Optimizing Average Reward Using Discounted Rewards

Sham Kakade

Gatsby Computational Neuroscience Unit
17 Queen Square
London WC1N 3AR
United Kingdom
sham@gatsby.ucl.ac.uk
http://www.gatsby.ucl.ac.uk/~sham/index.html

Abstract. In many reinforcement learning problems, it is appropriate to optimize the average reward. In practice, this is often done by solving the Bellman equations using a discount factor close to 1. In this paper, we provide a bound on the average reward of the policy obtained by solving the Bellman equations which depends on the relationship between the discount factor and the mixing time of the Markov chain. We extend this result to the direct policy gradient of Baxter and Bartlett, in which a discount parameter is used to find a biased estimate of the gradient of the average reward with respect to the parameters of a policy. We show that this biased gradient is an exact gradient of a related discounted problem and provide a bound on the optima found by following these biased gradients of the average reward. Further, we show that the exact Hessian in this related discounted problem is an approximate Hessian of the average reward, with equality in the limit the discount factor tends to 1. We then provide an algorithm to estimate the Hessian from a sample path of the underlying Markov chain, which converges with probability 1.

1 Introduction

Sequential decision making problems are usually formulated as dynamic programming problems in which the agent must maximize some measure of future reward. In many domains, it is appropriate to optimize the average reward. Often, discounted formulations with a discount factor γ close to 1 are used as a proxy to an average reward formulation. It is natural to inquire about the consequences of using a discount factor close to one. How does the quality of the policy, measured in an average reward sense, degrade as the discount factor is reduced? What are the benefits in using a smaller discount factor?

This papers focuses on the former issue by extending the results of Baxter and Bartlett[2, 1]. A key relationship proved in [2] shows that the discounted reward, scaled by $1 - \gamma$, is approximately the average reward, which suggests that maximizing discounted reward will be approximately maximizing average reward. We show if $\frac{1}{1-\gamma}$ is large compared to the mixing time of the Markov chain, then we would expect any policy that solves the Bellman equations to have a large average reward.

This interpretation of using discounted rewards to maximize average reward extends to the case of maximizing the average reward by following a gradient. We show that the approximate gradient of Baxter and Bartlett is an exact gradient of a related discounted, start state problem and provide a similar bound on the quality of the optima reached using this approximate gradient. These results naturally lead to an algorithm for computing an approximation to the Hessian. A slightly different, independent derivation of the Hessian is given in [3].

2 The Reinforcement Learning Problem

We consider the standard formulation of reinforcement learning, in which an agent interacts with a finite Markov decision process (MDP). An MDP is a tuple (S, A, R, P) where: S is finite set of states $S = \{1, \dots, n\}$, A is a finite set of actions, R is a reward function $R : S \to [0, R_{max}]$[1], and P is the transition model in which $p_{ij}(u)$ is the probability of transitioning to state j from state i under action u.[2]

The agent's decision making procedure is characterized by a stochastic policy $\mu : S \to A$, where $\mu_u(i)$ is the probability of taking action u in state i. For each policy μ there corresponds a Markov chain with a transition matrix $P(\mu)$, where $[P(\mu)]_{ij} = \sum_u p_{ij}(u)\mu_u(i)$. We assume that these Markov chains satisfy the following assumption:

Assumption 1. *Each $P(\mu)$ has a unique stationary distribution $\pi(\mu) \equiv [\pi(\mu, 1), \dots, \pi(\mu, n)]'$ satisfying:*

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

(where $\pi(\mu)'$ denotes the transpose of $\pi(\mu)$).

The *average reward* is defined by:

$$\eta(\mu) \equiv \lim_{N \to \infty} \frac{1}{N} \sum_i \pi(\mu, i) E_\mu\{\sum_{t=0}^{N-1} r(i_t) | i_0 = i\}$$

where i_t is the state at time t. The average reward can be shown to equal:

$$\eta(\mu) = \pi(\mu)' r$$

[1] It is a straightforward to extend these results to the case where the rewards are dependent on the actions, $R(s, a)$.

[2] We ignore the start state distribution, since the average reward is independent of the starting distribution under the current assumption of a unique stationary distribution.

where $r = [r(1), ..., r(n)]'$ (see [4]). The goal of the agent is to find a policy μ^* that returns the maximum average reward over all policies.

We define the γ-*discounted reward* from some starting state i as:

$$J_\gamma(\mu, i) \equiv E_\mu\{\sum_{t=0}^{\infty} \gamma^t r(i_t)|i_0 = i\}$$

where $\gamma \in [0, 1)$. These value functions satisfy the following consistency condition [8]:

$$J_\gamma(\mu) = r + \gamma P(\mu) J_\gamma(\mu) \tag{1}$$

(again we use the vector notation $J_\gamma(\mu) = [J_\gamma(\mu, 1), ..., J_\gamma(\mu, n)]'$). The *expected discounted reward* is defined as $\pi(\mu)' J_\gamma(\mu)$, where the expectation is taken over the stationary distribution . As shown in [6], the expected discounted reward is just a multiple of the average reward:

$$\pi(\mu)' J_\gamma(\mu) = \frac{\eta(\mu)}{1-\gamma} \ . \tag{2}$$

Thus, optimizing the expected discounted reward for any γ is equivalent to optimizing the average reward. Also, for all states i, $\lim_{\gamma \to 1}(1-\gamma) J_\gamma(i) = \eta$ (see [4]).

In discounted dynamic programming, we are concerned with finding a vector $J_\gamma \in \Re^n$ that satisfies the Bellman equations:

$$J_\gamma = \max_\mu (r + \gamma P(\mu) J_\gamma) \ . \tag{3}$$

Let $\mu^{\gamma *}$be a policy such that $J_\gamma(\mu^{\gamma *})$ satisfies this equation (there could be more than one such policy). Although the policy $\mu^{\gamma *}$ simultaneously maximizes the discounted reward starting from every state, it does not necessarily maximize the average discounted reward, which is sensitive to the stationary distribution achieved by this policy (see equation 2). The policies that solve the Bellman equations could lead to poor stationary distributions that do not maximize the average reward.

3 Appropriateness of Maximizing Discounted Reward

We extend the results of Baxter and Bartlett to show that if $\frac{1}{1-\gamma}$ is large compared to the mixing time of the Markov chain of the optimal policy then the solutions to the Bellman equations will have an average reward close to the maximum average reward. We use the following relation (shown by Baxter and Bartlett [2], modulo a typo), for any policy:

$$(1-\gamma) J_\gamma(\mu) = \eta(\mu) e + S(\mu)\, diag(0, \frac{1-\gamma}{1-\gamma|\lambda_2(\mu)|}, \ldots, \frac{1-\gamma}{1-\gamma|\lambda_n(\mu)|}) S(\mu)^{-1} r \tag{4}$$

where $e = [1, 1, \ldots, 1]'$ and $S(\mu) = (s_1 s_2 \cdots s_n)$ is the matrix of right eigenvectors of $P(\mu)$ with the corresponding eigenvalues $\lambda_1(\mu) = 1 >| \lambda_2(\mu)| \geq \cdots \geq$

$|\lambda_n(\mu)|$ (assuming that $P(\mu)$ has n distinct eigenvalues). This equation follows from separating $J_\gamma = \sum_{t=0}^{\infty} \gamma^t P^t r$ into the contribution associated with $\lambda_1 = 1$ and that coming from the remaining eigenvalues. Note that for γ near 1, the scaled discounted value function for each state is approximately the average reward, with an approximation error of order $1 - \gamma$.

Throughout the paper, $\|A\|_2$ denotes the spectral norm of a matrix A, defined as $\|A\|_2 \equiv \max_{x:\|x\|=1} \|Ax\|$, where$\|x\|$ denotes the Euclidean norm of x, and $\kappa_2(A)$ denotes the spectral condition number of a nonsingular matrix A, defined as $\kappa_2(A) \equiv \|A\|_2 \|A^{-1}\|_2$.

Theorem 1. *Let $\mu^{\gamma *}$be a policy such that $J_\gamma(\mu^{\gamma *})$ satisfies the Bellman equations (equation 3) and let μ^* be a policy such that $\eta(\mu^*)$ be the maximum average reward over all policies. Assume $P(\mu^*)$ has n distinct eigenvalues. Let $S = (s_1 s_2 \cdots s_n)$ be the matrix of right eigenvectors of $P(\mu^*)$ with the corresponding eigenvalues $\lambda_1 = 1 >| \lambda_2| \geq \cdots \geq |\lambda_n|$. Then*

$$\eta(\mu^{\gamma *}) \geq \eta(\mu^*) - \kappa_2(S)\|r\| \frac{1-\gamma}{1-\gamma|\lambda_2|} \; .$$

Proof. Since $J_\gamma(\mu^{\gamma *})$ satisfies the Bellman equations, we have

$$\forall \mu \;\; J_\gamma(\mu^{\gamma *}) \geq J_\gamma(\mu)$$

where the vector inequality is shorthand for the respective component wise inequality. As a special case, the inequality holds for a policy μ^* that maximizes the average reward, *ie* $J_\gamma(\mu^{\gamma *}) \geq J_\gamma(\mu^*)$. It follows from equation 2 and equation 4 (applied to μ^*) that

$$\begin{aligned}
\eta(\mu^{\gamma *}) &= (1-\gamma)\pi(\mu^{\gamma *})' J_\gamma(\mu^{\gamma *}) \\
&\geq (1-\gamma)\pi(\mu^{\gamma *})' J_\gamma(\mu^*) \\
&= \eta(\mu^*)\pi(\mu^{\gamma *})' e + \pi(\mu^{\gamma *})' S \operatorname{diag}(0, \frac{1-\gamma}{1-\gamma|\lambda_2|}, \ldots, \frac{1-\gamma}{1-\gamma|\lambda_n|}) S^{-1} r \\
&\geq \eta(\mu^*) - |\pi(\mu^{\gamma *})' S \operatorname{diag}(0, \frac{1-\gamma}{1-\gamma|\lambda_2|}, \ldots, \frac{1-\gamma}{1-\gamma|\lambda_n|}) S^{-1} r|
\end{aligned}$$

where we have used $\pi(\mu^{\gamma *})' e = 1$. The dependence of S and λ_i on μ^* is suppressed. Using the Cauchy-Schwartz inequality, we have

$$\begin{aligned}
\eta(\mu^{\gamma *}) &\geq \eta(\mu^*) - \|S\pi(\mu^{\gamma *})\| \, \|\operatorname{diag}(0, \frac{1-\gamma}{1-\gamma|\lambda_2|}, \ldots, \frac{1-\gamma}{1-\gamma|\lambda_n|}) S^{-1} r\| \\
&\geq \eta(\mu^*) - \|S\pi(\mu^{\gamma *})\| \, \|\operatorname{diag}(0, \frac{1-\gamma}{1-\gamma|\lambda_2|}, \ldots, \frac{1-\gamma}{1-\gamma|\lambda_n|})\|_2 \|S^{-1} r\| \; .
\end{aligned}$$

It is easy to show that $\|\operatorname{diag}(d_1, \ldots, d_n)\|_2 = \max_i |d_i|$. Using $\|\pi\| \leq 1$, it follows from the definition of the spectral norm and spectral condition number that $\|S\pi(\mu^{\gamma *})\| \, \|S^{-1} r\| \leq \kappa_2(S)\|r\|$. □

The previous theorem shows that if $1-\gamma$ is small compared to $1-|\lambda_2|$, then the solution to the Bellman equations will be close to the maximum average reward. Under assumption 1, from any initial state, the distribution of states of the Markov chain will converge at an exponential rate to the stationary distribution, and the rate of this will depend on the eigenvalues of the transition matrix. The second largest eigenvalue, $|\lambda_2|$, will determine an upper bound on this mixing time.

4 Direct Gradient Methods

A promising recent approach to finding the gradient of the average reward was presented by Baxter and Bartlett [2], where a discount parameter controls the bias and variance of the gradient estimate (also see a related approach by Marbach and Tsitsiklis [5]). We now relate this approximate gradient to an exact gradient for a modified discounted problem and provide a bound on the quality of the local optima reached by following this approximate gradient. To ensure the existence of certain gradients and the boundedness of certain random variables, we assume

Assumption 2. *The derivatives, ∇P_{ij} and $\nabla \mu_u(\theta, i)$, exist and the ratios, $\frac{\nabla P_{ij}}{P_{ij}}$ and $\frac{\nabla \mu_u(\theta,i)}{\mu_u(\theta,i)}$, are bounded by a constant for all $\theta \in \Re^k$.*

Let $\theta \in R^k$ be the parameters of a policy $\mu(\theta) : S \to A$, where $\mu_u(\theta, i)$ is the chance of taking action u in state i. These parameters implicitly parameterize the average reward, the stationary distribution, and the transition matrix, which we denote by $\eta(\theta)$, $\pi(\theta)$, and $P(\theta)$. Also let $J_\gamma(\theta)$ be the discounted value function under $P(\theta)$. The key result of Baxter and Bartlett shows that the exact gradient of the average reward, $\nabla\eta(\theta)$, can be approximated by

$$\nabla\eta(\theta) \approx \gamma\pi(\theta)'\nabla P(\theta) J_\gamma(\theta) \equiv \tilde{\nabla}_\gamma \eta(\theta)$$

where this approximation becomes exact as $\gamma \to 1$. We denote this approximate gradient by $\tilde{\nabla}_\gamma\eta(\theta)$ (the tilde makes it explicitly clear that $\tilde{\nabla}_\gamma$ is not differentiating with respect to γ). Further, they give an algorithm that estimates $\tilde{\nabla}_\gamma\eta(\theta)$ from a sample trajectory.

Before we state our theorem, we define $\nu_\gamma(\theta, \rho)$ to be the expected discounted reward received from a starting state chosen from the distribution ρ under $P(\theta)$, *ie*

$$\nu_\gamma(\theta, \rho) \equiv \rho' J_\gamma(\theta) \ .$$

Theorem 2. *Let $\nu_\gamma(\tilde{\theta}, \pi(\theta)) \equiv \pi(\theta)' J_\gamma(\tilde{\theta})$. Then*

$$(1-\gamma)\nabla_{\tilde{\theta}}\nu_\gamma(\tilde{\theta}, \pi(\theta))|_{\tilde{\theta}=\theta} = \tilde{\nabla}_\gamma\eta(\theta) \tag{5}$$

where $\nabla_{\tilde{\theta}}$ is the gradient with respect to $\tilde{\theta}$.

Proof. It follows from the fact that $\pi(\theta)$ is independent of $\tilde{\theta}$ that

$$\begin{aligned}\nabla_{\tilde{\theta}}\nu_\gamma(\tilde{\theta},\pi(\theta)) &= \pi(\theta)'\nabla_{\tilde{\theta}}J_\gamma(\tilde{\theta})\\ &= \pi(\theta)'\nabla_{\tilde{\theta}}(r+\gamma P(\tilde{\theta})J_\gamma(\tilde{\theta}))\\ &= \pi(\theta)'(\nabla_{\tilde{\theta}}P(\tilde{\theta})J_\gamma(\tilde{\theta})+\gamma P(\tilde{\theta})\nabla_{\tilde{\theta}}J_\gamma(\tilde{\theta}))\ .\end{aligned}$$

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

$$\begin{aligned}\nabla_{\tilde{\theta}}\nu_\gamma(\tilde{\theta},\pi(\theta))|_{\tilde{\theta}=\theta} &= \pi(\theta)'\nabla P(\theta)J_\gamma(\theta)+\gamma\pi(\theta)'P(\theta)\nabla_{\tilde{\theta}}J_\gamma(\tilde{\theta})|_{\tilde{\theta}=\theta}\\ &= \tilde{\nabla}_\gamma\eta(\theta)+\gamma\pi(\theta)'\nabla_{\tilde{\theta}}J_\gamma(\tilde{\theta})|_{\tilde{\theta}=\theta}\\ &= \tilde{\nabla}_\gamma\eta(\theta)+\gamma\nabla_{\tilde{\theta}}\nu_\gamma(\tilde{\theta},\pi(\theta))|_{\tilde{\theta}=\theta}\ .\end{aligned}$$

Collecting terms proves equation 5. □

Note that the approximate gradient at θ_1 is equivalent to the exact gradient in a start state problem under the starting distribution $\pi(\theta_1)$, whereas the approximate gradient at θ_2 is equivalent to the exact gradient in the start state problem with a different starting distribution, $\pi(\theta_2)$. If the approximate gradient is 0 at some point $\theta^{\gamma *}$ then this point will also be an extremum of the related problem, which allows us to make the following statement. In the following theorem, the basin of attraction of a maximum x of $f(x)$ is the set of all points which converge to x when taking infinitesimal steps in the direction of the gradient.

Theorem 3. *Let $\theta^{\gamma *}$ be a point such that $\nabla\nu_\gamma(\theta,\pi(\theta^{\gamma *}))|_{\theta=\theta^{\gamma *}}=0$. Assume that this extremum is a local maximum and let Ω be the basin of attraction of this maximum with respect to $\nu_\gamma(\theta,\pi(\theta^{\gamma *}))$. Let $\theta^* \in \Omega$ such that $\eta(\theta^*)$ is the maximum average reward over all θ in Ω. Assume $P(\mu^*)$ has n distinct eigenvectors. Let $S=(s_1 s_2 \cdots s_n)$ be the matrix of right eigenvectors of $P(\theta^*)$ with the corresponding eigenvalues $\lambda_1=1>|\ \lambda_2|\geq\cdots\geq|\lambda_n|$. Then*

$$\eta(\theta^{\gamma *})\geq\eta(\theta^*)-\kappa_2(S)\|r\|\frac{1-\gamma}{1-\gamma|\lambda_2|}\ .$$

Proof. By assumption that this is a local maximum,

$$\forall\theta\in\Omega\quad \pi(\theta^{\gamma *})'J_\gamma(\theta^{\gamma *})\geq\pi(\theta^{\gamma *})'J_\gamma(\theta)\ .$$

Let θ^* be a point in Ω which returns $\eta(\theta^*)$, the maximum average reward in Ω. As a special case, we have $\pi(\theta^{\gamma *})'J_\gamma(\theta^{\gamma *})\geq\pi(\theta^{\gamma *})'J_\gamma(\theta^*)$. Using equation 2,

$$\begin{aligned}\eta(\theta^{\gamma *}) &= (1-\gamma)\pi(\theta^{\gamma *})'J_\gamma(\theta^{\gamma *})\\ &\geq (1-\gamma)\pi(\theta^{\gamma *})'J_\gamma(\theta^*)\ .\end{aligned}$$

The remainder of the argument parallels the proof given in Theorem 1. □

The previous theorem gives a constraint on the quality of the maximum reached *if and when* the approximate gradient ascent converges. Note that this bound is essentially identical to the bound on the average reward of the policy obtained by solving the Bellman equations.

5 Direct Hessian Methods

Theorem 2 suggests that the natural choice for an approximate Hessian is $(1-\gamma)\nabla^2\nu_\gamma(\tilde{\theta},\pi(\theta))|_{\tilde{\theta}=\theta}$. We define

$$\tilde{\nabla}^2_\gamma \eta(\theta) \equiv (1-\gamma)\nabla^2_{\tilde{\theta}}\nu_\gamma(\tilde{\theta},\pi(\theta))|_{\tilde{\theta}=\theta} \ .$$

We make the following assumption.

Assumption 3. *The Hessians, $\nabla^2 P_{ij}$ and $\nabla^2\mu_u(\theta,i)$, exist and the ratios, $\frac{\nabla^2 P_{ij}}{P_{ij}}$ and $\frac{\nabla^2\mu_u(\theta,i)}{\mu_u(\theta,i)}$, are bounded by a constant for all $\theta \in \Re^k$.*

Theorem 4. *For all $\theta \in \Re^k$,*

$$\nabla^2\eta(\theta) = \lim_{\gamma\to 1}\tilde{\nabla}^2_\gamma\eta(\theta) \ .$$

Proof. Let $\partial_k \equiv \frac{\partial}{\partial\theta_k}$. Using equation 2 and suppressing the θ dependence,

$$\begin{aligned}
\lim_{\gamma\to 1}[\nabla^2\eta(\theta)]_{mn} &= \lim_{\gamma\to 1}(1-\gamma)\partial_m\partial_n(\pi' J_\gamma) \\
&= \lim_{\gamma\to 1}(1-\gamma)([\partial_m\partial_n\pi']J_\gamma + [\partial_m\pi'][\partial_n J_\gamma] + [\partial_n\pi'][\partial_m J_\gamma] \\
&\quad +\pi'[\partial_m\partial_n J_\gamma]) \\
&= \eta(\theta)\partial_m\partial_n\pi' e + \partial_n\eta\partial_m\pi' e + \partial_m\eta\partial_n\pi' e + \lim_{\gamma\to 1}(1-\gamma)\pi'\partial_m\partial_n J_\gamma \\
&= \lim_{\gamma\to 1}(1-\gamma)\pi'\partial_m\partial_n J_\gamma
\end{aligned}$$

where we have used $\lim_{\gamma\to 1}(1-\gamma)J_\gamma = \eta e$ (see [4]), $\lim_{\gamma\to 1}(1-\gamma)\partial_k J_\gamma = \partial_k\eta e$ (which is straightforward to prove), and $\partial_k\pi' e = \partial_k 1 = 0$. Following from the definition of ν_γ, $[\nabla^2_{\tilde{\theta}}\nu_\gamma(\tilde{\theta},\pi(\theta))|_{\tilde{\theta}=\theta}]_{mn} = \pi(\theta)'[\nabla^2 J_\gamma(\theta)]_{mn}$. □

The previous theorem shows that in the limit as γ tends to 1 the exact Hessian in the start state problem is the Hessian of the average reward. The following theorem gives an expression for $\tilde{\nabla}^2_\gamma\eta(\theta)$, which we later show how to estimate from Monte-Carlo samples.

Theorem 5. *For all θ and $\gamma \in [0,1)$,*

$$[\tilde{\nabla}^2_\gamma\eta(\theta)]_{mn} = \gamma\pi'([\partial_m\partial_n P]J_\gamma + [\partial_m P][\partial_n J_\gamma] + [\partial_n P][\partial_m J_\gamma]) \tag{6}$$

where

$$\partial_k J_\gamma = \sum_{t=1}^{\infty}\gamma^t[P^{t-1}\nabla P J_\gamma]_k \ . \tag{7}$$

Algorithm 1: HMDP (Hessian for a Markov Decision Process)

```
Obtain an arbitrary state $i_0$
Set $z_0 = \Delta_0 = 0 \in \Re^k$ and set $\tilde{z}_0 = y_0 = H_0 = 0 \in \Re^k \times \Re^k$
for $t = 0$ to $t = T - 1$ do
    Generate control $u_t$ according to $\mu_{u_t}(\theta, x_t)$
    Observe $r(x_{t+1})$ and $x_{t+1}$(generated according to $P_{x_t x_{t+1}}(u_t)$)
    $\mathrm{y}_{\mathrm{t+1}} = \gamma(y_t + \frac{\nabla \mu_{u_t}(\theta,x_t)}{\mu_{u_t}(\theta,x_t)} z_t' + z_t \frac{\nabla \mu_{u_t}(\theta,x_t)'}{\mu_{u_t}(\theta,x_t)})$
    $z_{t+1} = \gamma(z_t + \frac{\nabla \mu_{u_t}(\theta,x_t)}{\mu_{u_t}(\theta,x_t)})$
    $\tilde{z}_{t+1} = \gamma(\tilde{z}_t + \frac{\nabla^2 \mu_{u_t}(\theta,x_t)}{\mu_{u_t}(\theta,x_t)})$
    $\Delta_{t+1} = \Delta_t + r(x_{t+1}) z_{t+1}$
    $H_{t+1} = H_t + r(x_{t+1}) \tilde{z}_{t+1} + r(x_{t+1}) y_{t+1}$
end for
gradient $\Delta_T \leftarrow \Delta_T / T$
Hessian $H_T \leftarrow H_T / T$
```

Proof. Suppressing the θ dependence where it is clear,

$$\begin{aligned}
[\nabla^2_{\tilde{\theta}} \nu_\gamma(\tilde{\theta}, \pi(\theta))|_{\tilde{\theta}=\theta}]_{mn} &= \pi(\theta)'[\nabla^2_{\tilde{\theta}} J_\gamma(\tilde{\theta})|_{\tilde{\theta}=\theta}]_{mn} \\
&= \pi(\theta)'[\nabla^2_{\tilde{\theta}}(r + \gamma P(\tilde{\theta}) J_\gamma(\tilde{\theta}))|_{\tilde{\theta}=\theta}]_{mn} \\
&= \gamma\pi'(\partial_m \partial_n P J_\gamma + \partial_m P \partial_n J_\gamma + \partial_n P \partial_m J_\gamma + P \partial_m \partial_n J_\gamma) \\
&= \gamma\pi'(\partial_m \partial_n P J_\gamma + \partial_m P \partial_n J_\gamma + \partial_n P \partial_m J_\gamma) \\
&\quad + \gamma[\nabla^2_{\tilde{\theta}} \nu_\gamma(\tilde{\theta}, \pi(\theta))|_{\tilde{\theta}=\theta}]_{mn}
\end{aligned}$$

where we have used the stationarity of π' in the last line. Collecting terms proves equation 6. Using equation 1,

$$\begin{aligned}
\nabla J_\gamma &= \nabla(r + \gamma P J_\gamma) \\
&= \gamma \nabla P J_\gamma + \gamma P \nabla J_\gamma \ .
\end{aligned}$$

Equation 7 follows from unrolling the previous equation (see [7] for an equivalent proof of equation 7) □

Algorithm 1 introduces HMDP, which estimates the approximate Hessian $\tilde{\nabla}^2_\gamma \eta$ from a single sample path. It also includes the gradient algorithm of Baxter and Bartlett (with an additional factor of γ that [2] ignores). We outline the proof out its asymptotic correctness.

Theorem 6. *The HMDP sequence $\{H_T\}$ has the following property:*

$$\lim_{T \to \infty} H_T = \tilde{\nabla}^2_\gamma \eta \ .$$

Proof. The first term in equation 6, $\gamma\pi'\nabla^2 P J_\gamma$, is estimated in the algorithm by $\frac{1}{T}\sum_{t=0}^{T-1}\tilde{z}_t r(i_t)$. The proof of the asymptotic correctness of this term parallels the proof in [2] that $\frac{1}{T}\sum_{t=0}^{T-1} z_t r(i_t)$ is an asymptotically correct estimate of $\gamma\pi'\nabla P J_\gamma$ (see Algorithm 1 for definitions of $\tilde{z}_t$ and z_t).

We can write the other terms of equation 6 as

$$\gamma\pi'\partial_m P\partial_n J_\gamma = \gamma\sum_{i,,j,u}\pi(i)p_{ij}(u)\mu_u(\theta,i)\frac{\partial_m\mu_u(\theta,i)}{\mu_u(\theta,i)}\partial_n J_\gamma(\theta\, j)$$

where we have used $\partial_m P_{ij} = \sum_u p_{ij}(u)\mu_u(\theta,i)\frac{\partial_m\mu_u(\theta,i)}{\mu_u(\theta,i)}$. Let $x_0, x_1, \ldots$ be a sample trajectory corresponding to our Markov chain with x_0 chosen from the stationary distribution, which implies x_t is also a sample from the stationary distribution. Let $u_0, u_1, \ldots$ be the corresponding actions. Thus, $\gamma\frac{\partial_m\mu_{u_t}(\theta,x_t)}{\mu_{u_t}(\theta,x_t)} \times \partial_n J_\gamma(\theta, x_{t+1})$ is an unbiased estimate of $\gamma\pi'\partial_m P\partial_n J_\gamma$ for any t.

As shown in [7] (also see equation 7),

$$\partial_n J_\gamma(\theta, x_{t+1}) = \sum_{i,j,u}\sum_{\tau=t+1}^{\infty}\gamma^{\tau-t}P(X_\tau = i|X_{t+1} = x_{t+1})p_{ij}(u)$$
$$\times\mu_u(\theta,i)\frac{\partial_n\mu_u(\theta,i)}{\mu_u(\theta,i)}J_\gamma(\theta,j)$$

where X_τ is a random variable for the state of the system at time τ. For any t, $\sum_{\tau=t+1}^{\infty}\gamma^{\tau-t}\frac{\partial_n\mu_{u_\tau}(\theta,x_\tau)}{\mu_{u_\tau}(\theta,x_\tau)}J_\gamma(x_{\tau+1})$ is an unbiased estimate of $\partial_n J_\gamma(\theta, x_{t+1})$.

It follows from the Markov property that for any t,

$$\gamma\frac{\partial_m\mu_{u_t}(\theta,x_t)}{\mu_{u_t}(\theta,x_t)}\sum_{\tau=t+1}^{\infty}\gamma^{\tau-t}\frac{\partial_n\mu_{u_\tau}(\theta,x_\tau)}{\mu_{u_\tau}(\theta,x_\tau)}J_\gamma(\theta,x_{\tau+1})$$

is an unbiased sample of $\gamma\pi'\partial_m P\partial_n J_\gamma$. Since each x_t is a sample from the stationary distribution, the average

$$\frac{\gamma}{T}\sum_{t=0}^{T-1}\gamma\frac{\partial_m\mu_{u_t}(\theta,x_t)}{\mu_{u_t}(\theta,x_t)}\sum_{\tau=t+1}^{\infty}\gamma^{\tau-t}\frac{\partial_n\mu_{u_\tau}(\theta,x_\tau)}{\mu_{u_\tau}(\theta,x_\tau)}J_\gamma(\theta,x_{\tau+1})$$

almost surely converges to $\gamma\pi'\partial_m P\partial_n J_\gamma$ as $T\to\infty$. The previous expression depends on the exact values of $J_\gamma(\theta,x_t)$, which are not known. However, each $J_\gamma(\theta,x_t)$ can be estimated from the sample trajectory, and it is straightforward to show that

$$\frac{\gamma}{T}\sum_{t=0}^{T-1}\frac{\partial_m\mu_{u_t}(\theta,x_t)}{\mu_{u_t}(\theta,x_t)}\sum_{\tau=t+1}^{\infty}\gamma^{\tau-t}\frac{\partial_n\mu_{u_\tau}(\theta,x_\tau)}{\mu_{u_\tau}(\theta,x_\tau)}\sum_{\tilde{\tau}=\tau+1}^{\infty}\gamma^{\tilde{\tau}-\tau-1}r(x_{\tilde{\tau}}) \qquad (8)$$

almost surely converges to $\gamma\pi'\partial_m P\partial_n J_\gamma$ as $T\to\infty$. Using the ergodic theorem, the assumption that x_0 is chosen according to the stationary distribution can be relaxed to x_0 is an arbitrary state, since $\{X_t\}$ is asymptotically stationary (under assumption 1). Hence, equation 8 almost surely converges to $\gamma\pi'\partial_m P\partial_n J_\gamma$ for any start state x_0.

Unrolling the equations for H_T in the HMDP shows that

$$\frac{\gamma}{T}\sum_{t=0}^{T-2}\frac{\partial_m\mu_{u_t}(\theta,x_t)}{\mu_{u_t}(\theta,x_t)}\sum_{\tau=t+1}^{T-1}\gamma^{\tau-t}\frac{\partial_n\mu_{u_\tau}(\theta,x_\tau)}{\mu_{u_\tau}(\theta,x_\tau)}\sum_{\tilde{\tau}=\tau+1}^{T}\gamma^{\tilde{\tau}-\tau-1}r(x_{\tilde{\tau}})$$

is the estimate for $\gamma\pi'\partial_m P\partial_n J_\gamma$. It is straightforward to show that the error between the previous equation and equation 8 goes to 0 as $T \to \infty$. □

6 Discussion

Equation 4 suggests that the discount factor can be seen as introducing a bias-variance trade off. This equation shows that the scaled value function for every state is approximately the average reward, with an approximation error of $O(1-\gamma)$. Crudely, the variance of Monte-Carlo samples of the value function for each state is $\frac{1}{1-\gamma}$, since this is the horizon time over which rewards are added. Solving the Bellman equations will be simultaneously maximizing each biased estimate $J_\gamma(\mu,i) \approx \eta(\mu)$. We used the error in this approximation to bound the quality, measured by the average reward, of the policy obtained by solving the Bellman equations. This bound shows that good policies can be obtained if $\frac{1}{1-\gamma}$ is sufficiently larger than the mixing time of the Markov chain. This bound does not answer the question of which value of $\gamma < 1$ is large enough such that the policy that solves the Bellman equations is the optimal policy. This stems from using the worst case scenario for the error of the average reward approximation when deriving our bound, in which the stationary distribution is parallel to the second term in equation 4.

The idea of approximately maximizing the average reward using discounted rewards carries over to gradient methods. We have shown that Baxter and Bartlett's approximation to the gradient of the average reward is an exact gradient of a related discounted start state problem, and proved a similar bound on the quality of policies obtained by following this biased gradient. In [1], Bartlett and Baxter show that the bias of this approximate gradient is $O(1-\gamma)$ and the variance is $O(\frac{1}{1-\gamma})$.

Some average reward formulations exist and are essentially identical to discounted formulations with a discount factor sufficiently close to 1. Tsitsiklis and Van Roy [10] show that average reward temporal differencing (TD) (see [9]), with a discount factor sufficiently close to 1, is identical to discounted TD given appropriate learning rates and biases — converging to the same limit with the same transient behavior. An equivalent expression to the approximate gradient in the limit as $\gamma \to 1$ is given by Sutton *et al's* exact average reward gradient [7], which uses a reinforcement comparison term to obtain finite state-action values.

We have also presented an algorithm (HMDP) for computing arbitrarily accurate approximations to the Hessian from a single sample path. As suggested by [2], extensions include modifying the algorithm to compute the Hessian in a Partially Observable Markov Decision Process and in continuous state, action, and control spaces. Experimental and theoretical results are needed to better understand the approximation and estimation error of HMDP.

Acknowledgments

We thank Peter Dayan for discussions and comments on the paper. Funding is from the NSF and the Gatsby Charitable Foundation.

References

1. P. Bartlett and J. Baxter. Estimation and approximation bounds for gradient-based reinforcement learning. Technical report, Australian National University, 2000.
2. J. Baxter and P. Bartlett. Direct gradient-based reinforcement learning. Technical report, Australian National University, Research School of Information Sciences and Engineering, July 1999.
3. J. Baxter and P. Bartlett. Algorithms for infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 2001. (forthcoming).
4. D. P. Bertsekas. *Dynamic Programming and Optimal Control, Volumes 1 and 2.* Athena Scientific, 1995.
5. P. Marbach and J. Tsitsiklis. Simulation-based optimization of markov reward processes. Technical report, Massachusetts Institute of Technology, 1998.
6. S. Singh, T. Jaakkola, and M. I. Jordan. Learning without state-estimation in partially observable markovian decision processes. *Proc. 11th International Conference on Machine Learning*, 1994.
7. R. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. *Neural Information Processing Systems*, 13, 2000.
8. R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction.* MIT Press, 1998.
9. John N. Tsitsiklis and Benjamin Van Roy. Average cost temporal-difference learning. *Automatica*, 35:319–349, 1999.
10. John N. Tsitsiklis and Benjamin Van Roy. On average versus discounted reward temporal-difference learning. *Machine Learning*, 2001. (forthcoming).