&= \mu_{u_2}\mathbb{E}\left[\mathrm{r}|u_1\right] + \mu_{u_1}\mathbb{E}\left[\mathrm{r}|u_2\right] + R_\beta,
\end{aligned}$$

where we have used $\mathbb{E}\left[\mathrm{r}|u\right]$ to denote $\mathbb{E}_{j\sim\mathrm{p}(u)}\mathrm{r}(j)$. From (15) we know that β may be chosen arbitrarily. Choosing $\beta=0$ gives $R_\beta = 0$ and we regain the result of Dayan (1990).

In the special case of a controlled MDP we obtain the result that would be expected. This follows immediately from Theorem 13.

Corollary 15. *Let $D = (\mathcal{S},\mathcal{U},P,\mathrm{r},\mu)$ be a controlled MDP satisfying Assumptions 1, 2 and 3, with stationary distribution π. Then*

$$\inf_{\mathrm{b}_{\mathcal{Y}}\in\mathbb{R}^{\mathcal{S}}}\sigma^2_{\mathcal{Y}}(\mathrm{b}_{\mathcal{Y}}) = \sigma^2_{\mathcal{Y}} - \mathbb{E}_{i\sim\pi}\left[\frac{\left(\mathbb{E}\left[\left(\nabla\mu_u(i)/\mu_u(i)\right)^2\mathrm{J}_\beta(j)\middle| i\right]\right)^2}{\mathbb{E}\left[\left(\nabla\mu_u(i)/\mu_u(i)\right)^2\middle| i\right]}\right],$$

and this infimum is attained with the baseline

$$\mathrm{b}_{\mathcal{Y}}(i) = \frac{\mathbb{E}\left[\left(\nabla\mu_u(i)/\mu_u(i)\right)^2 \mathrm{J}_\beta(j)\middle| i\right]}{\mathbb{E}\left[\left(\nabla\mu_u(i)/\mu_u(i)\right)^2\middle| i\right]}.$$

The following theorem shows that, just as in the Markov chain case, the variance of an estimate with an arbitrary baseline can be expressed as the sum of the variance with the optimal baseline and a certain squared weighted distance between the baseline function and the optimal baseline function.

Theorem 16. *Let $(\mathcal{S},\mathcal{U},\mathcal{Y},P,\nu,\mathrm{r},\mu)$ be a controlled POMDP satisfying Assumptions 1, 2 and 3, with stationary distribution π. Then*

$$\sigma^2_{\mathcal{Y}}(\mathrm{b}_{\mathcal{Y}}) - \sigma^2_{\mathcal{Y}}(\mathrm{b}^*_{\mathcal{Y}}) = \mathbb{E}_\pi\left[\left(\frac{\nabla\mu_u(y)}{\mu_u(y)}\right)^2\left(\mathrm{b}_{\mathcal{Y}}(y) - \mathrm{b}^*_{\mathcal{Y}}(y)\right)^2\right].$$

Furthermore if the estimate using b^*, the optimal constant *baseline defined in Theorem 13, has variance $\sigma^2_{\mathcal{Y}}(b^*)$, we have that the variance $\sigma^2_{\mathcal{Y}}(b)$ of the gradient estimate with an arbitrary constant baseline is*

$$\sigma^2_{\mathcal{Y}}(b) - \sigma^2_{\mathcal{Y}}(b^*) = \mathbb{E}_\pi\left(\frac{\nabla\mu_u(y)}{\mu_u(y)}\right)^2 (b-b^*)^2.$$

Proof. For each $y \in \mathcal{Y}$, define S_y and W_y as

$$\begin{aligned} S_y &= \mathbb{E}\left[\left(\frac{\nabla\mu_u(y)}{\mu_u(y)}\right)^2\middle| y\right], \\ W_y &= \mathbb{E}\left[\left(\frac{\nabla\mu_u(y)}{\mu_u(y)}\right)^2 J_\beta(j)\middle| y\right]. \end{aligned}$$

Follow the steps in Theorem 10, replacing S_i with S_y, and W_i with W_y. The constant baseline case follows similarly by considering $S = \mathbb{E}_\pi S_y$ and $W = \mathbb{E}_\pi W_y$. ■

In Section 7.1 we will see how Theorem 16 can be used to construct a practical algorithm for finding a good baseline. In most cases it is not possible to calculate the optimal baseline, $\mathrm{b}^*_{\mathcal{Y}}$, a priori. However, for a parameterized class of baseline functions, a gradient descent approach could be used to find a good baseline. Section 7.1 explores this idea.

As before, Theorem 16 also gives us a tool to measure how far the baseline $b = \mathbb{E}_\pi \mathrm{J}_\beta(i)$ is from the optimum.

Corollary 17. *Let $D = (\mathcal{S},\mathcal{U},\mathcal{Y},P,\nu,\mathrm{r},\mu)$ be a controlled POMDP satisfying Assumptions 1, 2 and 3, with stationary distribution π. Then*

$$\sigma^2_{\mathcal{Y}}(\mathbb{E}_\pi \mathrm{J}_\beta(i)) - \inf_{b\in\mathbb{R}} \sigma^2_{\mathcal{Y}}(b) = \frac{\left(\mathbb{E}_\pi\left(\nabla\mu_u(y)/\mu_u(y)\right)^2 \mathbb{E}_\pi \mathrm{J}_\beta(j) - \mathbb{E}_\pi\left[\left(\nabla\mu_u(y)/\mu_u(y)\right)^2 \mathrm{J}_\beta(j)\right]\right)^2}{\mathbb{E}_\pi\left(\nabla\mu_u(y)/\mu_u(y)\right)^2}.$$

As in the case of a Markov reward process, the sub-optimality of the choice $b = \mathbb{E}_\pi \mathrm{J}_\beta(i)$ depends on the independence of the random variables $(\nabla\mu_u(y)/\mu_u(y))^2$ and $\mathrm{J}_\beta(j)$; if they are nearly independent, $\mathbb{E}_\pi \mathrm{J}_\beta(i)$ is a good choice.

6. Value Functions: Actor-Critic Methods

Consider the estimate produced by GPOMDP (see Equation (5)) in the MDP setting, where the state is observed. In this section we look at replacing J_t, the biased and noisy estimate of the discounted value function, in Δ_T with an arbitrary value function, that is, a function $\mathrm{V} : \mathcal{S} \to \mathbb{R}$. For a MDP, this gives the following estimate of $\nabla_\beta \eta$:

$$\Delta_T^{\mathrm{V}} \stackrel{\text{def}}{=} \frac{1}{T} \sum_{t=0}^{T-1} \frac{\nabla \mu_{U_t}(X_t)}{\mu_{U_t}(X_t)} \mathrm{V}(X_{t+1}). \tag{16}$$

Imagine that the discounted value function, J_β, is known. By replacing J_t with $\mathrm{J}_\beta(X_t)$ in Equation (5), that is, by choosing $\mathrm{V} = \mathrm{J}_\beta$, the bias and noise due to J_t is removed. This seems a good choice, but we may be able to do better. Indeed we will see that in some cases the selection of a value function differing from the discounted value function can remove *all* estimation variance, whilst introducing no bias.

6.1 Control Variate for a Value Function

Consider a control variate of the form

$$\varphi_\beta(i,u,j) \stackrel{\text{def}}{=} \pi_i \nabla \mu_u(i) \mathrm{p}_{ij}(u) \mathrm{A}_\beta(j)$$

where

$$\mathrm{A}_\beta(j) \stackrel{\text{def}}{=} \lim_{T\to\infty} \mathbb{E}\left[\sum_{k=1}^{T} \beta^{k-1} \mathrm{d}(X_{t+k}, X_{t+1+k}) \middle| X_{t+1} = j \right]$$

and

$$\mathrm{d}(i,j) \stackrel{\text{def}}{=} \mathrm{r}(i) + \beta \mathrm{V}(j) - \mathrm{V}(i).$$

We make the following assumption.

Assumption 5. *For all* $j \in \mathcal{S}$, $|\mathrm{V}(j)| \leq \mathbf{M} < \infty$.

Under this assumption, the estimation of the integral

$$\int_{(i,u,j)\in\mathcal{S}\times\mathcal{U}\times\mathcal{S}} \left(\pi_i \nabla \mu_u(i) \mathrm{p}_{ij}(u) \mathrm{J}_\beta(j) - \varphi_\beta(i,u,j) \right) \mathfrak{C}(di \times du \times dj) \tag{17}$$

has an expected bias from $\nabla_\beta \eta$ of

$$\int_{(i,u,j)\in\mathcal{S}\times\mathcal{U}\times\mathcal{S}} \varphi_\beta(i,u,j) \mathfrak{C}(di \times du \times dj) = \sum_{i\in\mathcal{S}, u\in\mathcal{U}, j\in\mathcal{S}} \pi_i \nabla \mu_u(i) \mathrm{p}_{ij}(u) \left(\mathrm{J}_\beta(j) - \mathrm{V}(j) \right).$$

This can be easily seen by noting that under Assumption 5, and as $\beta \in [0,1)$,

$$\begin{aligned} \mathrm{A}_\beta(j) &= \lim_{T\to\infty} \mathbb{E}\left[\sum_{k=1}^{T} \beta^{k-1} \left(\mathrm{r}(X_{t+k}) + \beta \mathrm{V}(X_{t+1+k}) - \mathrm{V}(X_{t+k}) \right) \middle| X_{t+1} = j \right] \\ &= \mathrm{J}_\beta(j) - \mathrm{V}(j) + \lim_{T\to\infty} \mathbb{E}\left[\beta^T \mathrm{V}(X_{t+1+T}) \middle| X_{t+1} = j \right] \\ &= \mathrm{J}_\beta(j) - \mathrm{V}(j). \end{aligned}$$

We see then that Δ_T^{V} gives an estimate of the integral in Equation (17). The following theorem gives a bound on the expected value of the squared Euclidean distance between this estimate and $\nabla_\beta \eta$. Notice that the bound includes both bias and variance terms.

Theorem 18. *Let $D = (\mathcal{S}, \mathcal{U}, P, \mathrm{r}, \mu)$ be a controlled MDP satisfying Assumptions 1, 2 and 3, with stationary distribution π. Let $\{X_t, U_t\}$ be a process generated by D, starting $X_0 \sim \pi$. Then*

$$\mathbb{E}\left(\Delta_T^{\mathrm{V}} - \nabla_\beta \eta\right)^2 = \mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1} \frac{\nabla \mu_{U_t}(X_t)}{\mu_{U_t}(X_t)} \mathrm{V}(X_{t+1})\right) + \left(\mathbb{E}_\pi\left[\frac{\nabla \mu_u(i)}{\mu_u(i)} \mathrm{A}_\beta(j)\right]\right)^2,$$

and hence there is an Ω^* such that

$$\mathbb{E}\left(\Delta_T^{\mathrm{V}} - \nabla_\beta \eta\right)^2 \leq \frac{\Omega^*}{T} \mathrm{Var}_\pi\left(\frac{\nabla \mu_u(i)}{\mu_u(i)} \mathrm{V}(j)\right) + \left(\mathbb{E}_\pi\left[\frac{\nabla \mu_u(i)}{\mu_u(i)} \mathrm{A}_\beta(j)\right]\right)^2.$$

Proof.

$$\begin{aligned}
&\mathbb{E}\left(\Delta_T^{\mathrm{V}} - \nabla_\beta \eta\right)^2 \\
&= \mathbb{E}\left(\frac{1}{T}\sum_{t=0}^{T-1} \frac{\nabla \mu_{U_t}(X_t)}{\mu_{U_t}(X_t)} \mathrm{V}(X_{t+1}) - \mathbb{E}_\pi\left[\frac{\nabla \mu_u(i)}{\mu_u(i)}\left(\mathrm{V}(j) + \mathrm{A}_\beta(j)\right)\right]\right)^2 \\
&= \mathbb{E}\left(\frac{1}{T}\sum_{t=0}^{T-1} \frac{\nabla \mu_{U_t}(X_t)}{\mu_{U_t}(X_t)} \mathrm{V}(X_{t+1}) - \mathbb{E}_\pi\left[\frac{\nabla \mu_u(i)}{\mu_u(i)} \mathrm{V}(j)\right]\right)^2 \\
&\quad - 2\left(\mathbb{E}_\pi\left[\frac{\nabla \mu_u(i)}{\mu_u(i)} \mathrm{A}_\beta(j)\right]\right)' \left(\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} \frac{\nabla \mu_{U_t}(X_t)}{\mu_{U_t}(X_t)} \mathrm{V}(X_{t+1}) - \mathbb{E}_\pi\left[\frac{\nabla \mu_u(i)}{\mu_u(i)} \mathrm{V}(j)\right]\right]\right) \\
&\quad + \left(\mathbb{E}\left[\frac{\nabla \mu_u(i)}{\mu_u(i)} \mathrm{A}_\beta(j)\right]\right)^2 \qquad (18) \\
&= \mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1} \frac{\nabla \mu_{U_t}(X_t)}{\mu_{U_t}(X_t)} \mathrm{V}(X_{t+1})\right) + \left(\mathbb{E}_\pi\left[\frac{\nabla \mu_u(i)}{\mu_u(i)} \mathrm{A}_\beta(j)\right]\right)^2 \\
&\leq \frac{\Omega^*}{T} \mathrm{Var}_\pi\left(\frac{\nabla \mu_u(i)}{\mu_u(i)} \mathrm{V}(j)\right) + \left(\mathbb{E}_\pi\left[\frac{\nabla \mu_u(i)}{\mu_u(i)} \mathrm{A}_\beta(j)\right]\right)^2. \qquad (19)
\end{aligned}$$

Note that

$$\mathbb{E}_\pi\left[\frac{\nabla \mu_u(i)}{\mu_u(i)} \mathrm{V}(j)\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} \frac{\nabla \mu_{U_t}(X_t)}{\mu_{U_t}(X_t)} \mathrm{V}(X_{t+1})\right],$$

which means that the second term of Equation (18) is zero, and the first term becomes the variance of the estimate. Equation (19), and hence Theorem 18, follow from Theorem 3. ■

Corollary 19. *Let $D = (\mathcal{S}, \mathcal{U}, P, \mathrm{r}, \mu)$ be a controlled MDP satisfying Assumptions 1, 2 and 3. Let $M = (\mathcal{S}, P)$ be the resultant chain of states, and let π be its stationary distribution; M has mixing*

time τ. *Let* $\{X_t, U_t\}$ *be a process generated by D, starting* $X_0 \sim \pi$. *Then for any* $0 < \varepsilon < e^{-1}$ *there is a* $C_\varepsilon \leq 1 + 50\tau(1+\mathbf{M}) + 8\tau \ln \varepsilon^{-1}$ *such that*

$$\mathbb{E}\left(\Delta_T^{\mathrm{V}} - \nabla_\beta \eta\right)^2 \leq K\varepsilon + \frac{C_\varepsilon}{T}\mathrm{Var}_\pi\left(\frac{\nabla\mu_u(i)}{\mu_u(i)}\mathrm{V}(j)\right) + \left(\mathbb{E}_\pi\left[\frac{\nabla\mu_u(i)}{\mu_u(i)}\mathrm{A}_\beta(j)\right]\right)^2.$$

Proof. Apply Theorem 4 to the first part of Theorem 18, for each of the K dimensions, noting that the mixing time of the process $\{X_t, U_t, X_{t+1}\}$ is at most $\tau\ln(2e) \leq 2\tau$ (Lemma 1). ■

6.2 Zero Variance, Zero Bias Example

Write $v = \mathrm{V} - \mathrm{J}_\beta$. The bias due to using V in place of J_β is given by

$$Gv,$$

where G is a $K \times |\mathcal{S}|$ matrix with its j^{th} column given by $\sum_{i\in\mathcal{S}, u\in\mathcal{U}} \pi_i \nabla\mu_u(i)\mathrm{p}_{ij}(u)$. If v is in the right null space of G then this bias is zero. An example of such a v is a constant vector; $v = (c, c, \ldots, c)'$. This can be used to construct a trivial example of how Δ_T^{V} (Equation (16)) can produce an unbiased, zero variance estimate. The observation that we need only consider value functions that span the range space of G to produce a "good" gradient estimate, in the sense that convergence results may be obtained, was made by Konda and Tsitsiklis (2003, 2000); Sutton et al. (2000). Here we wish to consider a richer class of value functions for the purpose of actively reducing the variance of gradient estimates.

Consider a controlled MDP $D = (\mathcal{S}, \mathcal{U}, P, \mathrm{r}, \mu)$ satisfying Assumptions 1, 2 and 3, and with $\mathrm{r}(i) = (1-\beta)c$, for some constant c, and all $i \in \mathcal{S}$. This gives a value function of $\mathrm{J}_\beta(i) = c$, for all $i \in \mathcal{S}$, and consequently

$$\nabla_\beta \eta = \sum_{i,u} \pi_i \nabla\mu_u(i) c = c \sum_i \pi_i \nabla \sum_u \mu_u(i) = 0.$$

With $v = (-c, -c, \ldots, -c)'$, and selecting the fixed value function $\mathrm{V} = \mathrm{J}_\beta + v$, we have

$$\frac{\nabla\mu_u(i)}{\mu_u(i)}\mathrm{V}(j) = 0, \qquad \forall i, u, j.$$

So Δ_T^{V} will produce a zero bias, zero variance estimate of $\nabla_\beta\eta$. Note also that if the MDP is such that there exists an i,u pair such that $\Pr\{X_t = i, U_t = u\} > 0$ and $\nabla\mu_u(i) \neq 0$ then selecting $\mathrm{V} = \mathrm{J}_\beta$ gives an estimate that, whilst still unbiased, has non-zero variance. The event

$$\left\{\frac{\nabla\mu_u(i)}{\mu_u(i)}\mathrm{V}(j) \neq 0\right\}$$

has a non-zero probability of occurrence.

A less trivial example is given in Appendix D.

7. Algorithms

In Section 5 and Section 6 we have seen bounds on squared error of gradient estimates when using various additive control variates. For the baseline control variates we have also seen the choice of baseline which minimizes this bound. Though it may not be possible to select the best baseline or value function a priori, data could be used to help us choose. For a parameterized baseline, or value function, we could improve the error bounds via gradient decent. In this section we explore this idea.

7.1 Minimizing Weighted Squared Distance to the Optimal Baseline

Given a controlled POMDP and a parameterized class of baseline functions

$$\left\{ \mathrm{b}_{\mathcal{Y}}(\cdot,\omega) : \mathcal{Y} \to \mathbb{R} \,\middle|\, \omega \in \mathbb{R}^L \right\},$$

we wish to choose a baseline function to minimize the variance of our gradient estimates. Theorem 16 expresses this variance as the sum of the optimal variance and a squared distance between the baseline function and the optimal one. It follows that we can minimize the variance of our gradient estimates by minimizing the distance between our baseline and the optimum baseline. The next theorem shows that we can use a sample path of the controlled POMDP to estimate the gradient (with respect to the parameters of the baseline function) of this distance. We need to make the following assumptions about the parameterized baseline functions.

Assumption 6. *There are bounds $\mathbf{M}, \mathbf{G} < \infty$ such that for all $y \in \mathcal{Y}$, and all $\omega \in \mathbb{R}^L$, the baseline function is bounded, $\left|\mathrm{b}_{\mathcal{Y}}(y,\omega)\right| \leq \mathbf{M}$, and the gradient of the baseline is bounded, $\left\|\nabla \mathrm{b}_{\mathcal{Y}}(y,\omega)\right\| \leq \mathbf{G}$.*

We drop ω in the notation, and, to avoid confusion, we write $\mathrm{g}^2(y,u)$ to denote $[(\nabla \mu_u(y))/\mu_u(y)]^2$, where the gradient is with respect to the parameters of the policy, θ.

Theorem 20. *Let $D = (\mathcal{S}, \mathcal{U}, \mathcal{Y}, P, \nu, \mathrm{r}, \mu)$ be a controlled POMDP satisfying Assumptions 1, 2 and 3. Let $\mathrm{b}_{\mathcal{Y}} : \mathcal{Y} \times \mathbb{R}^L \to \mathbb{R}$ be a parameterized baseline function satisfying Assumption 6. If $\{X_t, Y_t, U_t\}$ is a sample path of the controlled POMDP (for any X_0), then with probability* 1

$$\frac{1}{2}\nabla \sigma^2_{\mathcal{Y}}(\mathrm{b}_{\mathcal{Y}}) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \left(\mathrm{b}_{\mathcal{Y}}(Y_{t-1}) - \beta \mathrm{b}_{\mathcal{Y}}(Y_t) - \mathrm{r}(X_t)\right) \sum_{s=0}^{t-1} \beta^{t-s-1} \nabla \mathrm{b}_{\mathcal{Y}}(Y_s) \mathrm{g}^2(Y_s, U_s).$$

Proof. From Theorem 16,

$$\frac{1}{2}\nabla \sigma^2_{\mathcal{Y}}(\mathrm{b}_{\mathcal{Y}}) = \mathbb{E}_\pi \left[\mathrm{g}^2(y,u) \nabla \mathrm{b}_{\mathcal{Y}}(y) \left(\mathrm{b}_{\mathcal{Y}}(y) - \mathrm{b}^*_{\mathcal{Y}}(y) \right) \right],$$

but

$$\begin{aligned}
&\mathbb{E}_\pi \left[\mathrm{g}^2(y,u) \nabla \mathrm{b}_{\mathcal{Y}}(y) \mathrm{b}^*_{\mathcal{Y}}(y) \right] \\
&\quad = \sum_{i,y,u} \pi_i \nu_y(i) \mu_u(y) \mathrm{g}^2(y,u) \nabla \mathrm{b}_{\mathcal{Y}}(y) \\
&\qquad\qquad \times \frac{\sum_{\tilde{\imath},\tilde{u},\tilde{\jmath}} \pi_{\tilde{\imath}} \nu_y(\tilde{\imath}) \mu_{\tilde{u}}(y) \mathrm{p}_{\tilde{\jmath}}(\tilde{u}) \mathrm{g}^2(y,\tilde{u}) \mathrm{J}_\beta(\tilde{\jmath})}{\sum_{\tilde{\imath},\tilde{u}} \pi_{\tilde{\imath}} \nu_y(\tilde{\imath}) \mu_{\tilde{u}}(y) \mathrm{g}^2(y,\tilde{u})} \\
&\quad = \sum_y \nabla \mathrm{b}_{\mathcal{Y}}(y) \sum_{i,u,j} \pi_i \nu_y(i) \mu_u(y) \mathrm{p}_{ij}(u) \mathrm{g}^2(y,u) \mathrm{J}_\beta(j) \\
&\quad = \mathbb{E}_\pi \left[\nabla \mathrm{b}_{\mathcal{Y}}(y) \mathrm{g}^2(y,u) \mathrm{J}_\beta(j) \right].
\end{aligned}$$

Also, as $b_\gamma(Y_t)$ is uniformly bounded, we can write

$$b_\gamma(Y_t) = \sum_{s=t+1}^{\infty} \beta^{s-t-1}(b_\gamma(Y_{s-1}) - \beta b_\gamma(Y_s)).$$

The boundedness of r, and the dominated convergence theorem, likewise gives $J_\beta(X_{t+1}) = \mathbb{E}[\sum_{s=t+1}^{\infty} \beta^{s-t-1} r(X_s)|X_{t+1}]$. Now we have

$$\frac{1}{2}\nabla\sigma^2_\gamma(b_\gamma) = \mathbb{E}_\pi\left[g^2(Y_t,U_t)\nabla b_\gamma(Y_t)\sum_{s=t+1}^{\infty}\beta^{s-t-1}\left(b_\gamma(Y_{s-1}) - \beta b_\gamma(Y_s) - r(X_s)\right)\right]. \tag{20}$$

The rest of the proof is as the proof of Baxter and Bartlett (2001, Theorem 4): we use an ergodicity result to express the expectation as an average, then show that we can truncate the tail of the β decaying sum.

Assume $X_0 \sim \pi$. Write $\tilde{X}_t$ to denote the tuple (X_t, Y_t, U_t), write $\tilde{P}$ to denote the corresponding transition matrix, and write $\tilde{\pi}$ to denote the corresponding stationary distribution (so $\tilde{\pi}_{i,y,u} = \pi_i \nu_y(i)\mu_u(y)$). Now consider running the Markov chain on the process $\{\tilde{X}_t\}$ backwards. We have

$$\Pr\left\{\tilde{X}_{-1}|\tilde{X}_0,\tilde{X}_1,\ldots\right\} = \frac{\Pr\left\{\tilde{X}_{-1},\tilde{X}_0,\tilde{X}_1,\ldots\right\}}{\Pr\left\{\tilde{X}_0,\tilde{X}_1,\ldots\right\}} = \frac{\Pr\left\{\tilde{X}_{-1}\right\}\tilde{P}_{\tilde{X}_{-1}\tilde{X}_0}}{\tilde{\pi}_{\tilde{X}_0}} = \frac{\tilde{\pi}_{\tilde{X}_{-1}}\tilde{P}_{\tilde{X}_{-1}\tilde{X}_0}}{\tilde{\pi}_{\tilde{X}_0}},$$

as $\tilde{\pi}$ is the unique distribution such that $\tilde{\pi}'\tilde{P} = \tilde{\pi}'$. This gives the distribution for $\tilde{X}_{-1}$, and repeating this argument gives the distribution for $\tilde{X}_{-2}, \tilde{X}_{-3}, \ldots$. Denote this doubly infinite process by $\{\tilde{X}_t\}_{-\infty}^{\infty}$. We wish to look at the behavior of time averages of the function

$$f\left(\left\{\tilde{X}_t\right\}_{-\infty}^{\infty}\right) \stackrel{\text{def}}{=} g^2(Y_0,U_0)\nabla b_\gamma(Y_0)\sum_{s=1}^{\infty}\beta^{s-1}\left(b_\gamma(Y_{s-1}) - \beta b_\gamma(Y_s) - r(X_s)\right).$$

Specifically, we would like to show that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{m=0}^{T-1} f\left(\mathfrak{S}^m\left(\left\{\tilde{X}_t\right\}_{-\infty}^{\infty}\right)\right) = \mathbb{E}\left[f\left(\left\{\tilde{X}_t\right\}_{-\infty}^{\infty}\right)\right], \qquad \text{w.p.1} \tag{21}$$

where $\mathfrak{S}^m$ denotes m applications of the shift operator $\mathfrak{S}$, and where $\mathfrak{S}(\{\tilde{X}_t\}_{-\infty}^{\infty}) = \{W_t\}_{-\infty}^{\infty}$ with $W_t = \tilde{X}_{t+1}$ for all t. Doob (1994, L^2 Ergodic theorem, pg. 119) tells us, provided that $\mathfrak{S}$ is one-to-one and measure preserving, and that f is square integrable, the left hand side of Equation (21) is almost surely constant, and furthermore, provided that the only invariant sets of $\mathfrak{S}$ are sets of measure zero and their complements, this constant is equal to the right hand side of Equation (21). Expanding f and $\mathfrak{S}$ in Equation (21) then gives, with probability one,

$$\frac{1}{2}\nabla\sigma^2_\gamma(b_\gamma) = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1} g^2(Y_t,U_t)\nabla b_\gamma(Y_t)\sum_{s=t+1}^{\infty}\beta^{s-t-1}\left(b_\gamma(Y_{s-1}) - \beta b_\gamma(Y_s) - r(X_s)\right). \tag{22}$$

It remains to be shown that the conditions of the L^2 Ergodic theorem hold.

1. $\mathfrak{S}$ *is one-to-one.* By considering how $\mathfrak{S}$ behaves at each index, we see that it is a bijection.

2. $\mathfrak{S}$ *is measure preserving.* That is, for a set of sequences A, A and $\mathfrak{S}(A)$ have the same measure. This follows from the Markov property, and from the fact that the transition operator at time t, as well as the marginal distribution on X_t, is identical for all times t. Specifically, we have the following. Let $A^- \subset A$ be the smallest set such that $\Pr\{\{W_t\}_{-\infty}^{\infty} \in A\} = \Pr\{\{W_t\}_{-\infty}^{\infty} \in A^- \cap A\}$, and write $A_s = \{W_s : \{W_t\}_{-\infty}^{\infty} \in A^-\}$ and $A_{s_1}^{s_2} = \{\{W_t\}_{s_1}^{s_2} : \{W_t\}_{-\infty}^{\infty} \in A^-\}$, where $\{W_t\}_{s_1}^{s_2}$ is the process starting at $t = s_1$ and ending at $t = s_2$. For any s we have

$$\begin{aligned}
\Pr\left\{\mathfrak{S}\left(\{\tilde{X}_t\}_{-\infty}^{\infty}\right) \in A\right\} &= \int_{x \in A_s} \Pr\{\tilde{X}_{s+1} = x\} \Pr\left\{\{\tilde{X}_t\}_{s+2}^{\infty} \in A_{s+1}^{\infty} \middle| \tilde{X}_{s+1} = x\right\} \\
&\qquad \times \Pr\left\{\{\tilde{X}_t\}_{-\infty}^{s} \in A_{-\infty}^{s-1} \middle| \tilde{X}_{s+1} = x\right\} \mathfrak{C}(dx) \\
&= \int_{x \in A_s} \Pr\{\tilde{X}_s = x\} \Pr\left\{\{\tilde{X}_t\}_{s+1}^{\infty} \in A_{s+1}^{\infty} \middle| \tilde{X}_s = x\right\} \\
&\qquad \times \Pr\left\{\{\tilde{X}_t\}_{-\infty}^{s-1} \in A_{-\infty}^{s-1} \middle| \tilde{X}_s = x\right\} \mathfrak{C}(dx) \\
&= \Pr\left\{\{\tilde{X}_t\}_{-\infty}^{\infty} \in A\right\}.
\end{aligned}$$

   We also have that $\mathfrak{S}^{-1}$ (the inverse of $\mathfrak{S}$) is measure preserving; by the change of variables $\{W_t\}_{-\infty}^{\infty} = \mathfrak{S}(\{\tilde{X}_t\}_{-\infty}^{\infty})$.

3. f *is square integrable.* The measure on $\{\tilde{X}_t\}_{-\infty}^{\infty}$ is finite, and $|\mathrm{f}|$ is bounded.

4. *If set A is such that $\mathfrak{S}^{-1}(A) = A$ (where $\mathfrak{S}^{-1}(A) = \{\{W_t\}_{-\infty}^{\infty} : \mathfrak{S}(\{W_t\}_{-\infty}^{\infty}) \in A\}$), then either A has measure zero, or $\bar{A}$ has measure one.* Consider a set A of positive measure such that $\mathfrak{S}^{-1}(A) = A$, and write $\bar{A}$ for its complement. As $\mathfrak{S}$ is a bijection, we also have that $\mathfrak{S}^{-1}(\bar{A}) = \bar{A}$. Assumption 1 implies that $A_t = \tilde{S}$ (at least, this is true for the state component, and, without loss of generality, we may assume it is true for the extended space). If we assume that $A_0 \cap \bar{A}_0 = \emptyset$, and hence $A_t \cap \bar{A}_t = \emptyset$ for all t, then the measure of $\bar{A}$ must be zero. We will show that $A_0 \cap \bar{A}_0 = \emptyset$.

   Unless $\bar{A}$ has measure zero, for each $x \in A_0 \cap \bar{A}_0$ we must have that $\Pr\{\{W_t\}_{-\infty}^{-1} \in \bar{A}_{-\infty}^{-1} | W_0 = x\}$ and $\Pr\{\{W_t\}_1^{\infty} \in A_1^{\infty} | W_0 = x\}$ are both positive, by the Markov property. Hence if $A_0 \cap \bar{A}_0$ is non-empty there must be a set of positive measure, which we denote B, that follows sequences in $\bar{A}^-$ until time $t = 0$, and then follows sequences in A^-. Without loss of generality, let us assume that $B \subset A^-$. We will also assume that if $b \in B$ then $\mathfrak{S}^{-1}(b) \in B$, as the existence of B implies the existence of $\hat{B} = B \cup \{\mathfrak{S}^{-1}(b)\}$ with the same properties. We will show that such a B does not exist, and therefore $A_0 \cap \bar{A}_0 = \emptyset$.

   Let $A_{-\infty}^{s*} = \{\{W_t\}_{-\infty}^{\infty} : \{W_t\}_{-\infty}^{s} \in A_{-\infty}^{s}\}$, the set of sequences that follow A^- until time s, and then follow any sequence. We have that $B_{-\infty}^{0*} \subset \bar{A}_{-\infty}^{0*}$. Construct the set $B^* \subset B$ by $B^* = \lim_{t\to\infty} \mathfrak{S}^{-t}(B)$ (note, $\mathfrak{S}^{-t}(B)$ is a non-increasing sequence of sets, and hence its limit exists). We have that $\mathfrak{S}^{-t}(B) \subset \mathfrak{S}^{-t}(B)_{-\infty}^{t*} = \mathfrak{S}^{-t}(B_{-\infty}^{0*}) \subset \mathfrak{S}^{-t}(\bar{A}_{-\infty}^{0*}) = \bar{A}_{-\infty}^{t*}$, and so $B^* = \liminf_{t\to\infty} \mathfrak{S}^{-t}(B) \subset \limsup_{t\to\infty} \bar{A}_{-\infty}^{t*} = \bar{A}^-$, where the last equality follows from $\bar{A}_{-\infty}^{t*}$ being non-increasing. Furthermore, by the dominated convergence theorem we have $\Pr\{\{W_s\}_{-\infty}^{\infty} \in B^*\} = \lim_{t\to\infty} \Pr\{\{W_s\}_{-\infty}^{\infty} \in \mathfrak{S}^{-t}(B)\} = \Pr\{\{W_s\}_{-\infty}^{\infty} \in B\} > 0$. This means that the set B^* has positive measure and is a subset of both A and $\bar{A}$, which is impossible, and so such a B does not exist.

The statement given by Equation (21) is for a sample such that $X_0 \sim \pi$, but can be generalized to an arbitrary distribution using the convergence of $\{X_t\}$ to stationarity. Indeed, for the finite chains we consider, all states have positive π-measure, and hence Equation (22) holds for $X_0 \sim \pi$ only if it holds for all $X_0 \in \mathcal{S}$.

If we truncate the inner sum at T, the norm of the error is

$$\begin{aligned}
&\left\| \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathrm{g}^2(Y_t,U_t)\nabla \mathrm{b}_{\mathcal{Y}}(Y_t) \sum_{s=T+1}^{\infty} \beta^{s-t-1}\left(\mathrm{b}_{\mathcal{Y}}(Y_{s-1}) - \beta \mathrm{b}_{\mathcal{Y}}(Y_s) - \mathrm{r}(X_s)\right)\right\| \\
&= \left\| \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathrm{g}^2(Y_t,U_t)\nabla \mathrm{b}_{\mathcal{Y}}(Y_t)\beta^T \left(\mathrm{b}_{\mathcal{Y}}(Y_T) - \sum_{s=T+1}^{\infty} \beta^{s-t-1}\mathrm{r}(X_s)\right)\right\| \\
&\leq \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{B}^2\mathbf{G}\beta^T \left(\mathbf{M} + \frac{\mathbf{R}}{1-\beta}\right) \\
&= 0,
\end{aligned}$$

where we have used Assumptions 2, 3, and 6. This gives

$$\frac{1}{2}\nabla \sigma^2_{\mathcal{Y}}(\mathrm{b}_{\mathcal{Y}}) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathrm{g}^2(Y_t,U_t)\nabla \mathrm{b}_{\mathcal{Y}}(Y_t) \sum_{s=t+1}^{T} \beta^{s-t-1}\left(\mathrm{b}_{\mathcal{Y}}(Y_{s-1}) - \beta \mathrm{b}_{\mathcal{Y}}(Y_s) - \mathrm{r}(X_s)\right),$$

and changing the order of summation gives the result. ■

Theorem 20 suggests the use of Algorithm 1 to compute the gradient of $\sigma^2_{\mathcal{Y}}(\mathrm{b}_{\mathcal{Y}})$ with respect to the parameters of the baseline function $\mathrm{b}_{\mathcal{Y}}$. The theorem implies that, as the number of samples T gets large, the estimate produced by this algorithm approaches the true gradient.

Algorithm 1 Compute estimate of gradient of distance to optimal baseline

given

- A controlled POMDP $(\mathcal{S}, \mathcal{U}, \mathcal{Y}, P, \nu, \mathrm{r}, \mu)$.
- The sequence of states, observations and controls generated by the controlled POMDP, $\{i_0, y_0, u_0, i_1, y_1, \ldots, i_{T-1}, y_{T-1}, u_{T-1}, i_T, y_T\}$.
- A parameterized baseline function $\mathrm{b}_{\mathcal{Y}} : \mathcal{Y} \times \mathbb{R}^L \to \mathbb{R}$.

write $\mathrm{g}^2(y,u)$ to denote $[(\nabla \mu_u(y))/\mu_u(y)]^2$.
set $z_0 = 0$ ($z_0 \in \mathbb{R}^L$), $\Delta_0 = 0$ ($\Delta_0 \in \mathbb{R}^L$)
for all $\{i_t, y_t, u_t, i_{t+1}, y_{t+1}\}$ **do**
  $z_{t+1} = \beta z_t + \nabla \mathrm{b}_{\mathcal{Y}}(y_t,\omega)\mathrm{g}^2(y_t,u_t)$
  $\Delta_{t+1} = \Delta_t + \frac{1}{t+1}\left(\left(\mathrm{b}_{\mathcal{Y}}(y_t,\omega) - \beta \mathrm{b}_{\mathcal{Y}}(y_{t+1},\omega) - \mathrm{r}(x_{t+1})\right) z_{t+1} - \Delta_t\right)$
end for

In Bartlett and Baxter (2002) a variant of the GPOMDP algorithm is shown to give an estimate that, in finite time and with high probability, is close to $\nabla_\beta \eta$. A similar analysis could be performed for the estimate produced by Algorithm 1, in particular, we could replace $\nabla_t = (\nabla \mu_{U_t}(Y_t))/\mu_{U_t}(U_t)$ and $R_t = \mathrm{r}(X_t)$ in Bartlett and Baxter (2002) with $\tilde{\nabla}_t = \mathrm{g}^2(Y_t,U_t)\nabla \mathrm{b}_{\mathcal{Y}}(Y_t)$ and $\tilde{R}_t = \mathrm{b}_{\mathcal{Y}}(Y_{t-1}) - \beta \mathrm{b}_{\mathcal{Y}}(Y_t) - \mathrm{r}(X_t)$. Notice that ∇_t and R_t occur in precisely the same way in GPOMDP to produce an

estimate of

$$\mathbb{E}\left[\nabla_t \sum_{s=t+1}^{\infty} \beta^{s-t-1} R_s\right]$$

as $\tilde{\nabla}_t$ and $\tilde{R}_t$ occur in Algorithm 1 to produce an estimate of

$$\mathbb{E}\left[\tilde{\nabla}_t \sum_{s=t+1}^{\infty} \beta^{s-t-1} \tilde{R}_s\right].$$

Algorithm 2 gives an online version of Algorithm 1. The advantage of such an algorithm is that the baseline may be updated whilst the estimate of the performance gradient is being calculated. Such a strategy for updates would, however, affect the convergence of performance gradient estimates (for constant baselines this may be avoided, see Section 8.2). The question of the convergence of Algorithm 2, and the convergence of performance gradient estimates in the presence of online baseline updates, is not addressed in this paper; though simulations in Sections 8.2 and 8.3 show that performing such online baseline updates can give improvements.

Algorithm 2 Online version of Algorithm 1

given

- A controlled POMDP $(\mathcal{S}, \mathcal{U}, \mathcal{Y}, P, \nu, r, \mu)$.
- The sequence of states, observations and controls generated by the controlled POMDP, $\{i_0, y_0, u_0, i_1, y_1, \ldots, i_{T-1}, y_{T-1}, u_{T-1}, i_T, y_T\}$.
- A parameterized baseline function $b_{\mathcal{Y}} : \mathcal{Y} \times \mathbb{R}^L \to \mathbb{R}$.
- A sequence of step sizes, γ_t

write $g^2(y,u)$ to denote $[(\nabla \mu_u(y))/\mu_u(y)]^2$.
set $z_0 = 0$ $(z_0 \in \mathbb{R}^L)$
for all $\{i_t, y_t, u_t, i_{t+1}, y_{t+1}\}$ **do**
$z_{t+1} = \beta z_t + \nabla b_{\mathcal{Y}}(y_t, \omega_t) g^2(y_t, u_t)$
$\omega_{t+1} = \omega_t - \gamma_t \left(b_{\mathcal{Y}}(y_t, \omega_t) - \beta b_{\mathcal{Y}}(y_{t+1}, \omega_t) - r(x_{t+1})\right) z_{t+1}$
end for

7.2 Minimizing Bound on Squared Error when using a Value Function

Given a controlled MDP and a parameterized class of value functions,

$$\left\{ V(\cdot, \omega) : \mathcal{S} \to \mathbb{R} | \, \omega \in \mathbb{R}^L \right\},$$

we wish to choose a value function to minimize the expected squared error of our gradient estimates. Theorem 18 gives a bound on this error,

$$E_T = \frac{\Omega^*}{T} \mathrm{Var}_\pi\left(\frac{\nabla \mu_u(i)}{\mu_u(i)} V(j, \omega)\right) + \left(\mathbb{E}_\pi\left[\frac{\nabla \mu_u(i)}{\mu_u(i)} A_\beta(j, \omega)\right]\right)^2.$$

We drop ω in the notation, and write $g(i,u)$ to denote $(\nabla \mu_u(i))/\mu_u(i)$, where the gradient is with respect to the policy parameters, θ. We can compute the gradient of this bound:

$$\nabla \frac{1}{2} E_T \quad = \quad \nabla \frac{1}{2}\left(\frac{\Omega^*}{T} \mathrm{Var}_\pi(g(i,u) V(j)) + \left(\mathbb{E}_\pi\left[g(i,u) A_\beta(j)\right]\right)^2\right)$$

$$
\begin{aligned}
&= \frac{1}{2}\left(\frac{\Omega^*}{T}\nabla\mathbb{E}_\pi\left[(\mathrm{g}(i,u)\mathrm{V}(j))^2\right] - \frac{\Omega^*}{T}\nabla\left(\mathbb{E}_\pi\left[\mathrm{g}(i,u)\mathrm{V}(j)\right]\right)^2 + \nabla\left(\mathbb{E}_\pi\left[\mathrm{g}(i,u)\mathrm{A}_\beta(j)\right]\right)^2\right)\\
&= \left(\frac{\Omega^*}{T}\mathbb{E}_\pi\left[(\mathrm{g}(i,u)\mathrm{V}(j))'\left(\mathrm{g}(i,u)\left(\nabla\mathrm{V}(j)\right)'\right)\right]\right.\\
&\quad - \frac{\Omega^*}{T}\left(\mathbb{E}_\pi\left[\mathrm{g}(i,u)\mathrm{V}(j)\right]\right)'\left(\mathbb{E}_\pi\left[\mathrm{g}(i,u)\left(\nabla\mathrm{V}(j)\right)'\right]\right)\\
&\quad \left. - \left(\mathbb{E}_\pi\left[\mathrm{g}(i,u)\mathrm{A}_\beta(j)\right]\right)'\left(\mathbb{E}_\pi\left[\mathrm{g}(i,u)\left(\nabla\mathrm{V}(j)\right)'\right]\right)\right). \qquad (23)
\end{aligned}
$$

This gradient can be estimated from a single sample path of the controlled MDP, $\{X_s, U_s\}$. We need the following assumption on the value function.

Assumption 7. *There are bounds $\mathbf{M}, \mathbf{G} < \infty$ such that for all $i \in \mathcal{S}$, and all $\omega \in \mathbb{R}^L$, the value function is bounded, $|\mathrm{V}(i,\omega)| \leq \mathbf{M}$, and the gradient of the value function is bounded, $\|\nabla\mathrm{V}(i,\omega)\| \leq \mathbf{G}$.*

Algorithm 3 gives an estimate of (23) from a sample path of the controlled MDP; constructing this estimate from the following four estimations:

$$
\begin{aligned}
\Delta A_S &= \frac{1}{S}\sum_{s=0}^{S-1}\left(\mathrm{g}(X_s,U_s)\mathrm{V}(X_{s+1})\right)'\left(\mathrm{g}(X_s,U_s)\left(\nabla\mathrm{V}(X_{s+1})\right)'\right) && \in \mathbb{R}^{1\times L};\\
\Delta B_S &= \frac{1}{S}\sum_{s=0}^{S-1}\mathrm{g}(X_s,U_s)\mathrm{V}(X_{s+1}) && \in \mathbb{R}^{K};\\
\Delta C_S &= \frac{1}{S}\sum_{s=0}^{S-1}\left(\mathrm{r}(X_{s+1}) + \beta\mathrm{V}(X_{s+2}) - \mathrm{V}(X_{s+1})\right)z_{s+1} && \in \mathbb{R}^{K};\\
\Delta D_S &= \frac{1}{S}\sum_{s=0}^{S-1}\mathrm{g}(X_s,U_s)\left(\nabla\mathrm{V}(X_{s+1})\right)' && \in \mathbb{R}^{K\times L},
\end{aligned}
$$

where $z_0 = 0$ and $z_{s+1} = \beta z_s + \mathrm{g}(X_s,U_s)$. The estimate of the gradient then becomes

$$
\Delta_S = \left(\frac{\Omega^*}{T}\Delta A_S - \frac{\Omega^*}{T}\Delta B_S'\Delta D_S - \Delta C_S'\Delta D_S\right)'.
$$

Notice that $\Delta A_S, \Delta B_S$, and ΔD_S are simply sample averages (produced by the Markov chain) estimating the relevant expectations in Equation (23). We see from Theorem 3 and Corollary 5 that the variance of these estimates are $O(\ln(S)/S)$, giving swift convergence. By noting the similarity between the expectation in Equation (20) and the expectation estimated by ΔC_S, we see that the ergodicity and truncation arguments of Theorem 20, and the convergence discussion following, also hold for the ΔC_S estimate.

An online implementation is complicated by the multiplication of expectations. The online algorithm (Algorithm 4) uses a decaying window of time (normalized for the rate of decay) in the calculation of the expectations.

Algorithm 3 Compute estimate of gradient of squared error wrt value function parameter

given

- A controlled POMDP $(\mathcal{S}, \mathcal{U}, \mathcal{Y}, P, \nu, \mathrm{r}, \mu)$.
- The sequence of states, observations and controls generated by the controlled POMDP, $\{i_0, u_0, i_1, \ldots, i_S, u_S, i_{S+1}\}$.
- A parameterized value function $\mathrm{V} : \mathcal{S} \times \mathbb{R}^L \to \mathbb{R}$.

write $\mathrm{g}(i,u)$ to denote $(\nabla \mu_u(i))/\mu_u(i)$.
set $z_0 = 0$ ($z_0 \in \mathbb{R}^K$), $\Delta A_0 = 0$ ($\Delta A_0 \in \mathbb{R}^L$), $\Delta B_0 = 0$ ($\Delta B_0 \in \mathbb{R}^K$), $\Delta C_0 = 0$ ($\Delta C_0 \in \mathbb{R}^K$) and $\Delta D_0 = 0$ ($\Delta D_0 \in \mathbb{R}^{K \times L}$)
for all $\{i_s, u_s, i_{s+1}, i_{s+2}\}$ **do**

$$z_{s+1} = \beta z_s + \mathrm{g}(i_s, u_s)$$
$$\Delta A_{s+1} = \Delta A_s + \tfrac{1}{s+1}\left((\mathrm{g}(i_s,u_s)\mathrm{V}(i_{s+1}))' \left(\mathrm{g}(i_s,u_s)\left(\nabla \mathrm{V}(i_{s+1})\right)'\right) - \Delta A_s\right)$$
$$\Delta B_{s+1} = \Delta B_s + \tfrac{1}{s+1}\left(\mathrm{g}(i_s,u_s)\mathrm{V}(i_{s+1}) - \Delta B_s\right)$$
$$\Delta C_{s+1} = \Delta C_s + \tfrac{1}{s+1}\left(\left(\mathrm{r}(i_{s+1}) + \beta \mathrm{V}(i_{s+2}) - \mathrm{V}(i_{s+1})\right) z_{s+1} - \Delta C_s\right)$$
$$\Delta D_{s+1} = \Delta D_s + \tfrac{1}{s+1}\left(\mathrm{g}(i_s,u_s)\left(\nabla \mathrm{V}(i_{s+1})\right)' - \Delta D_s\right)$$

end for

$$\Delta_S = \left(\tfrac{\Omega^*}{T}\Delta A_S - \tfrac{\Omega^*}{T}\Delta B_S' \Delta D_S - \Delta C_S' \Delta D_S\right)'$$

Algorithm 4 Online version of Algorithm 3

given

- A controlled POMDP $(\mathcal{S}, \mathcal{U}, \mathcal{Y}, P, \nu, \mathrm{r}, \mu)$.
- The sequence of states, observations and controls generated by the controlled POMDP, $\{i_0, u_0, i_1, \ldots, i_S, u_S, i_{S+1}\}$.
- $\alpha \in \mathbb{R}$, $0 < \alpha < 1$
- A sequence of step sizes, γ_s
- A parameterized value function $\mathrm{V} : \mathcal{S} \times \mathbb{R}^L \to \mathbb{R}$.

write $\mathrm{g}(i,u)$ to denote $(\nabla \mu_u(i))/\mu_u(i)$.
set $z_0 = 0$ ($z_0 \in \mathbb{R}^K$), $\Delta A_0 = 0$ ($\Delta A_0 \in \mathbb{R}^L$), $\Delta B_0 = 0$ ($\Delta B_0 \in \mathbb{R}^K$), $\Delta C_0 = 0$ ($\Delta C_0 \in \mathbb{R}^K$) and $\Delta D_0 = 0$ ($\Delta D_0 \in \mathbb{R}^{K \times L}$)
for all $\{i_s, u_s, i_{s+1}, i_{s+2}\}$ **do**

$$z_{s+1} = \beta z_s + \mathrm{g}(i_s, u_s)$$
$$\Delta A_{s+1} = \alpha \Delta A_s + (\mathrm{g}(i_s,u_s)\mathrm{V}(i_{s+1}))' \left(\mathrm{g}(i_s,u_s)\left(\nabla \mathrm{V}(i_{s+1})\right)'\right)$$
$$\Delta B_{s+1} = \alpha \Delta B_s + \mathrm{g}(i_s,u_s)\mathrm{V}(i_{s+1})$$
$$\Delta C_{s+1} = \alpha \Delta C_s + \left(\mathrm{r}(i_{s+1}) + \beta \mathrm{V}(i_{s+2}) - \mathrm{V}(i_{s+1})\right) z_{s+1}$$
$$\Delta D_{s+1} = \alpha \Delta D_s + \mathrm{g}(i_s,u_s)\left(\nabla \mathrm{V}(i_{s+1})\right)'$$
$$\omega_{s+1} = \omega_s - \gamma_s \left(\left(\tfrac{1-\alpha}{1-\alpha^{s+1}}\right)\tfrac{\Omega^*}{T}\Delta A_S - \left(\tfrac{1-\alpha}{1-\alpha^{s+1}}\right)^2 \tfrac{\Omega^*}{T}\Delta B_S' \Delta D_S - \left(\tfrac{1-\alpha}{1-\alpha^{s+1}}\right)^2 \Delta C_S' \Delta D_S\right)'$$

end for

7.3 Minimizing the Bias Error when using a Value Function

As the number of steps T gets large, the error E_T of the gradient estimate becomes proportional to the square of the bias error,

$$E_\infty = \left(\mathbb{E}_\pi\left[\mathrm{g}(i,u)\mathrm{A}_\beta(j)\right]\right)^2.$$

The gradient of this quantity, with respect to the parameters of the value function, can be computed using Algorithm 3, with $\Omega^*/T = 0$. In this case, only ΔC_s and ΔD_s need to be computed.

7.4 Minimizing Bound on Sample Error when using a Value Function

A more restrictive approach is to minimize the error seen at each sample,

$$R = \mathbb{E}_\pi\left(\mathrm{g}(i,u)\mathrm{A}_\beta(j)\right)^2.$$

This approach directly drives V towards J_β and as such does not aim for additional beneficial correlation. It produces an algorithm that is very similar to TD Sutton (1988), but has the benefit that the relative magnitude of the gradient with respect to the policy parameters is taken into account. In this way, more attention is devoted to accuracy in regions of the state space with large gradients.

For a parameterized class of value functions, $\left\{\mathrm{V}(\cdot,\omega) : \mathcal{S} \to \mathbb{R} | \, \omega \in \mathbb{R}^L\right\}$, we can determine the gradient of this quantity.

$$\begin{aligned}
\nabla\frac{1}{2}R &= \nabla\frac{1}{2}\mathbb{E}_\pi\left(\mathrm{g}(i,u)\mathrm{A}_\beta(j)\right)^2 \\
&= -\mathbb{E}_\pi\left[\left(\mathrm{g}(i,u)\left(\nabla\mathrm{V}(j)\right)'\right)'\left(\mathrm{g}(i,u)\mathrm{A}_\beta(j)\right)\right] \\
&= -\mathbb{E}_\pi\left[\left(\mathrm{g}(i,u)\right)^2\nabla\mathrm{V}(j)\mathrm{A}_\beta(j)\right].
\end{aligned}$$

If the value function satisfies Assumption 7, the gradient may be estimated by a single sample path from a controlled MDP. The ergodicity and truncation argument is the same as that in the proof of Theorem 20.

$$\Delta R_T = \frac{1}{T}\sum_{t=1}^{T}\left(\mathrm{r}(X_t) + \beta\mathrm{V}(X_{t+1}) - \mathrm{V}(X_t)\right)z_t,$$

where $z_0 = 0$, and $z_{t+1} = \beta z_t + \left(\mathrm{g}(X_t,U_t)\right)^2\nabla\mathrm{V}(X_{t+1})$.

8. Simulation Examples

This section describes some experiments performed in simulated environments. First, the estimates suggested by Sections 5 and 6 are tested in a simple, simulated setting. This simple setting is then used to test the algorithms of Section 7. Finally, a larger, target-tracking setting is used to test a number of gradient estimates at various stages of the learning process.

8.1 Three State MDP, using Discounted Value Function

This section describes experiments comparing choices of control variate for a simple three state MDP. The system is the described in detail in Baxter et al. (2001). The gradient $\nabla\eta$ was compared to the gradient estimates produced with a variety of schemes: GPOMDP without any control variate; a constant baseline set to $\mathbb{E}_\pi\mathrm{J}_\beta(i)$; the optimum constant baseline, described in Theorem 11; the

optimum baseline function, described in Corollary 15; and a value function that was trained using Algorithm 3 with Ω^*/T set to 0.001. This value function had a distinct parameter for each state, all initially set to zero.

Because of its simplicity, a number of quantities can be computed explicitly, including the true gradient $\nabla\eta$, the discounted value function J_β, the expectation of the discounted value function, the optimal baseline, and the optimal constant baseline. All experiments used the precomputed discounted value function in their $\nabla_\beta\eta$ estimations rather than the discounted sum of future rewards, an estimate of the discounted value function. For each experiment, the data was collected over 500 independent runs, with $\beta = 0.95$.

Figures 2 and 3 plot the mean and standard deviation (respectively) of the relative norm difference of the gradient estimate from $\nabla\eta$, as a function of the number of time steps. The relative norm difference of a gradient estimate Δ from the true gradient $\nabla\eta$ is given by

$$\frac{\|\Delta - \nabla\eta\|}{\|\nabla\eta\|}.$$

It is clear from the figures that the use of these control variates gives significant variance reductions over GPOMDP. It is also clear that the optimum baseline gives better performance than the use of the expectation of the discounted value function as a baseline. For this MDP, the performance difference between the optimum baseline and the optimum constant baseline is small; the optimum baseline of this system, $b^*_{\mathcal{Y}} = (6.35254, 6.35254, 6.26938)'$, is close to a constant function. The optimum constant baseline is $b^* = 6.33837$.

Since the value of Ω^*/T was fixed when optimizing the value function, the asymptotic error of its associated gradient estimate is non-zero, as Figure 2 shows. However, the expected error remains smaller than that of GPOMDP for all but very large values of T, and the standard deviation is always smaller.

8.2 Online Training

Instead of precomputing the optimum baseline, and pretraining the value function, they could be learned online, whilst estimating $\nabla_\beta\eta$. Figures 4 and 5 show experiments on the same three state MDP as in Section 8.1, but here the baseline and value function were learned online, using Algorithm 2 and Algorithm 4 respectively. GPOMDP and baseline plots were over 500 independent runs, the value function plots were over 1000 independent runs. A β value of 0.95 was used, and the online training step size γ_t was set to $1/\ln(1+t)$. For the value function, Ω^*/T was set to 0.01 and α was set to 0.99. The baseline and the value function had a parameter for each state and were initially set to zero.

It is clear from the figures that the online baseline algorithm gives a significant improvement from the GPOMDP algorithm. Looking at the error using the online value function algorithm we see a performance increase over GPOMDP until T becomes large.

Note that the baseline, when trained online, is non-stationary, and the gradient estimate becomes

$$\Delta = \frac{1}{T}\sum_{t=0}^{T-1} \frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)} \left(\sum_{s=t+1}^{T} \beta^{s-t-1} r(X_s) - b_t(Y_t) \right).$$

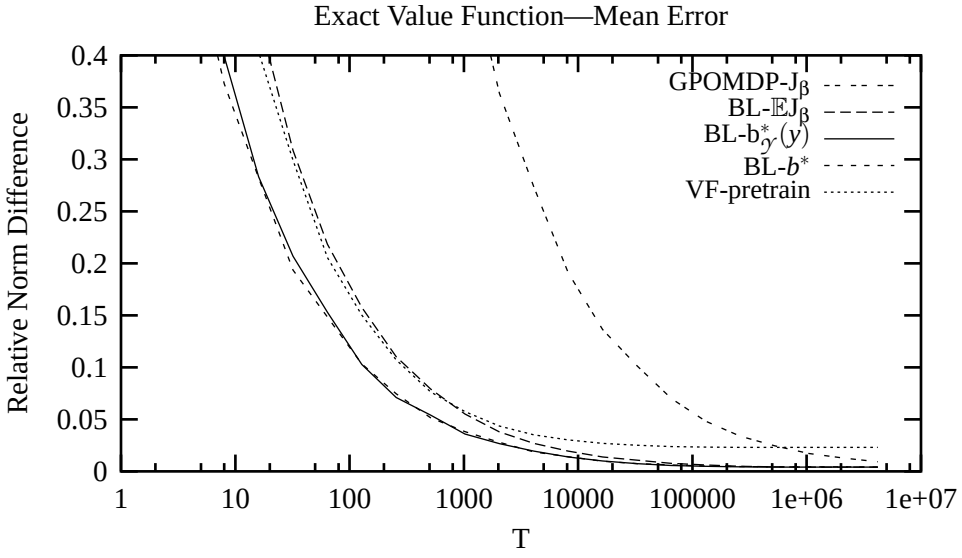


Figure 2: The mean of the relative norm difference from $\nabla\eta$: using no control variate (GPOMDP-J_β); using the expected discounted value function as a baseline (BL-$\mathbb{E}J_\beta$); using the optimum baseline (BL-$b^*_\gamma(y)$); using the optimal constant baseline (BL-b^*); and using a pretrained value function (VF-pretrained). In all cases the explicitly calculated discounted value function was used in place of the estimates J_t (except, of course, for the pretrained value function case, where the value function is used in place of the estimates J_t.)

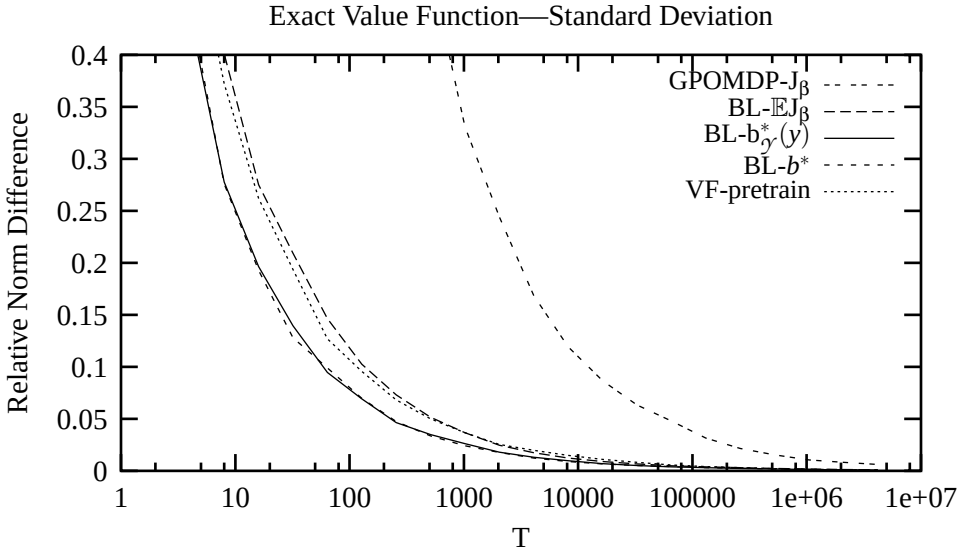


Figure 3: The standard deviation of the relative norm difference from $\nabla\eta$ (see Figure 2 for an explanation of the key).

This non-stationarity could mean an additional bias in the estimate, though we can see by the graphs that, at least in this case, this additional bias is small. The estimate that we actually use is

$$\Delta = \frac{1}{T}\sum_{t=1}^{T} z_t \left(\mathrm{r}(X_t) - (\mathrm{b}_{t-1}(Y_{t-1}) - \beta \mathrm{b}_{t-1}(Y_t))\right),$$

where z_t, the eligibility trace, is given by $z_0 = 0$ and $z_{t+1} = \beta z_t + (\nabla \mu_{U_t}(Y_t))/\mu_{U_t}(Y_t)$. One might argue that this additionally correlates our baseline with any errors due to the truncation of the sum of discounted future rewards. This should make little difference, except for small T; we have seen that, for the modified estimate $\Delta_T^{(+S)}(\mathrm{b}_\mathcal{Y})$, any influence this error has is exponentially decreasing.

Note that for any constant baseline we need not worry about non-stationarity, as we have

$$\sum_{t=1}^{T} z_t (b_T - \beta b_T) = \left(\sum_{t=1}^{T} z_t\right)(1-\beta) b_T,$$

so by additionally keeping track of $\Sigma_T = \sum_{t=1}^{T} z_t$ we have the estimate, at time T,

$$\frac{1}{T}\sum_{t=1}^{T} z_t \left(\mathrm{r}(X_t) - (b_T - \beta b_T)\right) = \frac{1}{T}\sum_{t=1}^{T} z_t \mathrm{r}(X_t) - \frac{1}{T}\Sigma_T (1-\beta) b_T,$$

an unbiased estimate of $\nabla_\beta \eta$; again, treating the error due to the truncation of the discounted sum of future rewards as negligible.

8.3 Locating a Target

These experiments deal with the task of a puck, moving in a plane, learning to locate a target. The puck had unit mass, 0.05 unit radius, and was controlled by applying a 5 unit force in either the positive or negative x direction and either the positive or negative y direction. The puck moved within a 5×5 unit area with elastic walls and a coefficient of friction of 0.0005; gravity being set to 9.8. The simulator worked at a granularity of $1/100$ of a second with controller updates at every $1/20$ of a second. The distance between the puck and the target location was given as a reward at each update time. Every 30 seconds this target and the puck was set to a random location, and the puck's x and y velocities set randomly in the range $[-10, 10]$.

The puck policy was determined by a neural network with seven inputs, no hidden layer, and four outputs; the outputs computing a tanh squashing function. The inputs to the controller were: the x and y location of the puck, scaled to be in $[-1, 1]$; the x and y location of the puck relative to the target, scaled by the dimension sizes; the velocity of the puck, scaled such that a speed of 10 units per second was mapped to a value of 1; and a constant input of 1 to supply an offset. The outputs of the neural network gave a weighting, $\xi_i \in (0, 1)$, to each of the (x, y) thrust combinations: $(-5, -5)$; $(-5, 5)$; $(5, -5)$; and $(5, 5)$. So, collating the seven inputs in the vector v, we have

$$\xi_i = \mathrm{sqsh}\left(\sum_{k=1}^{7} \theta_{i,k} v_k\right), \qquad i \in \{1, 2, 3, 4\},$$

where θ is a vector of 28 elements, one element, $\theta_{i,k}$, for each i, k pair, and the squashing function is $\mathrm{sqsh}(x) = (1 + \tanh(x))/2$. The probability of the i^{th} thrust combination is then given by

$$\mu_i(v, \theta) = \frac{\xi_i}{\sum_j \xi_j},$$

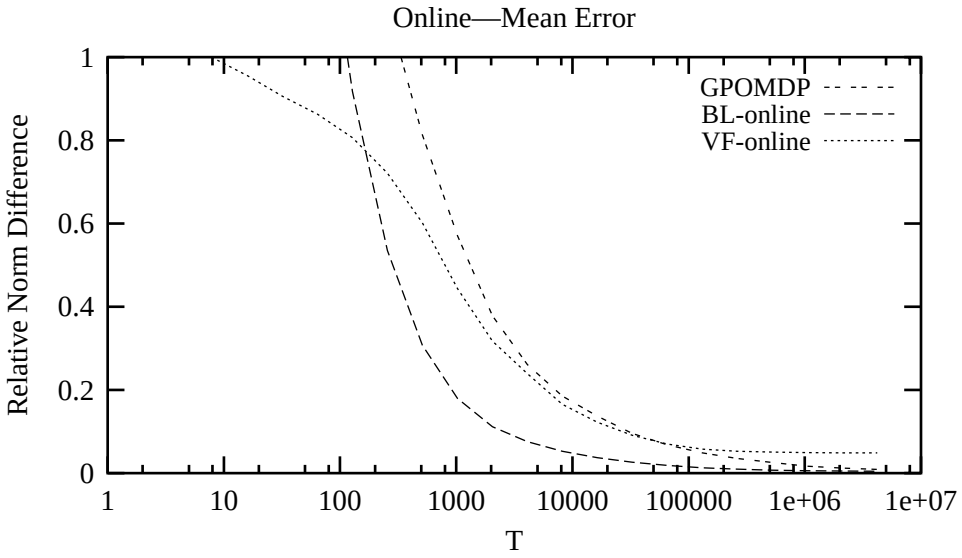


Figure 4: The mean of the relative norm difference from $\nabla\eta$: using no control variate (GPOMDP); using a baseline trained online (BL-online); and using a value function trained online (VF-online).

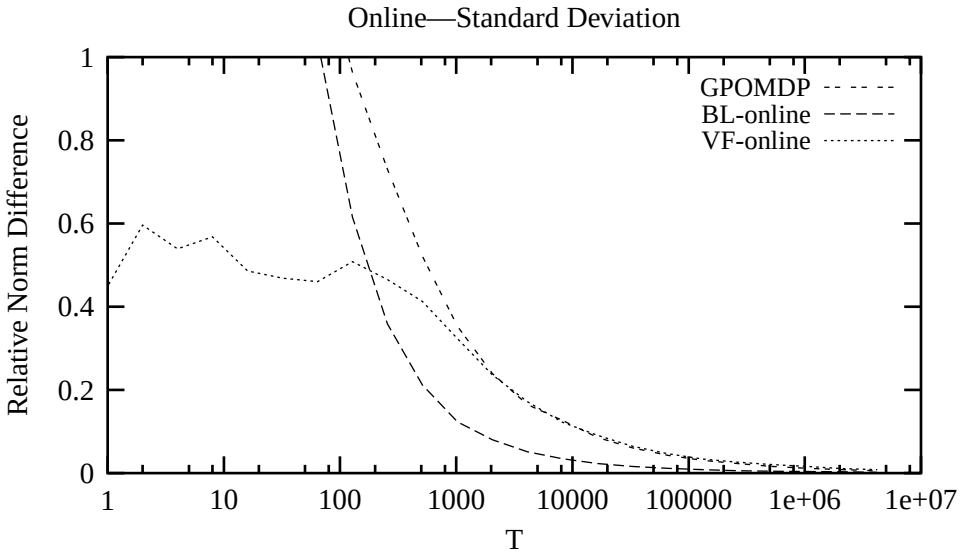


Figure 5: The standard deviation of the relative norm difference from $\nabla\eta$ (see Figure 4 for an explanation of the key).

where actions have been labelled with the associated thrust combination.

The puck was first trained using conjugate gradient ascent, with GPOMDP (with $T = 10^8$) used to estimate the gradient. The parameters of the policy were recorded at 28 points along the training curve: at the initial parameter setting ($\theta = 0$); and at each change in line search direction. Results for 4 of the 28 points are shown in Figure 6. The results show the mean, over 100 independent trials, of the relative norm difference between gradient estimates when using a range of different baselines, all learned online ($\gamma_t = 1/\ln(1+t)$) and initially set to zero, and an estimate of $\nabla_\beta \eta$. The second order baseline was a second order polynomial of the inputs, that is, again collating the inputs in the vector v,

$$b(v,\omega) = \omega_{0,0} + \sum_{k=1}^{7} \omega_{k,0} v_k + \sum_{k=1}^{7} \sum_{l=k}^{7} \omega_{k,l} v_k v_l,$$

where ω is a vector of 32 elements, with one element, $\omega_{k,l}$, for each second order term $v_k v_l$, one additional element, $\omega_{k,0}$, for each first order term v_k, and one additional element, $\omega_{0,0}$, for the constant term. The estimate of $\nabla_\beta \eta$ was produced by averaging the unbiased $\nabla_\beta \eta$ estimates at $T = 2^{23}$; an average over 400 samples.

Figure 6 shows that each baseline method performed better than GPOMDP, with the second order baseline performing the best of these. The estimated average reward and the estimated optimal constant baseline performed almost equally, and both performed better than the online constant baseline in this case. That the two estimation methods performed almost equally would suggest that, in this case, the random variables $(\nabla \mu_u(y)/\mu_u(y))^2$ and $J_\beta(j)$ are close to independent. It might be that for most policies, or at least policies at the θ values we tested, $\|\mathbb{E}_\pi J_\beta(i)\| \gg \|J_\beta(i) - \mathbb{E}_\pi J_\beta(i)\|$, since this implies

$$\begin{aligned}
\mathbb{E}_\pi \left[\left(\frac{\nabla \mu_u(y)}{\mu_u(y)} \right)^2 J_\beta(j) \right] &= \mathbb{E}_\pi \left(\frac{\nabla \mu_u(y)}{\mu_u(y)} \right)^2 \mathbb{E}_\pi J_\beta(i) + \mathbb{E}_\pi \left[\left(\frac{\nabla \mu_u(y)}{\mu_u(y)} \right)^2 \left(J_\beta(j) - \mathbb{E}_\pi J_\beta(i) \right) \right] \\
&\approx \mathbb{E}_\pi \left(\frac{\nabla \mu_u(y)}{\mu_u(y)} \right)^2 \mathbb{E}_\pi J_\beta(i).
\end{aligned}$$

9. Conclusions

We have shown that the use of control variate techniques can reduce estimation variance when estimating performance gradients. The first technique was to add a baseline. Here we analyzed the variance quantities of (8) and (9), the estimation variance when using a baseline under the assumption that the discounted value function is known and samples may be drawn independently. We have given the optimal baseline, the baseline that minimizes this variance, and have expressed the additional variance resulting from using an arbitrary baseline as a weighted squared distance from this optimum. Similar results have also been shown for a constant baseline. Here it was also shown how much additional variance results from using the expected discounted value function, a popular choice of baseline, in place of the optimal constant baseline. We have also shown that the estimation variance from $\Delta_T^{(+S)}(b_\mathcal{Y})$, a realizable estimate of $\nabla_\beta \eta$ formed from a single sample path of the associated POMDP, is bounded by the stationary variance plus a term independent of the choice of baseline, and another term of negligible magnitude.

A second control variate technique used to reduce estimation variance was to replace estimates of the discounted value function with some appropriate value function V. We have shown that, even

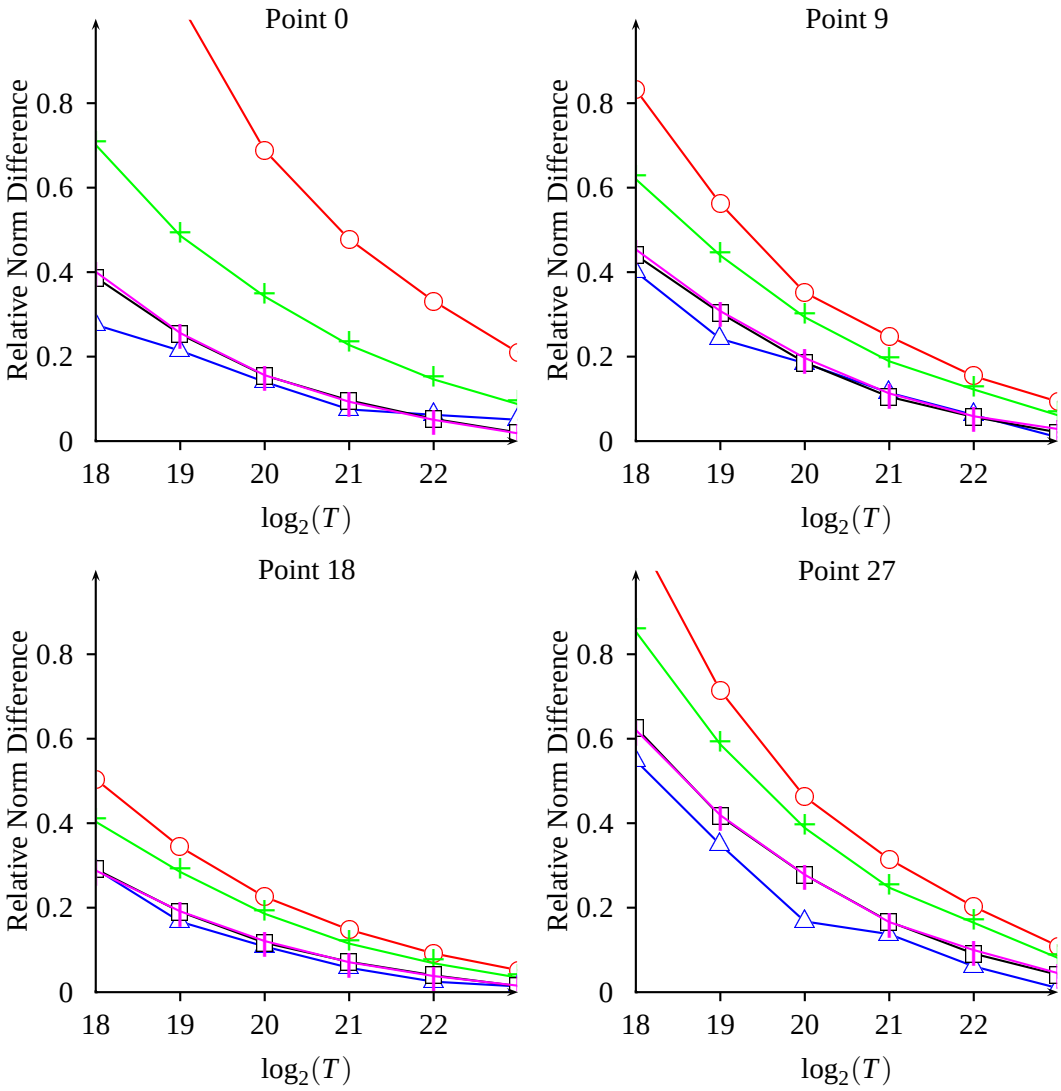


Figure 6: Each plot shows the mean, over 100 independent runs, of the relative norm difference from $\nabla_\beta \eta$: using no baseline ( ); using a constant baseline, trained online ( ); using a second order polynomial of the inputs as a baseline, trained online ( ); using $\mathbb{E}_\pi J_\beta(i)$ as a baseline, estimated online ( ); and using the optimal constant baseline, estimated online ( ). The reference $\nabla_\beta \eta$ is estimated by averaging the unbiased estimates at $T = 2^{23}$. The four plots show four of the 28 parameter values at the end points of each line search in the conjugate gradient ascent algorithm, when training on the target location example using GPOMDP (with $T = 10^8$) to produce gradient estimates. The remaining 24 parameter values give similar plots.

if the discounted value function is known, selecting V to be equal to the discounted value function is not necessarily the best choice. We have shown examples where this is the case; where additional reduction in estimation variance can be achieved by selecting V to be a function other than the discounted value function, with no addition of estimation bias. We have also given a bound on the expected squared error of the estimate Δ_T^{V}, an estimate of $\nabla_\beta \eta$ that uses V in place of discounted value function estimates and is formed from a single sample path of the associated MDP.

The gradient estimates $\Delta_T^{(+S)}(\mathrm{b}_\mathcal{Y})$ and Δ_T^{V} use a baseline $\mathrm{b}_\mathcal{Y}$ and a value function V, respectively, in their calculations. In experiments on a toy problem we investigated the improvements obtained when using the optimal choice of baseline in $\Delta_T^{(+S)}(\mathrm{b}_\mathcal{Y})$, and also when using the value function minimizing the bound on expected squared error of estimates in Δ_T^{V}. Significant improvement was shown.

In general the optimal choices for the baseline and the value function may not be known. We have explored the idea of using gradient descent on the error bounds derived in this paper to learn a good choice for a baseline, or for a value function. We have given realizable algorithms to obtain the appropriate gradient estimates, along with their online versions. In experiments on the toy problem, and in a target location problem, we have seen some improvements given by these algorithms.

In experiments we have looked at using the online versions of the algorithms in Section 7; updating the baseline (or value function) whilst estimating the gradient of the performance. Consequently some additional bias in the performance gradient estimate is likely to have occurred. The results of the experiments, however, would suggest that this bias is small. Further work of interest is the study of the convergence of these online algorithms, and also the convergence of the performance whilst using these online algorithms.

Acknowledgments

Most of this work was performed while the authors were with the Research School of Information Sciences and Engineering at the Australian National University, supported in part by the Australian Research Council.

Appendix A. Discussion of Assumption 1

The details in this section can be found in texts covering Markov chains; see, for example, Puterman (1994); Grimmett and Stirzaker (1992); Seneta (1981). We include them to: define the terms used in Assumption 1; show how Assumption 1 may be relaxed; and give an intuition of our use of Assumption 1.

The states of a Markov chain $M = (\mathcal{S}, P)$ can be divided into equivalence classes under the communicating relation $\leftrightarrow$. We define $i \leftrightarrow i$, and write $i \leftrightarrow j$ if there are integers $m, n > 0$ such that $p_{ij}^{(m)} > 0$ and $p_{ji}^{(n)} > 0$, where $p_{ij}^{(t)}$ is the ij^{th} entry of the t-step transition matrix P^t. We call a class $\tilde{\mathcal{S}} \subset \mathcal{S}$ recurrent if its states are recurrent, otherwise we call it transient. A state is recurrent if $\Pr\{X_t = i \text{ for some } t > 0 | X_0 = i\} = 1$, otherwise it is transient. Notice that this means that once the chain enters a recurrent class it never leaves, but rather visits all states of that class infinitely often. If the chain is finite then it will eventually leave every transient class and settle in some recurrent class.

We say a Markov chain $M = (\mathcal{S}, P)$ is irreducible if the space $\mathcal{S}$ forms a single class under $\leftrightarrow$; necessarily a recurrent class for finite Markov chains. We can relax the irreducibility condition, and instead allow any $\mathcal{S}$ that contains a single recurrent class $\mathcal{S}_R$ plus a set (possibly containing more than one class) of transient states $\mathcal{S}_T$ such that $\Pr\{X_t \in \mathcal{S}_R \text{ for some } t > 0 | X_0 = j\} = 1$ for all $j \in \mathcal{S}_T$ (guaranteed for finite chains).

The period, d, of a state $i \in \mathcal{S}$ of a Markov chain $M = (\mathcal{S}, P)$ is the greatest common divisor of the set of times $\{t > 0 : p_{ii}^{(t)} > 0\}$. It is uniform across the states of a class. A state, and consequently a class, is aperiodic if $d = 1$. We can relax the aperiodicity condition and allow arbitrary periods. Consider $\mathcal{S}_R$ to be constructed $\mathcal{S}_R = \mathcal{S}_0 \cup \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_{d-1}$, where d is the period of $\mathcal{S}_R$ and the sets $\mathcal{S}_k$ are chosen such that $\Pr\{X_{t+1} \in \mathcal{S}_{k+1 (\text{mod } d)} | X_t \in \mathcal{S}_k\} = 1$.

Our interest is in the existence and uniqueness of the stationary distribution π. The existence of π stems from the Markov chain reaching, and never leaving, a recurrent class, combined with the forgetfulness of the Markov property. The uniqueness of π stems from us allowing only a single recurrent class. So given a finite Markov chain $M = (\mathcal{S}, P)$ with the construction $\mathcal{S} = \mathcal{S}_T \cup \mathcal{S}_R$, and $\mathcal{S}_R = \mathcal{S}_0 \cup \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_{d-1}$, as above, we have, writing $N_{[0,T)}(i)$ to denote the number of times state i is visited before time T,

$$\lim_{T \to \infty} T^{-1} N_{[0,T)}(i) = \pi_i, \qquad \text{a.s.} \tag{24}$$

Equation (24) is helpful in two ways. Firstly, our choice of performance measure, (2), is the expected average of $\mathrm{r}(X_t)$, where $\{X_t\}$ is produced by the chain. We see from (24) that this value is independent of the initial state, and we could equivalently use the expected value of $\mathrm{r}(X)$, with $X \sim \pi$. Secondly, we are interested in calculating expectations over the stationary distribution (such as $\nabla_\beta \eta$), and we see from (24) that this expectation can be calculated by observing a single sample path generated by the Markov chain, almost surely. In Section 4 it is seen that we can even do well with a finite length sample path; it is here we use the assumption of irreducibility and aperiodicity.

The analytical results of Section 5 and Section 6 use Theorem 3, Theorem 4 and Corollary 5 of Section 4 to bound the variance terms of the form

$$\mathrm{Var}\left(\frac{1}{T} \sum_{t=0}^{T-1} \mathrm{f}(X_t) \right), \tag{25}$$

where X_t is generated by a Markov chain $M = (\mathcal{S}, P)$ starting in the stationary distribution $X_0 \sim \pi$, with variance terms of the form

$$\mathrm{Var}(\mathrm{f}(X)), \tag{26}$$

where $X \sim \pi$. The proofs of these results use the property

$$\lim_{T \to \infty} \Pr\{X_T = i\} = \pi_i, \tag{27}$$

which holds when $\mathcal{S}_R$ is aperiodic, and is stronger than Equation (24). In particular, Equation (27) holds with the addition of the set of transient states $\mathcal{S}_T$; indeed the variance quantities of Equation (25) and (26) are not affected by such an addition, as the set $\mathcal{S}_T$ has π-measure zero. Also, when $\mathcal{S}_R$ is periodic, writing $X_t^{(k)}$ for the d-step subprocess with elements in $\mathcal{S}_k$ and $\pi^{(k)}$ for the

stationary distribution corresponding to this irreducible aperiodic chain, we have

$$\begin{aligned}
&\mathrm{Var}\left(\frac{1}{dT}\sum_{t=0}^{dT-1}\mathrm{f}(X_t)\right)\\
&= \mathbb{E}\left[\left(\frac{1}{dT}\sum_{t=0}^{dT-1}\mathrm{f}(X_t)-\mathbb{E}\left[\frac{1}{dT}\sum_{t=0}^{dT-1}\mathrm{f}(X_t)\middle| X_0\sim\pi\right]\right)^2\middle| X_0\sim\pi\right]\\
&= \frac{1}{d^2}\sum_{k_1=0}^{d-1}\sum_{k_2=0}^{d-1}\mathbb{E}\left[\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathrm{f}\left(X_t^{(k_1)}\right)-\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\mathrm{f}\left(X_t^{(k_1)}\right)\middle| X_0\sim\pi\right]\right)\right.\\
&\qquad\left.\times\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathrm{f}\left(X_t^{(k_2)}\right)-\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\mathrm{f}\left(X_t^{(k_2)}\right)\middle| X_0\sim\pi\right]\right)\middle| X_0\sim\pi\right]\\
&\leq \frac{2}{d}\sum_{k=0}^{d-1}\mathbb{E}\left[\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathrm{f}\left(X_t^{(k)}\right)-\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\mathrm{f}\left(X_t^{(k)}\right)\middle| X_0\sim\pi\right]\right)^2\middle| X_0\sim\pi\right]\\
&= \frac{2}{d}\sum_{k=0}^{d-1}\mathbb{E}\left[\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathrm{f}\left(X_t^{(k)}\right)-\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\mathrm{f}\left(X_t^{(k)}\right)\middle| X_0^{(k)}\sim\pi^{(k)}\right]\right)^2\middle| X_0^{(k)}\sim\pi^{(k)}\right]\\
&= \frac{2}{d}\sum_{k=0}^{d-1}\mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathrm{f}\left(X_t^{(k)}\right)\right)
\end{aligned}$$

(that the distribution of $X_0^{(k)}$ is $\pi^{(k)}$ when $X_0\sim\pi$ is due to the sets $\mathcal{S}_k$ having equal π-measure.) It is now straightforward to give analogous results to those of Section 5 and 6 when the Markov chain consists of a single, possibly periodic, recurrent class plus a set of transient states.

The justification in studying the variance quantity (25) is that, after leaving the set of states $\mathcal{S}_T$, the distribution over states will approach π exponentially quickly. Whilst this does not hold for periodic chains, it does hold that the distribution over states restricted to the set of times $\{T_k+dt : t\in\{0,1,2\ldots\}\}$, where T_k is the first time X_t hits the set S_k, will approach $\pi^{(k)}$ exponentially quickly.

Appendix B. Proofs for Section 4

In this section we give the proofs for Theorem 2, Theorem 3, Theorem 4, and Corollary 5 of Section 4. Before giving the proof of Theorem 2 we first look at some properties of Markov chains. In particular, we look at the covariance decay matrix of a finite ergodic Markov chain.

Definition 4. *Let $M=(\mathcal{S},P)$ be a finite ergodic Markov chain, and let π be its stationary distribution. We denote the* covariance decay matrix *of this chain by $D(t)$, and define it by*

$$D(t)\stackrel{\text{def}}{=}\Pi^{\frac{1}{2}}\left(P^t-e\pi'\right)\Pi^{-\frac{1}{2}}$$

where, given $\mathcal{S}=\{1,2,\ldots,n\}$, $\Pi^{\frac{1}{2}}=\mathrm{diag}(\sqrt{\pi_1},\sqrt{\pi_2},\ldots,\sqrt{\pi_n})$, and $\Pi^{-\frac{1}{2}}=\left[\Pi^{1/2}\right]^{-1}$.

We will see that the gain of the auto-covariance over the variance can be bound by the spectral norm of the covariance decay matrix. First we will give a bound on the spectral norm of the covariance decay matrix for general finite ergodic Markov chains. Then we will give a much tighter

bound for reversible finite ergodic Markov chains. The spectral norm of a matrix is given by the following definition.

Definition 5. *The* spectral norm *of a matrix A is denoted* $\|A\|_\lambda$. *It is the matrix norm induced by the Euclidean norm,*

$$\|A\|_\lambda \stackrel{\text{def}}{=} \max_{\|x\|=1} \|Ax\|.$$

An equivalent definition is

$$\|A\|_\lambda = \max_{\|x\|=1} \|Ax\| = \max_{\|x\|=1} \sqrt{x'A'Ax} = \sqrt{\lambda_{\max}(A'A)},$$

where $\lambda_{\max}(A)$ *denotes the largest eigenvalue of the matrix A. As* $A'A$ *is symmetric and positive semi-definite, all of its eigenvalues are real and positive.*

Note 2. *We have that, for any matrix A*

$$\|Ax\| \leq \|A\|_\lambda \|x\|.$$

This can be seen from: for $\|x\| \neq 0$

$$\|Ax\| = \left\| A\left(\frac{x}{\|x\|}\right) \right\| \|x\|.$$

Recall the following two definitions.

Definition 6. *The* total variation distance *between two distributions* p, q *on the finite set* $\mathcal{S}$ *is given by*

$$\mathrm{d}_{TV}(p,q) \stackrel{\text{def}}{=} \frac{1}{2}\sum_{i\in\mathcal{S}} |p_i - q_i| = \sum_{i\in\mathcal{S}: p_i > q_i} (p_i - q_i).$$

Definition 7. *The* mixing time *of a finite ergodic Markov chain* $M = (\mathcal{S}, P)$ *is defined as*

$$\tau \stackrel{\text{def}}{=} \min\left\{ t > 0 : \max_{i,j} \mathrm{d}_{TV}\left(P_i^t, P_j^t\right) \leq e^{-1} \right\},$$

where P_i^t *denotes the* i^{th} *row of the t-step transition matrix* P^t.

Note 3. *Denoting* $d_t \stackrel{\text{def}}{=} \max_{i,j} \mathrm{d}_{TV}\left(P_i^t, P_j^t\right)$, *for* $s,t \geq 1$ *we have that* $d_{t+s} \leq d_t d_s$, *and hence*

$$d_t \leq \exp\left(-\lfloor t/\tau \rfloor\right).$$

Note 4. *We have that*

$$\max_{i\in\mathcal{S}} \mathrm{d}_{TV}\left(P_i^t, \pi\right) \leq d_t.$$

The sub-multiplicative property in Note 3 can be seen from

$$\begin{aligned}
\mathrm{d}_{TV}\left(P_i^{t+s},P_j^{t+s}\right) &= \sum_{l\in\mathcal{S}:p_{il}^{(t+s)}>p_{jl}^{(t+s)}}\left(p_{il}^{(t+s)}-p_{jl}^{(t+s)}\right)\\
&= \sum_{l\in\mathcal{S}:p_{il}^{(t+s)}>p_{jl}^{(t+s)}}\sum_{k\in\mathcal{S}}\left(p_{ik}^{(t)}-p_{jk}^{(t)}\right)p_{kl}^{(s)}\\
&= \sum_{k\in\mathcal{S}:p_{ik}^{(t)}>p_{jk}^{(t)}}\left(p_{ik}^{(t)}-p_{jk}^{(t)}\right)\sum_{l\in\mathcal{S}:p_{il}^{(t+s)}>p_{jl}^{(t+s)}}p_{kl}^{(s)}\\
&\quad -\sum_{k\in\mathcal{S}:p_{jk}^{(t)}>p_{ik}^{(t)}}\left(p_{jk}^{(t)}-p_{ik}^{(t)}\right)\sum_{l\in\mathcal{S}:p_{il}^{(t+s)}>p_{jl}^{(t+s)}}p_{kl}^{(s)}\\
&\leq d_{TV}\left(P_i^t,P_j^t\right)\max_{k_1,k_2\in\mathcal{S}}\sum_{l\in\mathcal{S}:p_{il}^{(t+s)}>p_{jl}^{(t+s)}}\left(p_{k_1l}^{(s)}-p_{k_2l}^{(s)}\right)\\
&\leq d_{TV}\left(P_i^t,P_j^t\right)d_s, \qquad (28)
\end{aligned}$$

where $p_{ij}^{(t)}$ denotes the ij^{th} component of P^t, and we have used that $\sum_l p_{il}^{(t)}p_{lk}^{(s)}=p_{ik}^{(t+s)}$. As $d_t\leq 1$, this also implies d_t is non-increasing (for $t\geq 1$). The inequality in Note 3 then follows from applying the sub-multiplicative property to $\tau\lfloor t/\tau\rfloor\leq t$, giving

$$d_t\leq\begin{cases} d_\tau^{\lfloor t/\tau\rfloor} & t\geq\tau,\\ 1 & t<\tau.\end{cases}$$

Note 4 can be seen from

$$\sum_{k\in\mathcal{S}}\left|p_{ik}^{(t)}-\pi_k\right|=\sum_{k\in\mathcal{S}}\left|\sum_{j\in\mathcal{S}}\pi_j\left(p_{ik}^{(t)}-p_{jk}^{(t)}\right)\right|\leq\sum_{j\in\mathcal{S}}\pi_j\sum_{k\in\mathcal{S}}\left|p_{ik}^{(t)}-p_{jk}^{(t)}\right|.$$

We will also consider the following, asymmetric, notion of distance.

Definition 8. *The* χ^2 distance *between the distribution* p *and the distribution* q *on the finite set* $\mathcal{S}$, *with* $q_i>0$ *for all* $i\in\mathcal{S}$, *is given by*

$$\mathrm{d}_{\chi^2}(p,q)\stackrel{\text{def}}{=}\left(\sum_{i\in\mathcal{S}}\frac{(p_i-q_i)^2}{q_i}\right)^{1/2}.$$

Lemma 21. *Let* $M=(\mathcal{S},P)$ *be a finite ergodic Markov chain, and let* π *be its stationary distribution. There exists a mixing time* τ, *which is a property of* M, *such that*

$$\|D(t)\|_\lambda\leq\sqrt{\mathbb{E}_{i\sim\pi}\left(\mathrm{d}_{\chi^2}(P_i^t,\pi)\right)^2}\leq\sqrt{2\left|\mathcal{S}\right|\max_{i\in\mathcal{S}}\mathrm{d}_{TV}(P_i^t,\pi)}\leq\sqrt{2\left|\mathcal{S}\right|\exp\left(-\lfloor t/\tau\rfloor\right)}.$$

Thus we have $\|D(t)\|_\lambda\leq\sqrt{2\left|\mathcal{S}\right|\exp\left(-\lfloor t/\tau\rfloor\right)}$.

Proof. Note that $D(t)'D(t)$ is symmetric and positive semi-definite and hence its eigenvalues are real and positive. Label them, in non-increasing order, $\lambda_1, \lambda_2, \ldots$. This combined with the relationship $\sum_i \lambda_i = \mathrm{tr}\left(D(t)'D(t)\right)$, where $\mathrm{tr}(A)$ denotes the trace of the matrix A, gives

$$0 \leq \lambda_1 \leq \mathrm{tr}\left(D(t)'D(t)\right).$$

Furthermore,

$$\begin{aligned}
\mathrm{tr}\left(D(t)'D(t)\right) &= \sum_{i\in\mathcal{S}}\sum_{k\in\mathcal{S}} \frac{\pi_i}{\pi_k}\left(p_{ik}^{(t)} - \pi_k\right)^2 = \mathbb{E}_{i\sim\pi}\left(\mathrm{d}_{\chi^2}\left(P_i^t, \pi\right)\right)^2 \\
&= \sum_{i\in\mathcal{S}}\sum_{k\in\mathcal{S}} \frac{\pi_i p_{ik}^{(t)}}{\pi_k}\left(p_{ik}^{(t)} - \pi_k\right) - \sum_{i\in\mathcal{S}}\sum_{k\in\mathcal{S}} \pi_i\left(p_{ik}^{(t)} - \pi_k\right) \\
&\leq \sum_{k\in\mathcal{S}} \max_{\tilde{i}\in\mathcal{S}} \left|p_{\tilde{i}k}^{(t)} - \pi_k\right| \left(\frac{1}{\pi_k}\sum_{i\in\mathcal{S}} \pi_i p_{ik}^{(t)}\right) \\
&\leq |\mathcal{S}| \max_{i\in\mathcal{S}} \sum_{k\in\mathcal{S}} \left|p_{ik}^{(t)} - \pi_k\right| \\
&= 2\,|\mathcal{S}| \max_{i\in\mathcal{S}} \mathrm{d}_{TV}\left(P_i^t, \pi\right) \\
&\leq 2\,|\mathcal{S}| \exp\left(-\lfloor t/\tau \rfloor\right).
\end{aligned}$$

The last inequality follows from Note 3 and Note 4. ■

Recall that a reversible Markov chain has a transition probability matrix and stationary distribution satisfying the detailed balance equations

$$\pi_i p_{ij} = \pi_j p_{ji},$$

for all i, j.

Lemma 22. *Let $M = (\mathcal{S}, P)$ be a finite ergodic reversible Markov chain, and let π be its stationary distribution. Order the eigenvalues of P such that $1 = \lambda_1 > |\lambda_2| \geq |\lambda_3| \geq \ldots$. Then*

$$\|D(t)\|_\lambda = |\lambda_2|^t.$$

Furthermore, if M has mixing time τ, we have that $|\lambda_2|^t \leq 2\exp\left(-\lfloor t/\tau \rfloor\right)$.

Proof. As P is reversible,

$$\sqrt{\frac{\pi_i}{\pi_j}} p_{ij}^{(t)} = \sqrt{\frac{\pi_i}{\pi_j}}\left(\frac{\pi_j}{\pi_i}\right) p_{ji}^{(t)} = \sqrt{\frac{\pi_j}{\pi_i}} p_{ji}^{(t)}, \tag{29}$$

and hence $D(t)' = D(t)$. Given a polynomial $\mathrm{f}(\cdot)$ and the symmetric matrix A, $Ax = \lambda x$ implies $\mathrm{f}(A)x = \mathrm{f}(\lambda)x$. Thus,

$$\|D(t)\|_\lambda = \sqrt{\lambda_{\max}\left(D(t)'D(t)\right)} = \sqrt{\lambda_{\max}\left(D(t)^2\right)} = \max_i \left|\lambda_i(D(t))\right|,$$

where $\lambda_1(D(t)), \lambda_2(D(t)), \ldots$ are the eigenvalues of $D(t)$. The matrix $D(t)$ is similar to $(P^t - e\pi')$ via the matrix $\Pi^{\frac{1}{2}}$, and hence has the same eigenvalues. Let $x_1, x_2, x_3, \ldots$ be the left eigenvectors of P, labelled with the indices of their associated eigenvalues. Then

$$x_i' \left(P^t - e\pi'\right) = x_i' \left(P^t - \lim_{n\to\infty} P^n\right) = \lambda_i^t x_i' - \lim_{n\to\infty} \lambda_i^n x_i' = \begin{cases} 0 & i = 1, \\ \lambda_i^t x_i' & i \neq 1. \end{cases}$$

Therefore λ_2^t is the greatest magnitude eigenvalue of $D(t)$. Furthermore, if $x \neq 0$ is a right eigenvector of $D(t)$ with eigenvalue λ, we have

$$\sum_{j\in\mathcal{S}} \sqrt{\frac{\pi_i}{\pi_j}} \left(p_{ij}^{(t)} - \pi_j\right) x_j = \frac{1}{\sqrt{\pi_i}} \sum_{j\in\mathcal{S}} \left(p_{ji}^{(t)} - \pi_i\right) \sqrt{\pi_j} x_j = \lambda x_i, \qquad \text{from (29)},$$

and so

$$\begin{aligned} |\lambda| \sum_{i\in\mathcal{S}} \sqrt{\pi_i}|x_i| &= \sum_{i\in\mathcal{S}} \left| \sum_{j\in\mathcal{S}} \left(p_{ji}^{(t)} - \pi_i\right) \sqrt{\pi_j} x_j \right| \\ &\leq \sum_{i\in\mathcal{S}} \sum_{j\in\mathcal{S}} \left|p_{ji}^{(t)} - \pi_i\right| \sqrt{\pi_j}|x_j| \\ &\leq \left(\max_{i\in\mathcal{S}} \sum_{k\in\mathcal{S}} \left|p_{ik}^{(t)} - \pi_k\right| \right) \sum_{j\in\mathcal{S}} \sqrt{\pi_j}|x_j| \\ &= 2 \max_{i\in\mathcal{S}} \mathrm{d}_{TV}\left(P_i^t, \pi\right) \sum_{j\in\mathcal{S}} \sqrt{\pi_j}|x_j|. \end{aligned}$$

So from Note 3 and Note 4 we have that $|\lambda| \leq 2\exp(-\lfloor t/\tau \rfloor)$. ■

Lemma 23. *Let $M = (S, P)$ be a finite ergodic Markov chain, and let π be its stationary distribution. Let $\{X_t\}$ be the process generated by M starting $X_0 \sim \pi$. For any two functions $\mathrm{f}, \mathrm{g} : \mathcal{S} \to \mathbb{R}$*

$$\left|\mathbb{E}\left[\left(\mathrm{f}(X_s) - \mathbb{E}_\pi \mathrm{f}(i)\right)\left(\mathrm{g}(X_{s+t}) - \mathbb{E}_\pi \mathrm{g}(i)\right)\right]\right| \leq \|D(t)\|_\lambda \sqrt{\mathbb{E}_\pi\left(\mathrm{f}(i) - \mathbb{E}_\pi \mathrm{f}(i)\right)^2 \mathbb{E}_\pi\left(\mathrm{g}(i) - \mathbb{E}_\pi \mathrm{g}(i)\right)^2}.$$

Proof. Denoting $\underline{\mathrm{f}}$ to be the column vector of $\mathrm{f}(x) - \mathbb{E}_\pi \mathrm{f}(i)$ over the states $x \in \mathcal{S}$, then

$$\begin{aligned} \left|\mathbb{E}\left[\underline{\mathrm{f}}_{X_s} \underline{\mathrm{g}}_{X_{s+t}}\right]\right| &= \left|\underline{\mathrm{f}}' \Pi P^t \underline{\mathrm{g}}\right| \\ &= \left|\underline{\mathrm{f}}' \Pi \left(P^t - e\pi'\right) \underline{\mathrm{g}}\right| \\ &= \left|\underline{\mathrm{f}}' \Pi^{\frac{1}{2}} D(t) \Pi^{\frac{1}{2}} \underline{\mathrm{g}}\right| \\ &\leq \left\|\underline{\mathrm{f}}' \Pi^{\frac{1}{2}}\right\|_2 \left\|D(t) \Pi^{\frac{1}{2}} \underline{\mathrm{g}}\right\|_2 \qquad \text{(Schwartz)} \\ &\leq \|D(t)\|_\lambda \left\|\underline{\mathrm{f}}' \Pi^{\frac{1}{2}}\right\|_2 \left\|\Pi^{\frac{1}{2}} \underline{\mathrm{g}}\right\|_2 \qquad \text{(Note 2)} \\ &= \|D(t)\|_\lambda \sqrt{\mathbb{E}_\pi\left(\underline{\mathrm{f}}_i\right)^2 \mathbb{E}_\pi\left(\underline{\mathrm{g}}_i\right)^2}. \end{aligned}$$

■

Lemma 23 shows how covariance terms can be bounded by the variance under the stationary distribution attenuated by the spectral norm of the covariance decay matrix. Combining this with Lemma 23 (or Lemma 22 for reversible chains) gives us Theorem 2.

Proof of Theorem 2. By the application of Lemma 21 and Lemma 23,

$$\begin{aligned} |\mathrm{Cov}_\pi(t;\mathrm{f})| &\leq \|D(t)\|_\lambda \mathrm{Var}_\pi(\mathrm{f}) \qquad \text{(Lemma 23)} \\ &\leq \sqrt{2|\mathcal{S}|\exp(-\lfloor t/\tau \rfloor)}\mathrm{Var}_\pi(\mathrm{f}) \qquad \text{(Lemma 21)} \\ &\leq \sqrt{2|\mathcal{S}|e}\sqrt{\exp(-t/\tau)}\mathrm{Var}_\pi(\mathrm{f}). \end{aligned}$$

This shows that Theorem 2 holds with some $\mathbf{L} \leq \sqrt{2|\mathcal{S}|e}$ and $0 \leq \alpha \leq \exp(-1/(2\tau))$. If the chain is reversible, then similarly, using Lemma 22, the bound of Theorem 2 holds with $\mathbf{L} = 2e$ and $\alpha = \exp(-1/\tau)$. ■

We can use the result of Theorem 2 to prove Theorem 3. Recall that Theorem 3 shows how the variance of an average of dependent samples can be bounded by $O(1/T)$ times the variance of a sample distributed according to the stationary distribution.

Proof of Theorem 3.

$$\begin{aligned} \mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathrm{f}(X_t)\right) &= \frac{1}{T^2}\mathbb{E}\left(\sum_{t=0}^{T-1}(\mathrm{f}(X_t)-\mathbb{E}\mathrm{f}(X_t))\right)^2 \\ &= \frac{1}{T^2}\sum_{t_1=0}^{T-1}\sum_{t_2=0}^{T-1}\mathbb{E}\left[(\mathrm{f}(X_{t_1})-\mathbb{E}\mathrm{f}(X_{t_1}))(\mathrm{f}(X_{t_2})-\mathbb{E}\mathrm{f}(X_{t_2}))\right] \\ &= \frac{1}{T^2}\sum_{t_1=0}^{T-1}\sum_{t_2=0}^{T-1}\mathrm{Cov}_\pi(|t_2-t_1|;\mathrm{f}) \\ &= \frac{1}{T^2}\sum_{t=-(T-1)}^{T-1}(T-|t|)\,\mathrm{Cov}_\pi(|t|;\mathrm{f}). \end{aligned}$$

Then, using Theorem 2,

$$\begin{aligned} \frac{1}{T^2}\sum_{t=-(T-1)}^{T-1}(T-|t|)\,\mathrm{Cov}_\pi(|t|;\mathrm{f}) &\leq \frac{1}{T^2}\sum_{t=-(T-1)}^{T-1}(T-|t|)\,\mathbf{L}\alpha^{|t|}\mathrm{Var}(\mathrm{f}(X)) \\ &= \frac{\mathbf{L}}{T^2}\left(\frac{T(1+\alpha)}{1-\alpha}-\frac{2\alpha(1-\alpha^T)}{(1-\alpha)^2}\right)\mathrm{Var}(\mathrm{f}(X)) \\ &\leq \frac{1}{T}\left(\frac{\mathbf{L}(1+\alpha)}{1-\alpha}\right)\mathrm{Var}(\mathrm{f}(X)), \end{aligned}$$

where the equality follows from

$$\sum_{t=-(T-1)}^{T-1}(T-|t|)\,\alpha^{|t|} = T + 2T\sum_{t=1}^{T-1}\alpha^t - 2\sum_{t=1}^{T-1}t\alpha^t$$

$$
\begin{aligned}
&= T + \frac{2T\alpha\left(1-\alpha^{T-1}\right)}{1-\alpha} - 2\sum_{s=1}^{T-1}\sum_{t=s}^{T-1}\alpha^t \\
&= \left(T + \frac{2T\alpha}{1-\alpha}\right) - \frac{2T\alpha^T}{1-\alpha} - 2\sum_{s=1}^{T-1}\sum_{t=s}^{T-1}\alpha^t \\
&= \frac{T\left(1+\alpha\right)}{1-\alpha} - \frac{2T\alpha^T}{1-\alpha} - 2\sum_{s=1}^{T-1}\alpha^s\frac{1-\alpha^{T-s}}{1-\alpha} \\
&= \frac{T\left(1+\alpha\right)}{1-\alpha} - \frac{2T\alpha^T}{1-\alpha} - \frac{2\alpha\left(1-\alpha^{T-1}\right)}{\left(1-\alpha\right)^2} + \frac{2\left(T-1\right)\alpha^T}{1-\alpha} \\
&= \frac{T\left(1+\alpha\right)}{1-\alpha} - \frac{2\alpha\left(1-\alpha^{T-1}\right) + 2\alpha^T\left(1-\alpha\right)}{\left(1-\alpha\right)^2} \\
&= \frac{T\left(1+\alpha\right)}{1-\alpha} - \frac{2\alpha\left(1-\alpha^{T}\right)}{\left(1-\alpha\right)^2}.
\end{aligned}
$$

We may set the Ω^* in Theorem 3 to $\Omega^* = \mathbf{L}\left(1+\alpha\right)/\left(1-\alpha\right)$. Furthermore, recalling that $\alpha \leq \exp(-1/(2\tau))$, we have

$$\Omega^* = \mathbf{L}\frac{1+\alpha}{1-\alpha} \leq 2\mathbf{L}\frac{1}{1-\exp\left(-1/(2\tau)\right)} \leq 6\mathbf{L}\tau,$$

where the last inequality uses $\left[1-\exp(-1/(2\tau))\right]^{-1} \leq \frac{8}{3}\tau$. Note that for $x = 1/(2\tau)$ we have $0 \leq x \leq 1/2$, and that for such an x

$$
\begin{aligned}
\exp(-x) &\leq 1 - x + \frac{x^2}{2} \\
\Leftrightarrow \quad 1-\exp(-x) &\geq x\left(1-\frac{x}{2}\right) \\
\Leftrightarrow \quad \frac{1}{1-\exp(-x)} &\leq \frac{1}{x}\cdot\frac{2}{2-x} \\
\Rightarrow \quad \frac{1}{1-\exp(-x)} &\leq \frac{4}{3x}.
\end{aligned}
\tag{30}
$$

■

Theorem 4 gives a result similar to Theorem 3, but without relying on Theorem 2, and hence without relying on the size of the state space. For the proof we find it useful to define the following.

Definition 9. *The* triangular discrimination *(Topsøe, 2000) between two distributions p,q on the finite set $\mathcal{S}$ is given by*

$$\mathrm{d}_{\triangle}(p,q) \stackrel{\text{def}}{=} \sum_{i\in\mathcal{S}}\frac{\left(p_i-q_i\right)^2}{p_i+q_i}.$$

Note 5. *We have that* $\mathrm{d}_{\triangle}(p,q) \leq 2\mathrm{d}_{TV}(p,q)$.

Note 5 can be seen from $\sum_{i\in\mathcal{S}}\frac{\left(p_i-q_i\right)^2}{p_i+q_i} = \sum_{i\in\mathcal{S}}\frac{|p_i-q_i|}{p_i+q_i}|p_i-q_i| \leq \sum_{i\in\mathcal{S}}|p_i-q_i|$.

Proof of Theorem 4. Write $g_i = \mathrm{f}(i) - \mathbb{E}\mathrm{f}(X)$, and write $V = \sum_{i\in\mathcal{S}} \pi_i g_i^2$, the variance of $\mathrm{f}(X)$. We have that $|g_i| \leq 2c$ for all $i \in \mathcal{S}$. Now, we have for any $s \geq 0$ and $t \geq 0$,

$$\begin{aligned}
&\mathbb{E}\left(\mathrm{f}(X_s) - \mathbb{E}\mathrm{f}(X_s)\right)\left(\mathrm{f}(X_{s+t}) - \mathbb{E}\mathrm{f}(X_{s+t})\right) \\
&= \sum_{i,j\in\mathcal{S}} \pi_i g_i \left(p_{ij}^{(t)} - \pi_j\right) g_j \\
&= \sum_{i,j\in\mathcal{S}} \sqrt{\pi_i} \frac{p_{ij}^{(t)} - \pi_j}{\sqrt{p_{ij}^{(t)} + \pi_j}} \sqrt{\pi_i} g_i \sqrt{p_{ij}^{(t)} + \pi_j}\, g_j \\
&\leq \left(\sum_{i\in\mathcal{S}} \pi_i \sum_{j\in\mathcal{S}} \frac{\left(p_{ij}^{(t)} - \pi_j\right)^2}{p_{ij}^{(t)} + \pi_j}\right)^{1/2} \left(\sum_{i\in\mathcal{S}} \pi_i g_i^2 \sum_{j\in\mathcal{S}} \left(p_{ij}^{(t)} + \pi_j\right) g_j^2\right)^{1/2} \qquad \text{(Schwartz)} \\
&= \left(\sum_{i\in\mathcal{S}} \pi_i \mathrm{d}_{\triangle}\left(P_i^t, \pi\right)\right)^{1/2} \left(2V^2 + \sum_{i\in\mathcal{S}} \pi_i g_i^2 \sum_{j\in\mathcal{S}} \left(p_{ij}^{(t)} - \pi_j\right) g_j^2\right)^{1/2} \\
&\leq (2d_t)^{1/2} \left(2V^2 + (2c)^2 \sum_{i\in\mathcal{S}} \pi_i g_i^2 \sum_{j\in\mathcal{S}} \left|p_{ij}^{(t)} - \pi_j\right|\right)^{1/2} \qquad \text{(Note 5)} \\
&\leq 2d_t^{1/2} \left(V^2 + 2c^2 V d_t\right)^{1/2}. \qquad (31)
\end{aligned}$$

Consider the case where $V = \mathrm{Var}\left(\mathrm{f}(X)\right) > \varepsilon$. If $d_t \leq \varepsilon$, from Equation (31), we have

$$\mathbb{E}\left(\mathrm{f}(X_s) - \mathbb{E}\mathrm{f}(X_s)\right)\left(\mathrm{f}(X_{s+t}) - \mathbb{E}\mathrm{f}(X_{s+t})\right) \leq 2\sqrt{2}(1+c) d_t^{1/2} V. \qquad (32)$$

This holds for all s,t such that $d_t \leq \varepsilon$, which is implied by

$$\begin{aligned}
& \exp\left(-t/\tau + 1\right) \leq \varepsilon \\
\Leftrightarrow \quad & -t/\tau \leq \ln\varepsilon - 1 \\
\Leftrightarrow \quad & t \geq \tau\left(1 + \ln\frac{1}{\varepsilon}\right) \\
\Leftarrow \quad & t \geq 2\tau \ln\frac{1}{\varepsilon},
\end{aligned}$$

as $\varepsilon \leq e^{-1}$. For all s,t we have

$$\mathbb{E}\left(\mathrm{f}(X_s) - \mathbb{E}\mathrm{f}(X_s)\right)\left(\mathrm{f}(X_{s+t}) - \mathbb{E}\mathrm{f}(X_{s+t})\right) \leq V, \qquad (33)$$

which is a Cauchy-Schwartz inequality:

$$\begin{aligned}
&\mathbb{E}\left(\mathrm{f}(X_s) - \mathbb{E}\mathrm{f}(X_s)\right)\left(\mathrm{f}(X_{s+t}) - \mathbb{E}\mathrm{f}(X_{s+t})\right) \\
&= \sum_{i,j\in\mathcal{S}} \pi_i g_i p_{ij}^{(t)} g_j \\
&= \sum_{i,j\in\mathcal{S}} \sqrt{\pi_i p_{ij}^{(t)}}\, g_i \sqrt{\pi_i p_{ij}^{(t)}}\, g_j \\
&\leq \left(\sum_{i\in\mathcal{S}} \pi_i g_i^2 \sum_{j\in\mathcal{S}} p_{ij}^{(t)}\right)^{1/2} \left(\sum_{j\in\mathcal{S}} \left(\sum_{i\in\mathcal{S}} \pi_i p_{ij}^{(t)}\right) g_j^2\right)^{1/2}
\end{aligned}$$

$$
\begin{aligned}
&= \left(\sum_{i\in\mathcal{S}} \pi_i g_i^2\right)^{1/2} \left(\sum_{j\in\mathcal{S}} \pi_j g_j^2\right)^{1/2} \\
&= V.
\end{aligned}
$$

So from Equation (32) and Equation (33) we have

$$
\begin{aligned}
&\mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1} \mathrm{f}(X_t)\right) \\
&= \frac{1}{T^2}\sum_{t_1=0}^{T-1}\sum_{t_2=0}^{T-1} \mathbb{E}\left(\mathrm{f}(X_{t_1}) - \mathbb{E}\mathrm{f}(X_{t_1})\right)\left(\mathrm{f}(X_{t_2}) - \mathbb{E}\mathrm{f}(X_{t_2})\right) \\
&= \frac{1}{T^2}\sum_{t=0}^{T-1} \mathbb{E}\left(\mathrm{f}(X_t) - \mathbb{E}\mathrm{f}(X_t)\right)^2 + \frac{2}{T^2}\sum_{s=0}^{T-2}\sum_{t=1}^{T-s-1} \mathbb{E}\left(\mathrm{f}(X_s) - \mathbb{E}\mathrm{f}(X_s)\right)\left(\mathrm{f}(X_{s+t}) - \mathbb{E}\mathrm{f}(X_{s+t})\right) \\
&= \frac{1}{T^2}\sum_{t=0}^{T-1} \mathbb{E}\left(\mathrm{f}(X) - \mathbb{E}\mathrm{f}(X)\right)^2 + \frac{2}{T^2}\sum_{t=1}^{T-1} (T-t)\mathbb{E}\left(\mathrm{f}(X_0) - \mathbb{E}\mathrm{f}(X_0)\right)\left(\mathrm{f}(X_t) - \mathbb{E}\mathrm{f}(X_t)\right) \\
&= \frac{1}{T}V + \frac{2}{T^2}\sum_{t=1}^{\lfloor 2\tau\ln(1/\varepsilon)\rfloor} (T-t)\mathbb{E}\left(\mathrm{f}(X_0) - \mathbb{E}\mathrm{f}(X_0)\right)\left(\mathrm{f}(X_t) - \mathbb{E}\mathrm{f}(X_t)\right) \\
&\qquad + \frac{2}{T^2}\sum_{t=\lfloor 2\tau\ln(1/\varepsilon)\rfloor+1}^{T-1} (T-t)\mathbb{E}\left(\mathrm{f}(X_0) - \mathbb{E}\mathrm{f}(X_0)\right)\left(\mathrm{f}(X_t) - \mathbb{E}\mathrm{f}(X_t)\right) \\
&\leq \frac{1}{T}V + 4\tau\ln\left(\varepsilon^{-1}\right)\frac{1}{T}V + 4\sqrt{2}(1+c)\sum_{t=\lfloor 2\tau\ln(1/\varepsilon)\rfloor+1}^{\infty} d_t^{1/2}\,\frac{1}{T}V \\
&\leq \left(1 + 4\tau\ln\frac{1}{\varepsilon} + 25\tau(1+c)\varepsilon\right)\frac{1}{T}V, \qquad (34)
\end{aligned}
$$

where the last line follows from

$$
\begin{aligned}
\sum_{t=\lfloor 2\tau\ln(1/\varepsilon)\rfloor+1}^{\infty} d_t^{1/2} &\leq \sum_{t=\lfloor 2\tau\ln(1/\varepsilon)\rfloor+1}^{\infty} \exp\left(-t/(2\tau) + 1/2\right) \\
&= \sqrt{e}\sum_{t=\lfloor 2\tau\ln(1/\varepsilon)\rfloor+1}^{\infty} \exp\left(-t/(2\tau)\right) \\
&\leq \sqrt{e}\exp\left(-\ln\frac{1}{\varepsilon}\right)\sum_{t=0}^{\infty} \left(\exp\left(-1/(2\tau)\right)\right)^t \\
&= \sqrt{e}\varepsilon\frac{1}{1-\exp(-1/(2\tau))} \\
&\leq \frac{8\sqrt{e}}{3}\tau\varepsilon,
\end{aligned}
$$

where we have again used Equation (30). For the case where $\mathrm{Var}\left(\mathrm{f}(X)\right) \leq \varepsilon$ we have

$$
\mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1} \mathrm{f}(X_t)\right) = \frac{1}{T^2}\sum_{t_1=0}^{T-1}\sum_{t_2=0}^{T-1} \mathbb{E}\left(\mathrm{f}(X_{t_1}) - \mathbb{E}\mathrm{f}(X_{t_1})\right)\left(\mathrm{f}(X_{t_2}) - \mathbb{E}\mathrm{f}(X_{t_2})\right)
$$

$$\begin{aligned} &\leq \frac{1}{T^2}\sum_{t_1=0}^{T-1}\sum_{t_2=0}^{T-1} V \\ &\leq \varepsilon. \end{aligned} \tag{35}$$

As the variance is bounded either by Equation (34) or by Equation (35), taking their sum gives the result. ■

Lastly, we prove the corollary to Theorem 4, which shows the essential rate of decrease of the bound.

Proof of Corollary 5. Selecting ε such that, writing $V = \mathrm{Var}(\mathrm{f}(X))$,

$$\frac{1}{\varepsilon} = \frac{T}{4\tau V} + \frac{25}{4}(1+c),$$

satisfies $0 \leq \varepsilon \leq e^{-1}$. Substituting this into the result of Theorem 4 gives

$$\begin{aligned} \mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathrm{f}(X_t)\right) &\leq \frac{4\tau V}{T+25(1+c)\tau V} + \left(1 + \frac{100(1+c)\tau^2 V}{T+25(1+c)\tau V}\right. \\ &\qquad \left. + 4\tau\ln\left(\frac{T}{4\tau V} + \frac{25}{4}(1+c)\right)\right)\frac{V}{T} \\ &\leq \frac{4\tau V}{T} + \left(1 + 4\tau + 4\tau\ln\left(\frac{T}{4\tau V} + \frac{25}{4}(1+c)\right)\right)\frac{V}{T} \\ &\leq (1+8\tau)\frac{V}{T} + 4\tau\ln\left(7(1+c) + \frac{1}{4\tau}\left(\frac{V}{T}\right)^{-1}\right)\frac{V}{T}. \end{aligned}$$

■

Appendix C. Proofs for Section 5.1

In this section we give the proofs for Lemma 6 and Theorem 7 in Section 5.1. A few auxiliary lemmas are also given.

Proof of Lemma 6. Consider $\mathcal{F}$-measurable random variables A, B, with $\mathcal{F}$ being some σ-algebra. If B is also $\mathcal{G}$-measurable for some $\mathcal{G} \subset \mathcal{F}$ such that $\mathbb{E}[A|\mathcal{G}] = B$ almost surely, then we have:

$$\mathbb{E}[A-B] = 0; \qquad \text{and} \qquad \mathbb{E}[B(A-B)] = 0$$

(Note that $\mathbb{E}[B(A-B)|\mathcal{G}] = B\mathbb{E}[A-B|\mathcal{G}] = 0$, almost surely). This gives us

$$\begin{aligned} \mathrm{Var}(A) &= \mathbb{E}\left[(A-\mathbb{E}[A])^2\right] \\ &= \mathbb{E}\left[((B-\mathbb{E}[B]) + (A-B) - \mathbb{E}[A-B])^2\right] \\ &= \mathbb{E}\left[((B-\mathbb{E}[B]) + (A-B))^2\right] \\ &= \mathbb{E}\left[(B-\mathbb{E}[B])^2 + 2(B-\mathbb{E}[B])(A-B) + (A-B)^2\right] \\ &= \mathbb{E}\left[(B-\mathbb{E}[B])^2\right] + 2\mathbb{E}[B(A-B)] - 2\mathbb{E}[B]\mathbb{E}[A-B] + \mathbb{E}\left[(A-B)^2\right] \\ &= \mathrm{Var}(B) + \mathbb{E}\left[(A-B)^2\right]. \end{aligned} \tag{36}$$

Now choosing $\mathcal{F}$ to be the smallest σ-algebra such that the random variable $(X_0,\ldots,X_{T-1},J_0,\ldots,J_{T-1})$ and our functions $\mathrm{f},\mathrm{J},\mathrm{a}$, for all X_t, are measurable, and $\mathcal{G}$ such that $(X_0,\ldots,X_{T-1})$, and the functions on X_t, are measurable, we have that for

$$A=\frac{1}{T}\sum_{t=0}^{T-1}\mathrm{f}(X_t)\left(J_t-\mathrm{a}(X_t)\right)\qquad\text{and}\qquad B=\frac{1}{T}\sum_{t=0}^{T-1}\mathrm{f}(X_t)\left(\mathrm{J}(X_t)-\mathrm{a}(X_t)\right),$$

A and B are $\mathcal{F}$-measurable, B is $\mathcal{G}$-measurable, and $\mathcal{G}\subset\mathcal{F}$. Furthermore, we have

$$\begin{aligned}
\mathbb{E}[A|\,\mathcal{G}] &= \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\mathrm{f}(X_t)\left(J_t-\mathrm{a}(X_t)\right)\middle|X_0,\ldots,X_{T-1}\right]\\
&= \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\mathrm{f}(X_t)\left(J_t-\mathrm{a}(X_t)\right)|X_t\right]\\
&= \frac{1}{T}\sum_{t=0}^{T-1}\mathrm{f}(X_t)\left(\mathbb{E}\left[J_t|X_t\right]-\mathrm{a}(X_t)\right)\\
&= \frac{1}{T}\sum_{t=0}^{T-1}\mathrm{f}(X_t)\left(\mathrm{J}(X_t)-\mathrm{a}(X_t)\right)\\
&= B.
\end{aligned}$$

The proof then follows from Equation 36. ■

The proof of Theorem 7 requires some additional tools. In addition to (10) we also consider a variation of GPOMDP where a fixed length chain is used to estimate the discounted value function:

$$\Delta_T^{(S)}\stackrel{\text{def}}{=}\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}J_{t+1}^{(S)},\qquad J_t^{(S)}\stackrel{\text{def}}{=}\sum_{s=t}^{t+S-1}\beta^{s-t}\mathrm{r}(X_s).$$

Lemma 24. *Let $D=(\mathcal{S},\mathcal{U},\mathcal{Y},P,\nu,\mathrm{r},\mu)$ be a controlled POMDP satisfying Assumptions 1, 2 and 3. Then*

$$\left\|\Delta_T^{(+S)}-\Delta_T^{(S)}\right\|\leq\frac{\mathbf{BR}}{1-\beta}\beta^S,$$

and similarly,

$$\left\|\Delta_T^{(\infty)}-\Delta_T^{(S)}\right\|\leq\frac{\mathbf{BR}}{1-\beta}\beta^S,$$

where $\Delta_T^{(\infty)}$ denotes $\Delta_T^{(S)}$ in the limit as $S\to\infty$.

Proof.

$$\begin{aligned}
\left\|\Delta_T^{(+S)}-\Delta_T^{(S)}\right\| &= \left\|\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}J_{t+1}^{(+S)}-\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}J_{t+1}^{(S)}\right\|\\
&= \left\|\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\sum_{s=t+1+S}^{c}\beta^{s-t-1}\mathrm{r}(X_s)\right\|,\qquad c=T+S
\end{aligned}$$

$$\begin{aligned}
&\leq \mathbf{BR}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{s=t+1+S}^{c}\beta^{s-t-1}\\
&= \mathbf{BR}\frac{1}{T}\sum_{t=0}^{T-1}\frac{\beta^S(1-\beta^{c-S-t-1})}{1-\beta}\\
&\leq \frac{\mathbf{BR}}{1-\beta}\beta^S.
\end{aligned}$$

Obtain the bound $\left\|\Delta_T^{(\infty)}-\Delta_T^{(S)}\right\|$ similarly by considering the limit as $c\to\infty$. ■

Lemma 25. *Let $D=(\mathcal{S},\mathcal{U},\mathcal{Y},P,\nu,r,\mu)$ be a controlled POMDP satisfying Assumptions 1, 2 and 3. Let $\{Z_t\}=\{X_t,Y_t,U_t,X_{t+1}\}$ be the process generated by D. For any $a:\mathcal{S}\times\mathcal{Y}\times\mathcal{U}\times\mathcal{S}\to\mathbb{R}$ satisfying $|a(\cdot)|\leq\mathbf{M}$, we have*

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(J_{t+1}^{(+S)}-a(Z_t)\right)\right) &\leq \mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(J_{t+1}^{(S)}-a(Z_t)\right)\right)\\
&\quad+\frac{5\mathbf{B}^2\mathbf{R}(\mathbf{R}+\mathbf{M})}{(1-\beta)^2}\beta^S
\end{aligned}$$

and similarly,

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(J_{t+1}^{(S)}-a(Z_t)\right)\right) &\leq \mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(J_{t+1}^{(\infty)}-a(Z_t)\right)\right)\\
&\quad+\frac{5\mathbf{B}^2\mathbf{R}(\mathbf{R}+\mathbf{M})}{(1-\beta)^2}\beta^S,
\end{aligned}$$

where $J_t^{(\infty)}$ denotes $J_t^{(S)}$ in the limit as $S\to\infty$.

Proof.

$$\begin{aligned}
&\mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(J_{t+1}^{(+S)}-a(Z_t)\right)\right)\\
&= \mathbb{E}\left(\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(J_{t+1}^{(+S)}-a(Z_t)\right)-\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(J_{t+1}^{(+S)}-a(Z_t)\right)\right]\right)^2\\
&= \mathbb{E}\left(\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(J_{t+1}^{(S)}-a(Z_t)\right)-\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(J_{t+1}^{(S)}-a(Z_t)\right)\right]\right.\\
&\qquad\left.+\left(\Delta_T^{(+S)}-\Delta_T^{(S)}\right)-\mathbb{E}\left[\Delta_T^{(+S)}-\Delta_T^{(S)}\right]\right)^2\\
&= \mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(J_{t+1}^{(S)}-a(Z_t)\right)\right)+\mathbb{E}\left(\Delta_T^{(+S)}-\Delta_T^{(S)}\right)^2-\left(\mathbb{E}\left[\Delta_T^{(+S)}-\Delta_T^{(S)}\right]\right)^2\\
&\quad+2\mathbb{E}\left[\left(\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(J_{t+1}^{(S)}-a(Z_t)\right)-\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(J_{t+1}^{(S)}-a(Z_t)\right)\right]\right)'\right.\\
&\qquad\left.\times\left(\Delta_T^{(+S)}-\Delta_T^{(S)}\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(J_{t+1}^{(S)}-\mathrm{a}(Z_t)\right)\right)+\mathbb{E}\left\|\Delta_T^{(+S)}-\Delta_T^{(S)}\right\|^2 \\
&\quad+2\mathbb{E}\left[\left(\left\|\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(J_{t+1}^{(S)}-\mathrm{a}(Z_t)\right)\right\|+\mathbb{E}\left\|\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(J_{t+1}^{(S)}-\mathrm{a}(Z_t)\right)\right\|\right)\right. \\
&\qquad\left.\times\left\|\Delta_T^{(+S)}-\Delta_T^{(S)}\right\|\right] \\
&\leq \mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(J_{t+1}^{(S)}-\mathrm{a}(Z_t)\right)\right)+\left(\frac{\mathbf{BR}}{1-\beta}\beta^S\right)^2 \\
&\quad+4\mathbb{E}\left\|\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(J_{t+1}^{(S)}-\mathrm{a}(Z_t)\right)\right\|\frac{\mathbf{BR}}{1-\beta}\beta^S \qquad \text{(Lemma 24)} \\
&\leq \mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(J_{t+1}^{(S)}-\mathrm{a}(Z_t)\right)\right)+\left(\frac{\mathbf{BR}}{1-\beta}\beta^S\right)^2 \\
&\quad+4\left(\frac{\mathbf{B}(\mathbf{R}+\mathbf{M}(1-\beta))}{1-\beta}\right)\left(\frac{\mathbf{BR}}{1-\beta}\beta^S\right) \\
&\leq \mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(J_{t+1}^{(S)}-\mathrm{a}(Z_t)\right)\right)+\frac{5\mathbf{B}^2\mathbf{R}(\mathbf{R}+\mathbf{M})}{(1-\beta)^2}\beta^S.
\end{aligned}
$$

Obtain the second result by replacing $J_t^{(+S)}$ with $J_t^{(S)}$, and $J_t^{(S)}$ with $J_t^{(\infty)}$; then $\Delta_T^{(+S)}-\Delta_T^{(S)}$ becomes $\Delta_T^{(S)}-\Delta_T^{(\infty)}$. ■

Using these Lemmas, and Theorem 4, we can now prove Theorem 7.

Proof of Theorem 7. In this proof we will apply Theorem 4 to show that the variance of the sample average is $O(\ln(T)/T)$ times the variance of a single sample, and we will apply Lemma 6 to show that the additional variance due to estimating the value function need not be considered. We first use Lemma 25 to convert each of the samples within the average to be functions on a fixed length of the chain, that is, functions on states of the Markov process $\{X_t, Y_t, U_t, \ldots, U_{t+S-1}, X_{t+S}\}$. We can then use Theorem 4 for the sample average of functions on this process. Write

$$V = \mathrm{Var}_\pi\left(\frac{\nabla\mu_u(y)}{\mu_u(y)}\left(\mathrm{J}_\beta(j)-\mathrm{a}(i,y,u,j)\right)\right),$$

$$E = \mathbb{E}_\pi\left[\left(\frac{\nabla\mu_u(y)}{\mu_u(y)}\left(g(j)-\mathrm{J}_\beta(j)\right)\right)^2\right],$$

and

$$C = \frac{5\mathbf{B}^2\mathbf{R}(\mathbf{R}+\mathbf{M})}{(1-\beta)^2}\beta^S,$$

note that

$$1+\left\|\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(J_{t+1}^{(S)}-\mathrm{a}(Z_t)\right)\right\|_\infty \leq \frac{1}{7}\cdot\frac{C_1}{1-\beta},$$

where $\|a\|_\infty$ is the maximum of the magnitudes of the components of vector a, and denote the mixing time of the process $\{X_t, Y_t, U_t, \ldots, U_{t+S-1}, X_{t+S}\}$ by $\tilde{\tau}$. We have

$$\begin{aligned}
&\mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(J_{t+1}^{(+S)}-\mathrm{a}(Z_t)\right)\right)\\
&\leq\ \mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(J_{t+1}^{(S)}-\mathrm{a}(Z_t)\right)\right)+C \qquad \text{(Lemma 25)}\\
&\leq\ K\varepsilon+\left(1+\frac{25}{7}\tilde{\tau}\frac{C_1}{1-\beta}\varepsilon+4\tilde{\tau}\ln\frac{1}{\varepsilon}\right)\frac{1}{T}\mathrm{Var}\left(\frac{\nabla\mu_{U_0}(Y_0)}{\mu_{U_0}(Y_0)}\left(J_1^{(S)}-\mathrm{a}(Z_0)\right)\right)\\
&\qquad +C \qquad \text{(Theorem 4)}\\
&\leq\ K\varepsilon+\left(1+\frac{25}{7}\tilde{\tau}\frac{C_1}{1-\beta}\varepsilon+4\tilde{\tau}\ln\frac{1}{\varepsilon}\right)\frac{1}{T}\mathrm{Var}\left(\frac{\nabla\mu_{U_0}(Y_0)}{\mu_{U_0}(Y_0)}\left(J_1^{(\infty)}-\mathrm{a}(Z_0)\right)\right)\\
&\qquad +\left(1+\frac{25}{7}\tilde{\tau}\frac{C_1}{1-\beta}\varepsilon+4\tilde{\tau}\ln\frac{1}{\varepsilon}\right)\frac{C}{T}+C \qquad \text{(Lemma 25)}\\
&=\ K\varepsilon+\left(1+\frac{25}{7}\tilde{\tau}\frac{C_1}{1-\beta}\varepsilon+4\tilde{\tau}\ln\frac{1}{\varepsilon}\right)\left(\frac{V}{T}+\frac{E}{T}+\frac{C}{T}\right)+C \qquad \text{(Lemma 6).}
\end{aligned}$$

Here, Theorem 4 was applied to each of the K dimensions of the quantity the variance is taken over (recall that we consider the variance of a vector quantity to be the sum of the variance of its components). Now, similar to the proof of Corollary 5, we choose

$$\frac{1}{\varepsilon}=\frac{K}{4\tilde{\tau}}\left(\frac{V}{T}+\frac{E}{T}+\frac{C}{T}\right)^{-1}+\frac{25}{28}\cdot\frac{C_1}{1-\beta},$$

giving

$$\begin{aligned}
&\mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu_{U_t}(Y_t)}{\mu_{U_t}(Y_t)}\left(J_{t+1}^{(+S)}-\mathrm{a}(Z_t)\right)\right)\\
&\leq\ K\varepsilon+\left(1+\frac{25}{7}\tilde{\tau}\frac{C_1}{1-\beta}\varepsilon+4\tilde{\tau}\ln\frac{1}{\varepsilon}\right)\left(\frac{V}{T}+\frac{E}{T}+\frac{C}{T}\right)+C\\
&\leq\ 4\tilde{\tau}\left(\frac{V}{T}+\frac{E}{T}+\frac{C}{T}\right)+\Bigg[1+4\tilde{\tau}\\
&\qquad +4\tilde{\tau}\ln\left(\frac{25}{28}\cdot\frac{C_1}{1-\beta}+\frac{K}{4\tilde{\tau}}\left(\frac{V}{T}+\frac{E}{T}+\frac{C}{T}\right)^{-1}\right)\Bigg]\left(\frac{V}{T}+\frac{E}{T}+\frac{C}{T}\right)+C\\
&\leq\ \mathrm{h}\left(\frac{\tilde{\tau}}{T}V\right)+\mathrm{h}\left(\frac{\tilde{\tau}}{T}E\right)+\mathrm{h}\left(\frac{\tilde{\tau}}{T}C\right)+C\\
&\leq\ \mathrm{h}\left(\frac{\tau\ln(e(S+1))}{T}V\right)+\mathrm{h}\left(\frac{\tau\ln(e(S+1))}{T}E\right)+\mathrm{h}\left(\frac{\tau\ln(e(S+1))}{T}C\right)+C,
\end{aligned}$$

where the last line follows from $\tilde{\tau}\leq\tau\ln(e(S+1))$ (Lemma 1), and from h being an increasing function. Lastly, we have

$$\mathrm{h}\left(\frac{\tau\ln(e(S+1))}{T}C\right)+C$$

$$
\begin{aligned}
&\leq \left(\frac{1}{T}+8\tau\frac{\ln e(S+1)}{T}\right.\\
&\quad \left.+4\tau\frac{\ln(e(S+1))}{T}\ln\left(\frac{C_1}{1-\beta}+\frac{K(1-\beta)^2}{20\tau\mathbf{B}^2\mathbf{R}(\mathbf{R}+\mathbf{M})\ln(e(S+1))}\left(\frac{\beta^S}{T}\right)^{-1}\right)+1\right)C\\
&\leq \left(\frac{1}{T}+8\tau\frac{\ln e(S+1)}{T}+4\tau\frac{\ln(T)\ln(e(S+1))}{T}\right.\\
&\quad \left.+4\tau\frac{S\ln(e(S+1))}{T}\ln\frac{1}{\beta}+4\tau\frac{\ln(e(S+1))}{T}\ln\left(\frac{C_1}{1-\beta}+\frac{K(1-\beta)^2}{20\tau\mathbf{B}^2\mathbf{R}(\mathbf{R}+\mathbf{M})}\right)+1\right)C\\
&\leq \frac{2C_2}{(1-\beta)^2}\left[\ln\frac{1}{\beta}+\ln\left(\frac{C_1}{1-\beta}+\frac{K(1-\beta)^2}{C_2}\right)\right]\frac{(T+S)\ln(e(S+1))}{T}\beta^S.
\end{aligned}
$$

The second step has used the increasing property of ln, along with $\ln(e(S+1)) \geq 1$ and $\beta^S/T \leq 1$. This gives us, for any $A, B \geq 0$,

$$
\ln\left(A+\frac{B}{\ln(e(S+1))}\left(\frac{\beta^S}{T}\right)^{-1}\right) \leq \ln\left(A+B\left(\frac{\beta^S}{T}\right)^{-1}\right) \leq \ln\left((A+B)\left(\frac{\beta^S}{T}\right)^{-1}\right).
$$

■

Appendix D. Value Function Example

Here we consider a somewhat less trivial example than that presented in Section 6.2—an example of reducing variance through appropriate choice of value function. A toy MDP is shown in Figure 7. Here action a_1 causes the MDP to have a tendency to stay in state s_1, and action a_2 causes the MDP to have a tendency to move away from s_1 and stay in state s_2 and s_3.

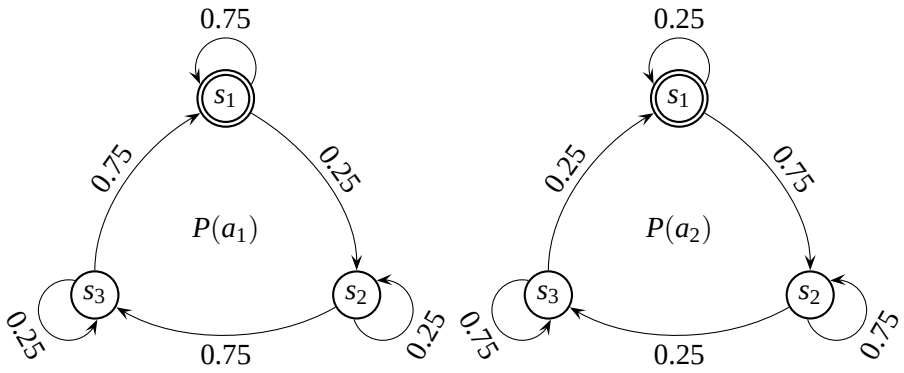


Figure 7: Transition probabilities for a toy 3 state, 2 action Markov decision process

Now consider the resultant controlled MDP when the single parameter, state independent policy

$$
\mu_{a_1} = \frac{e^\theta}{e^\theta + e^{-\theta}} \qquad \mu_{a_2} = 1 - \mu_{a_1} = \frac{e^{-\theta}}{e^\theta + e^{-\theta}}
$$

along with any reward function satisfying Assumption 2 is used. Note that this controlled MDP satisfies Assumptions 1, 2 and 3 for all θ. For the policy at $\theta = 0$ we have $\mu_{a_1} = \mu_{a_2} = 0.5$ and

$$
\nabla\mu_{a_1} = 0.5 \qquad \nabla\mu_{a_2} = -0.5.
$$

The transition matrix and stationary distribution of the resultant chain are:

$$P = \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \end{bmatrix} \qquad \pi = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix}.$$

In this case the 1×3 matrix $G = (1/6, -1/6, 0)$, and the right null space of G is $\{\alpha_1 v_1 + \alpha_2 v_2 : \alpha_1, \alpha_2 \in \mathbb{R}\}$, where

$$v_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \qquad v_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Any value function of the form $\mathrm{V} = \mathrm{J}_\beta + \alpha_1 v_1 + \alpha_2 v_2$ will produce an unbiased estimate of $\nabla_\beta \eta$. In this case we have that, writing $r_i = \mathrm{r}(s_i)$,

$$\mathrm{J}_\beta = (I - \beta P)^{-1} \mathrm{r} = \frac{2}{(2-\beta)^3 - \beta^3} \begin{bmatrix} (2-\beta)^2 & \beta(2-\beta) & \beta^2 \\ \beta^2 & (2-\beta)^2 & \beta(2-\beta) \\ \beta(2-\beta) & \beta^2 & (2-\beta)^2 \end{bmatrix} \begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix}.$$

If we select $\beta = 0.9$ this becomes

$$\mathrm{J}_{0.9} = \frac{1}{0.301} \begin{bmatrix} 1.21 & 0.99 & 0.81 \\ 0.81 & 1.21 & 0.99 \\ 0.99 & 0.81 & 1.21 \end{bmatrix} \begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix} = \frac{1}{0.301} \begin{pmatrix} 1.21 r_1 + 0.99 r_2 + 0.81 r_3 \\ 0.81 r_1 + 1.21 r_2 + 0.99 r_3 \\ 0.99 r_1 + 0.81 r_2 + 1.21 r_3 \end{pmatrix}.$$

If we had $\mathrm{r} = (1/10, 2/11, 0)'$ then we would again have $\mathrm{J}_{0.9} [= (1, 1, 81/99)']$ in the right null space of G, and we could again choose $\mathrm{V} = 0$ to obtain a zero bias, zero variance estimate of $\nabla_\beta \eta$. Consider instead the reward function

$$\mathrm{r}(i) = \begin{cases} 4.515 & i = s_1 \\ 0 & \text{otherwise,} \end{cases}$$

so that $\mathrm{J}_{0.9} = (18.15, 12.15, 14.85)'$ and $\nabla_{0.9} \eta = 1$. We now have

$$\begin{aligned} \mathrm{Var}_\pi \left(\frac{\nabla \mu_u(i)}{\mu_u(i)} \mathrm{J}_{0.9}(j) \right) &= \mathbb{E}_\pi \left(\frac{\nabla \mu_u(i)}{\mu_u(i)} \mathrm{J}_{0.9}(j) \right)^2 - \left(\mathbb{E}_\pi \left[\frac{\nabla \mu_u(i)}{\mu_u(i)} \mathrm{J}_{0.9}(j) \right] \right)^2 \\ &= \mathbb{E}_\pi \left(\mathrm{J}_{0.9}(j) \right)^2 - 1 \\ &= \pi' \begin{pmatrix} 18.15^2 \\ 12.15^2 \\ 14.85^2 \end{pmatrix} - 1 \\ &= 231.5225. \end{aligned}$$

The second line is obtained from $|\nabla \mu_u(i) / \mu_u(i)| = 1$ and $\nabla_{0.9} \eta = 1$. If we choose $\alpha_1 = -15.15\sqrt{2}$ and $\alpha_2 = -14.85$ then, for the value function $\mathrm{V} = \mathrm{J}_\beta + \alpha_1 v_1 + \alpha_2 v_2$, we have

$$\begin{aligned} \mathrm{Var}_\pi \left(\frac{\nabla \mu_u(i)}{\mu_u(i)} \mathrm{V}(j) \right) &= \mathbb{E}_\pi \left(\frac{\nabla \mu_u(i)}{\mu_u(i)} \mathrm{V}(j) \right)^2 - \left(\mathbb{E}_\pi \left[\frac{\nabla \mu_u(i)}{\mu_u(i)} \mathrm{V}(j) \right] \right)^2 \\ &= \pi' \begin{pmatrix} (18.15 - 15.15)^2 \\ (12.15 - 15.15)^2 \\ 0 \end{pmatrix} - 1 \\ &= 5; \end{aligned}$$

a significant reduction in variance, with no additional bias.

References

D. Aberdeen. A survey of approximate methods for solving partially observable markov decision processes. Technical report, Research School of Information Science and Engineering, Australian National University, Australia, 2002.

L. Baird. Gradient descent for general reinforcement learning. In S. A. Solla M. S Kearns and D. A. Cohn, editors, *Advances in Neural Information Processing Systems*, volume 11. The MIT Press, 1999.

P. L. Bartlett and J. Baxter. Estimation and approximation bounds for gradient-based reinforcement learning. *Journal of Computer and System Sciences*, 64(1):133–150, February 2002.

A. G. Barto, R. S. Sutton, and C. W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, 13:834–846, 1983.

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

J. Baxter, P. L. Bartlett, and L. Weaver. Experiments with infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:351–381, 2001.

S. J. Bradtke and A. Barto. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 262:33–57, 1996.

P. Dayan. Reinforcement comparison. In *Proceedings of the 1990 Connectionist Models Summer School*, pages 45–51. Morgan Kaufmann, 1990.

J. L. Doob. *Measure Theory*. Number 143 in Graduate Texts in Mathematics. Springer-Verlag, New York, 1994.

M. Evans and T. Swartz. *Approximating integrals via Monte Carlo and deterministic methods*. Oxford statistical science series. Oxford University Press, Oxford; New York, 2000.

G. S. Fishman. *Monte Carlo: Concepts, Algorithms and Applications*. Springer series in operations research. Springer-Verlag, New York, 1996.

P. W. Glynn. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM*, 33(10):75–84, 1990.

P. W. Glynn and P. L'Ecuyer. Likelihood ratio gradient estimation for regenerative stochastic recursions. *Advances in Applied Probability*, 12(4):1019–1053, 1995.

G. R. Grimmett and D. R. Stirzaker. *Probability and Random Processes*. Oxford University Press, Oxford, 1992.

J. M. Hammersley and D. C. Handscomb. *Monte Carlo Methods*. Chapman and Hall, New York, 1965.

T. Jaakkola, S. P. Singh, and M. I. Jordan. Reinforcement learning algorithm for partially observable Markov decision problems. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 6, pages 345–352. The MIT Press, 1995.

L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1998.

H. Kimura and S. Kobayashi. An analysis of actor/critic algorithms using eligibility traces: Reinforcement learning with imperfect value functions. In *International Conference on Machine Learning*, pages 278–286, 1998a.

H. Kimura and S. Kobayashi. Reinforcement learning for continuous action using stochastic gradient ascent. In *Intelligent Autonomous Systems*, volume 5, pages 288–295, 1998b.

H. Kimura, K. Miyazaki, and S. Kobayashi. Reinforcement learning in POMDPs with function approximation. In D. H. Fisher, editor, *International Conference on Machine Learning*, pages 152–160, 1997.

H. Kimura, M. Yamamura, and S. Kobayashi. Reinforcement learning by stochastic hill climbing on discounted reward. In *International Conference on Machine Learning*, pages 295–303, 1995.

V. R. Konda and J. N. Tsitsiklis. Actor-critic algorithms. In T. K. Leen S. A. Solla and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. The MIT Press, 2000.

V. R. Konda and J. N. Tsitsiklis. On actor-critic algorithms. *SIAM Journal on Control and Optimization*, 42(4):1143–1166, 2003.

W. S. Lovejoy. A survey of algorithmic methods for partially observed markov decision processes. *Annals of Operations Research*, 28:47–66, 1991.

P. Marbach and J. N. Tsitsiklis. Simulation-based optimization of markov reward processes. *IEEE Transactions on Automatic Control*, 46(2):191–209, February 2001.

M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley series in probability and mathematical statistics. Applied probability and statistics. John Wiley & Sons, New York, 1994.

M. I. Reiman and A. Weiss. Sensitivity analysis for simulations via likelihood ratios. *Operations Research*, 37, 1989.

R. Y. Rubinstein. How to optimize complex stochastic systems from a single sample path by the score function method. *Annals of Operations Research*, 27:175–211, 1991.

E. Seneta. *Non-negative Matrices and Markov Chains*. Springer series in statistics. Springer-Verlag, New York, 1981.

S. P. Singh, T. Jaakkola, and M. I. Jordan. Learning without state-estimation in partially observable markovian decision processes. In *International Conference on Machine Learning*, pages 284–292, 1994.

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

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge MA, 1998.

R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In T. K. Leen S. A. Solla and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. The MIT Press, 2000.

F. Topsøe. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on Information Theory*, 46:1602–1609, 2000.

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