1 Proofs

1.1 Theorem 1 and Table 2

We combine the proofs of the Theorem 1 and the variance analysis in Table 2 due to substantial overlap. In particular, the results of Table 2 entail Theorem 2. First, we consider the advantage estimator derived from $v^\theta(O_t)$:

$$\begin{aligned}
\mathrm{Var}(G_t - v^\theta(O_t)) &\overset{(a)}{=} \mathbb{E}[\mathrm{Var}(G_t - v^\theta(O_t)|O_t)] + \mathrm{Var}(\mathbb{E}[G_t - v^\theta(O_t)|O_t]) \\
&\overset{(b)}{=} \mathbb{E}[\mathrm{Var}(G_t|O_t)] + \mathrm{Var}(\mathbb{E}[G_t - v^\theta(O_t)|O_t]) \\
&\overset{(c)}{=} \mathbb{E}[\mathrm{Var}(G_t|O_t)] + \mathrm{Var}(\mathbb{E}[G_t|O_t] - \mathbb{E}[v^\theta(O_t)|O_t]) \\
&\overset{(d)}{=} \mathbb{E}[\mathrm{Var}(G_t|O_t)] + \mathrm{Var}(\underbrace{\mathbb{E}[G_t|O_t] - \mathbb{E}[\mathbb{E}[G_t|O_t]|O_t]}_{0}) \\
&= \mathbb{E}[\mathrm{Var}(G_t|O_t)].
\end{aligned}$$

where **(a)** follows from the law of total variance, **(b)** follows because the conditional variance of $v^\theta(O_t)$ given O_t is zero, **(c)** follows from the linearity of expectation, and **(d)** follows from the definition of $v^\theta(O_t)$. An identical series of transformations can be applied for $v^\theta(H_t)$ and $v^\theta(S_t)$, except by applying the law of total variance using H_t and S_t respectively, instead of O_t. $\mathrm{Var}(G_t - u_t^\theta(H))$ cannot be substantially simplified, however, the value listed in the table may be derived as follows:

$$\mathrm{Var}(G_t - u_t^\theta(H)) \overset{(a)}{=} \mathrm{Var}(\mathbb{E}[G_t|H] - \mathbb{E}[v^\theta(S_t)|H]) \overset{(b)}{=} \mathrm{Var}(\mathbb{E}[G_t - v^\theta(S_t)|H]),$$

where **(a)** follows because $G_t = \mathbb{E}[G_t|H]$ and from the definition of u_t^θ, and **(b)** follows by linearity of expectation.

Next we consider the difference between each baseline, starting with $v^\theta(O_t)$ and $v^\theta(H_t)$:

$$\begin{aligned}
\mathrm{Var}(G_t - v^\theta(O_t)) - \mathrm{Var}(G_t - v^\theta(H_t)) &\overset{(a)}{=} \mathbb{E}[\mathrm{Var}(G_t|O_t)] - \mathbb{E}[\mathrm{Var}(G_t|H_t)] \\
&\overset{(b)}{=} \mathbb{E}[\mathrm{Var}(\mathbb{E}[G_t|H_t, O_t]|O_t)] + \mathbb{E}[\mathrm{Var}(G_t|H_t, O_t)] - \mathbb{E}[\mathrm{Var}(G_t|H_t)] \\
&\overset{(c)}{=} \mathbb{E}[\mathrm{Var}(\mathbb{E}[G_t|H_t]|O_t)] + \underbrace{\mathbb{E}[\mathrm{Var}(G_t|H_t)] - \mathbb{E}[\mathrm{Var}(G_t|H_t)]}_{0} \\
&\overset{(d)}{=} \mathbb{E}[\mathrm{Var}(v^\theta(H_t)|O_t)],
\end{aligned}$$

where **(a)** follows from the previously derived expressions, **(b)** follows from law of total conditional variance, **(c)** follows because O_t is a component of H_t, so conditioning on O_t and H_t together is equivalent to conditioning on H_t, and **(d)** follows from the definition of $v^\theta(H_t)$. The same series of transformations applies without issue when comparing $v^\theta(H_t)$ and $v^\theta(S_t)$:

$$\begin{aligned}
\mathrm{Var}(G_t - v^\theta(H_t)) - \mathrm{Var}(G_t - v^\theta(S_t)) &\overset{(a)}{=} \mathbb{E}[\mathrm{Var}(G_t|H_t)] - \mathbb{E}[\mathrm{Var}(G_t|S_t)] \\
&\overset{(b)}{=} \mathbb{E}[\mathrm{Var}(\mathbb{E}[G_t|H_t, S_t]|H_t)] + \mathbb{E}[\mathrm{Var}(G_t|H_t, S_t)] - \mathbb{E}[\mathrm{Var}(G_t|S_t)] \\
&\overset{(c)}{=} \mathbb{E}[\mathrm{Var}(\mathbb{E}[G_t|S_t]|H_t)] + \underbrace{\mathbb{E}[\mathrm{Var}(G_t|S_t)] - \mathbb{E}[\mathrm{Var}(G_t|S_t)]}_{0} \\
&\overset{(d)}{=} \mathbb{E}[\mathrm{Var}(v^\theta(S_t)|H_t)].
\end{aligned}$$

A subtle difference in reasoning occurs in **(c)**, being that we are able to condition on just S_t instead of H_t and S_t together due to the Markov property, i.e., because G_t is conditionally independent of H_t given S_t. Lastly, we compare $u_t^\theta(H)$ and $v^\theta(S_t)$:

$$\begin{aligned}
\mathrm{Var}(G_t - v^\theta(S_t)) - \mathrm{Var}(G_t - u_t^\theta(H)) &\overset{\text{(a)}}{=} \mathrm{Var}(G_t - v^\theta(S_t)) - \mathrm{Var}(\mathbb{E}[G_t - v^\theta(S_t)|H]) \\
&\overset{\text{(b)}}{=} \mathbb{E}[\mathrm{Var}(G_t - v^\theta(S_t)|H)] + \mathrm{Var}(\mathbb{E}[G_t - v^\theta(S_t)|H]) - \mathrm{Var}(\mathbb{E}[G_t - v^\theta(S_t)|H]) \\
&\overset{\text{(c)}}{=} \mathbb{E}[\mathrm{Var}(G_t - v^\theta(S_t)|H)] \\
&\overset{\text{(d)}}{=} \mathbb{E}[\mathrm{Var}(v^\theta(S_t)|H)],
\end{aligned}$$

where **(a)** substitutes the previously derived expression for $\mathrm{Var}(G_t - u_t^\theta(H))$, **(b)** follows from the law of total variance, **(c)** results from terms canceling, and **(d)** follows because the conditional variance of G_t given H is zero and because for any X, $\mathrm{Var}(-X) = \mathrm{Var}(X)$. Thus, the proof of the results listed in Table 2 is complete. Theorem 1 follows from the above results because as variance and conditional variance are always non-negative.

1.2 Theorem 2

By linearity of expectation, we can write the above expression as:

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

where we define ψ as the compatible features, $\psi(X_t, A_t) \coloneqq \partial \ln \pi_\theta(X_t, U_t)/\partial\theta$. The first component is $\nabla J(\theta)$, and we will prove the latter component is zero. Therefore, the entire expression is equal to $\nabla J(\theta)$, proving the theorem. For all t:

$$\begin{aligned}
\mathbb{E}\Big[u_t^\theta(H)\psi(O_t, A_t)\Big] &\overset{\text{(a)}}{=} \mathbb{E}\Big[\mathbb{E}[v^\theta(S_t)|H]\psi(O_t, A_t)\Big] \\
&\overset{\text{(b)}}{=} \mathbb{E}\Big[\mathbb{E}[v^\theta(S_t)\psi(O_t, A_t)|H]\Big] \\
&\overset{\text{(c)}}{=} \mathbb{E}\Big[v^\theta(S_t)\psi(O_t, A_t)\Big],
\end{aligned}$$

where **(a)** follows from the definition of u_t^θ and **(c)** holds due to the tower property. The critical step is **(b)**. Because O_t and A_t are components of H, $\psi(O_t, A_t)$ may be moved inside the inner expectation. The remainder of the proof is essentially equivalent to the standard proof that the $v^\theta(S_t)$ baseline is unbiased, which can be found in any standard RL text, but we will complete the proof here in our notation.

Continuing, we again use the tower property, essentially reversing the steps above, except this time conditioning on S_t instead of H:

$$\mathbb{E}\Big[v^\theta(S_t)\psi(O_t, A_t)\Big] = \mathbb{E}\Big[\mathbb{E}[v^\theta(S_t)\psi(O_t, A_t)|S_t]\Big] = \mathbb{E}\Big[v^\theta(S_t)\underbrace{\mathbb{E}[\psi(O_t, A_t)|S_t]}_{0}\Big].$$

Again, we can move $v^\theta(S_t)$ out of the inner expectation because it is constant given S_t. The remaining inner expression is zero because:

$$
\begin{aligned}
&\mathbb{E}[\psi(O_t, A_t)|S_t]\\
&\overset{\text{(a)}}{=} \sum_{o\in\mathcal{O}} \Pr(O_t = o|S_t) \sum_{a\in\mathcal{A}} \underbrace{\Pr(A_t = a|O_t = o, S_t)}_{\pi_\theta(o,a)} \psi(o, a)\\
&\overset{\text{(b)}}{=} \sum_{o\in\mathcal{O}} \Pr(O_t = o|S_t) \sum_{a\in\mathcal{A}} \pi_\theta(o, a) \frac{\partial \ln \pi_\theta(o, a)}{\partial\theta}\\
&\overset{\text{(c)}}{=} \sum_{o\in\mathcal{O}} \Pr(O_t = o|S_t) \sum_{a\in\mathcal{A}} \pi_\theta(o, a) \frac{1}{\pi_\theta(o, a)} \frac{\partial \pi_\theta(o, a)}{\partial\theta}\\
&\overset{\text{(d)}}{=} \sum_{o\in\mathcal{O}} \Pr(O_t = o|S_t) \sum_{a\in\mathcal{A}} \frac{\partial \pi_\theta(o, a)}{\partial\theta}\\
&\overset{\text{(e)}}{=} \sum_{o\in\mathcal{O}} \Pr(O_t = o|S_t) \frac{\partial}{\partial\theta} \sum_{a\in\mathcal{A}} \pi_\theta(o, a)\\
&\overset{\text{(f)}}{=} \sum_{o\in\mathcal{O}} \Pr(O_t = o|S_t) \frac{\partial}{\partial\theta} \underbrace{\sum_{a\in\mathcal{A}} \Pr(A_t = a|O_t = o, \theta)}_{1}\\
&\overset{\text{(g)}}{=} 0,
\end{aligned}
$$

where **(a)** follows from the law of total expectation, **(b)** follows from the definition of π_θ and ψ, **(c)** follows from a well-known property of the derivative of a logarithmic function, **(d)** follows trivially by canceling terms, **(e)** follows from the linearity of differentiation, **(f)** follows from the definition of π_θ, and **(g)** follows because the sum over the probability of all possible actions is 1, and the derivative of a constant is 0.

1.3 Theorem 3

Let $\psi(O_t, A_t) \coloneqq \partial \ln \pi_\theta(O_t, A_t)/\partial\theta$. Then:

$$
\begin{aligned}
\text{Var}\big((G_t - u_t^\theta(H))\psi(O_t, A_t)\big) &\overset{\text{(a)}}{=} \text{Var}\big(\big(\mathbb{E}[G_t|H] - \mathbb{E}[v^\theta(S_t)|H]\big)\psi(O_t, A_t)\big)\\
&\overset{\text{(b)}}{=} \text{Var}\big(\mathbb{E}[G_t - v^\theta(S_t)|H]\psi(O_t, A_t)\big)\\
&\overset{\text{(c)}}{=} \text{Var}\big(\mathbb{E}[(G_t - v^\theta(S_t))\psi(O_t, A_t)|H]\big)\\
&\overset{\text{(d)}}{\leq} \text{Var}\big(\mathbb{E}[(G_t - v^\theta(S_t))\psi(O_t, A_t)|H]\big) + \mathbb{E}\big[\text{Var}\big((G_t - v^\theta(S_t))\psi(O_t, A_t)|H\big)\big]\\
&\overset{\text{(e)}}{=} \text{Var}\big((G_t - v^\theta(S_t))\psi(O_t, A_t)\big),
\end{aligned}
$$

where **(a)** follows because $G_t = \mathbb{E}[G_t|H]$ always and by the definition of u_t^θ, **(b)** is due to the linearity of expectation, **(c)** holds because O_t and A_t are components of H, **(d)** holds because variance is always positive, and **(e)** holds by the law of total variance. This completes the proof.

1.4 Theorem 4

The layout of the proof is as follows: We first show that, for all $h \in \mathcal{H}$, $\Pr(H = h) = \Pr(\tilde{H} = h)$. Next, we show that $\nabla J(\theta) = \nabla \tilde{J}(\theta)$. Finally, we show that the given expectation is equal to $\nabla \tilde{J}(\theta)$, and therefore, by substitution, $\nabla J(\theta)$.

First, recall the two conditions are that for all $o \in \mathcal{O}$, partial histories $h = (o_0, a_0, r_0, \dots, o_t)$, $h' = (o_0, a_0, r_0, \dots, o_{t+1})$, and actions $a \in \mathcal{A}$:

$$\Pr(O_0 = o) = \Pr(\tilde{O}_0 = o) \tag{1}$$

$$\Pr(H_{t+1} = h' | H_t = h, A_t = a) = \Pr(\tilde{H}_{t+1} = h' | \tilde{H}_t = h, \tilde{A}_t = a) \tag{2}$$

Additionally, we define the following helper functions for notational convenience: For any history, $h = (o_0, a_0, r_0, \dots, o_i)$, such that $i \geq t$, we define $o_t(h) \coloneqq o_t$. For any history such that $i > t$, $a_t(h) \coloneqq a_t$. Finally, for any complete history, we define $g_t(h) \coloneqq \sum_{i=t}^{|h|} r_i$, where $|h|$ is the length of h in terms of timesteps.

1.4.1 Probability of Histories

We first prove by induction that the two conditions given are sufficient to prove that for all histories h, $\Pr(H = h) = \Pr(\tilde{H} = h)$. We show that for all partial histories h and h', if $\Pr(H_t = h) = \Pr(\tilde{H}_t = h)$, then $\Pr(H_{t+1} = h') = \Pr(\tilde{H}_{t+1} = h')$. Because $H_0 = O_0$, the initial condition, $\Pr(H_0 = h) = \Pr(\tilde{H}_0 = h)$, holds trivially by (1). Then:

$$\begin{aligned}
\Pr(H_{t+1} = h') &\overset{(a)}{=} \sum_h \Pr(H_t = h) \Pr(H_{t+1} = h' | H_t = h) \\
&\overset{(b)}{=} \sum_h \Pr(\tilde{H}_t = h) \Pr(H_{t+1} = h' | H_t = h) \\
&\overset{(c)}{=} \sum_h \Pr(\tilde{H}_t = h) \sum_{a \in \mathcal{A}} \Pr(A_t = a | H_t = h) \Pr(H_{t+1} = h' | H_t = h, A_t = a) \\
&\overset{(d)}{=} \sum_h \Pr(\tilde{H}_t = h) \sum_{a \in \mathcal{A}} \pi_\theta(o_t(h), a) \Pr(H_{t+1} = h' | H_t = h, A_t = a) \\
&\overset{(e)}{=} \sum_h \Pr(\tilde{H}_t = h) \sum_{a \in \mathcal{A}} \pi_\theta(o_t(h), a) \Pr(\tilde{H}_{t+1} = h' | \tilde{H}_t = h, \tilde{A}_t = a) \\
&\overset{(f)}{=} \sum_h \Pr(\tilde{H}_t = h) \sum_{a \in \mathcal{A}} \Pr(\tilde{A}_t = a | \tilde{H}_t = h) \Pr(\tilde{H}_{t+1} = h' | \tilde{H}_t = h, \tilde{A}_t = a) \\
&\overset{(g)}{=} Pr(\tilde{H}_{t+1} = h'),
\end{aligned}$$

where **(a)** follows from the law of total probability, **(b)** follows from the inductive assumption, **(c)** again follows from the law of total probability, **(d)** follows from the definition of π_θ, **(e)** follows from (2), **(f)** follows again from the definition of π_θ, and **(g)** follows from the law of total probability again.

1.4.2 Equality of Policy Gradients

First of all, we note that for any complete history h:

$$\mathbb{E}[G_0 | H = h] = g_0(h) = \mathbb{E}[\tilde{G}_0 | \tilde{H} = h], \tag{3}$$

because for both M and $\tilde{M}$, G_0 is simply the deterministic sum of rewards in h. Therefore:

$$
\begin{aligned}
J(\theta) &= \mathbb{E}[G_0] \\
&\overset{(a)}{=} \sum_h \Pr(H = h)\mathbb{E}[G_0 | H = h] \\
&\overset{(b)}{=} \sum_h \Pr(\tilde{H} = h)\mathbb{E}[G_0 | H = h] \\
&\overset{(c)}{=} \sum_h \Pr(\tilde{H} = h)\mathbb{E}[\tilde{G}_0 | \tilde{H} = h] \\
&\overset{(d)}{=} \mathbb{E}[\tilde{G}_0] \\
&= \tilde{J}(\theta),
\end{aligned}
$$

where **(a)** follows from the law of total expectation, **(b)** was proved in the previous subsection, **(c)** follows from (3), and **(d)** again follows from the law of total expectation. Therefore, $\nabla J(\theta) = \nabla \tilde{J}(\theta)$.

1.4.3 Completion of Proof

We have:

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=0}^{\infty}(G_t - \tilde{u}_t^\theta(H))\frac{\partial \ln \pi_\theta(O_t, A_t)}{\partial \theta}\right] &\overset{(a)}{=} \sum_h \Pr(H = h)\mathbb{E}\left[\sum_{t=0}^{\infty}(G_t - \tilde{u}_t^\theta(H))\frac{\partial \ln \pi_\theta(O_t, A_t)}{\partial \theta}\middle| H = h\right] \\
&\overset{(b)}{=} \sum_h \Pr(H = h)\sum_{t=0}^{\infty}(g_t(h) - \tilde{u}_t^\theta(h))\frac{\partial \ln \pi_\theta(o_t(h), a_t(h))}{\partial \theta} \\
&\overset{(c)}{=} \sum_h \Pr(H = h)\mathbb{E}\left[\sum_{t=0}^{\infty}(\tilde{G}_t - \tilde{u}_t^\theta(H))\frac{\partial \ln \pi_\theta(\tilde{O}_t, \tilde{A}_t)}{\partial \theta}\middle| \tilde{H} = h\right] \\
&\overset{(d)}{=} \sum_h \Pr(\tilde{H} = h)\mathbb{E}\left[\sum_{t=0}^{\infty}(\tilde{G}_t - \tilde{u}_t^\theta(H))\frac{\partial \ln \pi_\theta(\tilde{O}_t, \tilde{A}_t)}{\partial \theta}\middle| \tilde{H} = h\right] \\
&\overset{(e)}{=} \mathbb{E}\left[\sum_{t=0}^{\infty}(\tilde{G}_t - \tilde{u}_t^\theta(H))\frac{\partial \ln \pi_\theta(\tilde{O}_t, \tilde{A}_t)}{\partial \theta}\right] \\
&\overset{(f)}{=} \nabla \tilde{J}(\theta) \\
&\overset{(g)}{=} \nabla J(\theta),
\end{aligned}
$$

where **(a)** follows from the law of total expectation, **(b)** and **(c)** follows from the definitions of g_t, o_t, and a_t, **(d)** follows from the proof in Section 1.4.1, **(e)** follows from the law of total expectation again, **(f)** follows from Theorem 2, and **(g)** was shown in Section 1.4.2. Thus, the proof is completed.

1.5 Theorem 5

Following the proof of Theorem 4, it is straightforward but tedious to prove that $\mathrm{Var}(G_t - \tilde{u}_t^\theta(H)) = \mathrm{Var}(\tilde{G}_t - \tilde{u}_t^\theta(\tilde{H}))$ and $\mathrm{Var}(G_t - v^\theta(H_t)) = \mathrm{Var}(\tilde{G}_t - \tilde{v}^\theta(\tilde{H}_t))$, where $\tilde{v}^\theta(H_t) \coloneqq \mathbb{E}[\tilde{G}_t|\tilde{H}_t]$. We know from Theorem 1 that $\mathrm{Var}(\tilde{G}_t - \tilde{u}_t^\theta(\tilde{H})) \leq \mathrm{Var}(\tilde{G}_t - \tilde{v}^\theta(\tilde{H}_t))$. Therefore, by substitution, we have immediately that $\mathrm{Var}(G_t - \tilde{u}_t^\theta(H)) \leq \mathrm{Var}(G_t - v^\theta(H_t))$.

However, we must still prove the prior statements. First we prove $\mathrm{Var}(G_t - \tilde{u}_t^\theta(H)) = \mathrm{Var}(\tilde{G}_t - \tilde{u}_t^\theta(\tilde{H}))$:

$$\begin{aligned}
\mathrm{Var}(G_t - \tilde{u}_t^\theta(H)) &\overset{(a)}{=} \mathbb{E}[(G_t - \tilde{u}_t^\theta(H) - \mathbb{E}[G_t - \tilde{u}_t^\theta(H)])^2] \\
&\overset{(b)}{=} \sum_h \Pr(H = h)\mathbb{E}[(G_t - \tilde{u}_t^\theta(H) - \mathbb{E}[G_t - \tilde{u}_t^\theta(H)])^2|H = h] \\
&\overset{(c)}{=} \sum_h \Pr(H = h)(g_t(h) - \tilde{u}_t^\theta(h) - \mathbb{E}[G_t - \tilde{u}_t^\theta(H)])^2 \\
&\overset{(d)}{=} \sum_h \Pr(\tilde{H} = h)(g_t(h) - \tilde{u}_t^\theta(h) - \mathbb{E}[G_t - \tilde{u}_t^\theta(\tilde{H})])^2 \\
&\overset{(e)}{=} \sum_h \Pr(\tilde{H} = h)(g_t(h) - \tilde{u}_t^\theta(h) - \mathbb{E}[\tilde{G}_t - \tilde{u}_t^\theta(\tilde{H})])^2 \\
&\overset{(f)}{=} \sum_h \Pr(\tilde{H} = h)\mathbb{E}[(\tilde{G}_t - \tilde{u}_t^\theta(\tilde{H}) - \mathbb{E}[\tilde{G}_t - \tilde{u}_t^\theta(\tilde{H})])^2|\tilde{H} = h] \\
&\overset{(g)}{=} \mathrm{Var}(\tilde{G}_t - \tilde{u}_t^\theta(\tilde{H})),
\end{aligned}$$

where **(a)** follows from the definition of variance, **(b)** follows from the law of total expectation, **(c)** follows from the definition of $g_t(h)$, **(d)** follows from the proof found in Section 1.4.1, **(f)** follows from the definition of $g_t(h)$, **(g)** follows again from the definition of variance. **(e)** follows from a similar series of transformations:

$$\begin{aligned}
\mathbb{E}[G_t - \tilde{u}_t^\theta(H)] &= \sum_h \Pr(H = h)\mathbb{E}[G_t - \tilde{u}_t^\theta(H)|H = h] \\
&= \sum_h \Pr(\tilde{H} = h)(g_t(h) - \tilde{u}_t^\theta(h)) \\
&= \sum_h \Pr(\tilde{H} = h)\mathbb{E}[\tilde{G}_t - \tilde{u}_t^\theta(\tilde{H})|\tilde{H} = h] \\
&= \mathbb{E}[\tilde{G}_t - \tilde{u}_t^\theta(\tilde{H})].
\end{aligned}$$

Next we prove $\mathrm{Var}(G_t - v^\theta(H_t)) = \mathrm{Var}(\tilde{G}_t - \tilde{v}^\theta(\tilde{H}_t))$:

$$\begin{aligned}
\mathrm{Var}(G_t - v^\theta(H_t)) &\overset{(a)}{=} \mathbb{E}[(G_t - v^\theta(H_t) - \mathbb{E}[G_t - v^\theta(H_t)])^2] \\
&= \mathbb{E}[(G_t - v^\theta(H_t))^2] \\
&= \sum_h \Pr(H = h)\mathbb{E}[G_t - v^\theta(H_t))^2|H = h] \\
&= \sum_h \Pr\big(H = h\big)\big(g_t(h) - v^\theta(h_t(h))\big)^2 \\
&\overset{(b)}{=} \sum_h \Pr\big(H = h\big)\big(g_t(h) - \tilde{v}^\theta(h_t(h))\big)^2 \\
&= \sum_h \Pr(\tilde{H} = h)\mathbb{E}[\tilde{G}_t - \tilde{v}^\theta(\tilde{H}_t))^2|\tilde{H} = h] \\
&= \mathbb{E}[(\tilde{G}_t - \tilde{v}^\theta(\tilde{H}_t))^2] \\
&\overset{(c)}{=} \mathbb{E}[(\tilde{G}_t - \tilde{v}^\theta(\tilde{H}_t) - \mathbb{E}[\tilde{G}_t - \tilde{v}^\theta(\tilde{H}_t)])^2] \\
&= \mathrm{Var}(\tilde{G}_t - v^\theta(\tilde{H}_t)).
\end{aligned}$$

where **(a)** and **(c)** follow from:

$$\mathbb{E}[G_t - v^\theta(H_t)] = \mathbb{E}[\mathbb{E}[G_t - v^\theta(H_t)|H_t]] = \mathbb{E}[\mathbb{E}[G_t - \mathbb{E}[G_t|H_t]|H_t]] = \mathbb{E}[\mathbb{E}[G_t|H_t] - \mathbb{E}[G_t|H_t]] = 0,$$

and **(b)** follows because

$$\begin{aligned} v^\theta(h_t) = \mathbb{E}[G_t|H_t = h_t] &= \sum_h \Pr(H = h|H_t = h_t)\mathbb{E}[G_t|H = h, H_t = h_t] \\ &= \sum_h \Pr(\tilde{H} = h|\tilde{H}_t = h_t) g_t(h) \\ &= \sum_h \Pr(\tilde{H} = h|\tilde{H}_t = h_t)\mathbb{E}[\tilde{G}_t|\tilde{H} = h, \tilde{H}_t = h_t] \\ &= \tilde{v}^\theta(h_t). \end{aligned}$$

This completes the proof. The supplemental material continues on the next page.

1.6 Theorem 6

We assume that $\hat{v}$ is a tabular representation with parameters ω, such that for all $o \in \mathcal{O}$ and $z \in \mathcal{Z}$, $\hat{v}(o,z) = \omega_{o,z}$. Consider the following loss function:

$$\mathcal{L}(\omega) = \sum_{t=0}^{\infty} \sum_{z \in \mathcal{Z}} \sum_{o \in \mathcal{O}} \Pr(O_t = o, Z_t = z)\big(v^\theta(o,z) - \hat{v}(o,z)\big)^2. \tag{4}$$

This loss function is convex with a global minimum where $\forall o \in \mathcal{O}, z \in \mathcal{Z} : \omega_{o,z} = v^\theta(o,z)$, and therefore a good candidate for stochastic approximation. The gradient of $\mathcal{L}$ is:

$$\nabla\mathcal{L}(\omega) = 2\sum_{t=0}^{\infty} \sum_{z \in \mathcal{Z}} \sum_{o \in \mathcal{O}} \Pr(O_t = o, Z_t = z)\big(v^\theta(o,z) - \hat{v}(o,z)\big)\frac{\partial \hat{v}(o,z)}{\partial \omega}. \tag{5}$$

The proposed update direction could be rewritten as:

$$\sum_{t=0}^{\infty} \sum_{z \in \mathcal{Z}} \Pr(Z_t = z|H)(G_t - \hat{v}(O_t,z))\frac{\partial \hat{v}(O_t,z)}{\partial \omega}.$$

In expectation, this update is:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{\infty} \sum_{z \in \mathcal{Z}} \Pr(Z_t = z|H)(G_t - \hat{v}(O_t,z))\frac{\partial \hat{v}(O_t,z)}{\partial \omega}\right] \\
=&\sum_{t=0}^{\infty} \sum_{z \in \mathcal{Z}} \mathbb{E}\left[\Pr(Z_t = z|H)(G_t - \hat{v}(O_t,z))\frac{\partial \hat{v}(O_t,z)}{\partial \omega}\right] \\
=&\sum_{t=0}^{\infty} \sum_{z \in \mathcal{Z}} \sum_{o \in \mathcal{O}} \Pr(O_t = o)\mathbb{E}\left[\Pr(Z_t = z|H)(G_t - \hat{v}(o,z))\frac{\partial \hat{v}(o,z)}{\partial \omega}\middle| O_t = o\right] \\
=&\sum_{t=0}^{\infty} \sum_{z \in \mathcal{Z}} \sum_{o \in \mathcal{O}} \Pr(O_t = o)\sum_{h} \Pr(H = h|O_t = o)\Pr(Z_t = z|H = h)(g_t(h) - \hat{v}(o,z))\frac{\partial \hat{v}(o,z)}{\partial \omega} \\
=&\sum_{t=0}^{\infty} \sum_{z \in \mathcal{Z}} \sum_{o \in \mathcal{O}} \Pr(O_t = o)\sum_{h} \Pr(H = h|O_t = o)\Pr(Z_t = z|H = h, O_t = o)(g_t(h) - \hat{v}(o,z))\frac{\partial \hat{v}(o,z)}{\partial \omega} \\
=&\sum_{t=0}^{\infty} \sum_{z \in \mathcal{Z}} \sum_{o \in \mathcal{O}} \Pr(O_t = o)\sum_{h} \Pr(H = h|O_t = o)\Pr(Z_t = z|H = h, O_t = o) \\
&\times \mathbb{E}\left[(G_t - \hat{v}(o,z))\frac{\partial \hat{v}(o,z)}{\partial \omega}\middle| H = h, O_t = o, Z_t = z\right] \\
=&\sum_{t=0}^{\infty} \sum_{z \in \mathcal{Z}} \sum_{o \in \mathcal{O}} \Pr(O_t = o)\Pr(Z_t = z|O_t = o)\mathbb{E}\left[(G_t - \hat{v}(o,z))\frac{\partial \hat{v}(o,z)}{\partial \omega}\middle| O_t = o, Z_t = z\right]
\end{aligned}$$

Continuing:

$$
\begin{aligned}
&\sum_{t=0}^{\infty}\sum_{z\in\mathcal{Z}}\sum_{o\in\mathcal{O}} \Pr(O_t=o)\Pr(Z_t=z|O_t=o)\mathbb{E}\left[(G_t-\hat{v}(o,z))\frac{\partial\hat{v}(o,z)}{\partial\omega}\middle|O_t=o,Z_t=z\right]\\
=&\sum_{t=0}^{\infty}\sum_{z\in\mathcal{Z}}\sum_{o\in\mathcal{O}} \Pr(O_t=o,Z_t=z)\mathbb{E}\left[(G_t-\hat{v}(o,z))\frac{\partial\hat{v}(o,z)}{\partial\omega}\middle|O_t=o,Z_t=z\right]\\
=&\sum_{t=0}^{\infty}\sum_{z\in\mathcal{Z}}\sum_{o\in\mathcal{O}} \Pr(O_t=o,Z_t=z)(v^\theta(o,z)-\hat{v}(o,z))\frac{\partial\hat{v}(o,z)}{\partial\omega}\\
=&\frac{1}{2}\nabla\mathcal{L}(\omega).
\end{aligned}
$$

Therefore, we can see that the proposed update is an unbiased estimator for the gradient of the given loss times a positive constant. Therefore, given an appropriate step-size schedule for α_i, we are given the standard guarantees of gradient descent [1]. Further, because $\mathcal{L}$ is convex with respect to its parameters, the method will converge to the global optimum almost surely, i.e.:

$$\forall o\in\mathcal{O}, z\in\mathcal{Z}:\Pr\left(\lim_{i\to\infty}\hat{v}_i(o,z)=v^\theta(o,z)\right)=1.$$

Because $\hat{v}(o,z)=v_i^\theta(o,z) \implies \hat{u}_{t,i}(h)=u_t^\theta(h)$ by the construction of $\hat{u}_{t,i}$ and the definition of u_t^θ, we conclude:

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

This completes the proof. The supplemental material continues on the next page.

1.7 Theorem 7

We again consider the loss function given by (4). We can rewrite the more efficient update as:

$$\sum_{t=0}^{\infty}(G_t - \hat{v}(O_t, \hat{Z}_t))\frac{\partial \hat{v}(O_t, \hat{Z}_t)}{\partial \omega}.$$

In expectation, we have:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{\infty}(G_t - \hat{v}(O_t, \hat{Z}_t))\frac{\partial \hat{v}(O_t, \hat{Z}_t)}{\partial \omega}\right] \\
=&\sum_{t=0}^{\infty}\mathbb{E}\left[(G_t - \hat{v}(O_t, \hat{Z}_t))\frac{\partial \hat{v}(O_t, \hat{Z}_t)}{\partial \omega}\right] \\
=&\sum_{t=0}^{\infty}\sum_{h}\Pr(H = h)\sum_{z\in\mathcal{Z}}\Pr(\hat{Z} = z|H = h)\mathbb{E}\left[(G_t - \hat{v}(O_t, \hat{Z}_t))\frac{\partial \hat{v}(O_t, \hat{Z}_t)}{\partial \omega}\middle| H = h, \hat{Z} = z\right] \\
=&\sum_{t=0}^{\infty}\sum_{h}\Pr(H = h)\sum_{z\in\mathcal{Z}}\Pr(\hat{Z} = z|H = h)\left((g_t(h) - \hat{v}(o_t(h), z))\frac{\partial \hat{v}(o_t(h), z)}{\partial \omega}\right) \\
=&\sum_{t=0}^{\infty}\sum_{h}\Pr(H = h)\sum_{z\in\mathcal{Z}}\Pr(Z = z|H = h)\left((g_t(h) - \hat{v}(o_t(h), z))\frac{\partial \hat{v}(o_t(h), z)}{\partial \omega}\right) \\
=&\sum_{t=0}^{\infty}\mathbb{E}\left[(G_t - \hat{v}(O_t, Z_t))\frac{\partial \hat{v}(O_t, Z_t)}{\partial \omega}\right] \\
=&\sum_{t=0}^{\infty}\sum_{z\in\mathcal{Z}}\sum_{o\in\mathcal{O}}\Pr(O_t = o, Z_t = z)\mathbb{E}\left[(G_t - \hat{v}(o, z))\frac{\partial \hat{v}(o, z)}{\partial \omega}\middle| O_t = o, Z_t = z\right] \\
=&\sum_{t=0}^{\infty}\sum_{z\in\mathcal{Z}}\sum_{o\in\mathcal{O}}\Pr(O_t = o, Z_t = z)(v^{\theta}(o, z) - \hat{v}(o, z))\frac{\partial \hat{v}(o, z)}{\partial \omega} \\
=&\frac{1}{2}\nabla\mathcal{L}(\omega).
\end{aligned}$$

Thus, by the same argument as for Theorem 6:

$$\forall o \in \mathcal{O}, z \in \mathcal{Z} : \Pr\left(\lim_{i\to\infty}\hat{v}_i(o, z) = v^{\theta}(o, z)\right) = 1.$$

However, this time we must also consider $\hat{u}$. We again consider a tabular representation, $\hat{u}(o, \phi) = \psi_{o,\phi}$, where ψ is a parameter vector. Consider the loss:

$$\mathcal{L}(\psi) = \sum_{t=0}^{\infty}\sum_{h}\Pr(H = h)\big(u_t^{\theta}(h) - \hat{u}_t(o_t(h), \phi(h))\big)^2 \tag{6}$$

$$\nabla\mathcal{L}(\psi) = 2\sum_{t=0}^{\infty}\sum_{h}\Pr(H = h)\big(u_t^{\theta}(h) - \hat{u}_t(o_t(h), \phi_t(h))\big)\frac{\partial \hat{u}_t(o_t(h), \phi_t(h))}{\partial \psi}. \tag{7}$$

Again, notice this loss is convex, with a global optimum at $\hat{u}_t(o_t(h), \phi_t(h)) = u_t^{\theta}(h)$. In the limit as $i \to \infty$, our update is:

$$\sum_{t=0}^{\infty}\Big(v^{\theta}(O_t,\tilde{Z}_t)-\hat{u}_t(O_t,\phi_t)\Big)\frac{\partial\hat{u}_t(O_t,\phi_t)}{\partial\psi}.$$

In expectation, this is:

$$\begin{aligned}
&\sum_{t=0}^{\infty}\mathbb{E}\left[\Big(v^{\theta}(O_t,\tilde{Z}_t)-\hat{u}_t(O_t,\phi_t)\Big)\frac{\partial\hat{u}_t(O_t,\phi_t)}{\partial\psi}\right]\\
=&\sum_{t=0}^{\infty}\sum_{h}\Pr(H=h)\sum_{z}\Pr(\tilde{Z}=z|H=h)\Big(v^{\theta}(o_t(h),z)-\hat{u}_t(o_t(h),\phi_t(h))\Big)\frac{\partial\hat{u}_t(o_t(h),\phi_t(h))}{\partial\psi}\\
&\sum_{t=0}^{\infty}\sum_{h}\Pr(H=h)\Big(\underbrace{\sum_{z}\Pr(\tilde{Z}=z|H=h)v^{\theta}(o_t(h),z)}_{u_t^{\theta}(h)}-\underbrace{\sum_{z}\Pr(\tilde{Z}=z|H=h)}_{1}\hat{u}_t(o_t(h),\phi_t(h))\Big)\frac{\partial\hat{u}_t(o_t(h),\phi_t(h))}{\partial\psi}\\
&\sum_{t=0}^{\infty}\sum_{h}\Pr(H=h)\Big(u_t^{\theta}(h)-\hat{u}_t(o_t(h),\phi_t(h))\Big)\frac{\partial\hat{u}_t(o_t(h),\phi_t(h))}{\partial\psi}\\
&=\frac{1}{2}\nabla L(\psi).
\end{aligned}$$

However, recall that we cannot sample $v^{\theta}(O_t,\tilde{Z}_t)$ directly, and instead must rely on $\hat{v}(O_t,\tilde{Z}_t)$. However, due to the fact that the above estimators are unbiased, by choosing appropriate learning rate schedules for the two updates, we achieve the convergence guarantees of two-timescale stochastic gradient descent [2]. Due to the convexity of our loss functions, the system will converge to the global optimum almost surely.

2 Notes on POMDP Formulation

In order to clarify the concepts presented in this paper, maintain greater consistency with other work in the field of RL in the MDP setting, and to simplify certain proofs, we modified the standard POMDP formulation slightly. In this section, we briefly show that this modification causes no loss of generality, and show how to represent a standard POMDP is represented in our notation.

In the standard formulation, a POMDP is a tuple, $(\mathcal{S},\mathcal{A},T,R,\Omega,\mathcal{O})$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $T:\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to[0,1]$ is the transition probability function, $R:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$ is the reward function, Ω is the set of observations, and $O:\mathcal{S}\times\Omega\to[0,1]$ is the observation probability function. In our formulation, the observations are a component of $\mathcal{S}$, and the observation probabilities are defined by a single transition function, P.

Any standard POMDP can be converted to our formulation as follows: First, we define a new state set, $\mathcal{S}'=\mathcal{S}\times\Omega$, such that the hidden component set is $\mathcal{Z}=\mathcal{S}$ and the observable component set is $\mathcal{O}=\Omega$. Then, for any s and s' in $\mathcal{S}'$, where $s=(o,z)$ and $s'=(o',z')$, and any $a\in\mathcal{A}$, we define a new transition function $P(s,a,s')=T(z,a,z')O(z,o)$ and a reward function $R'(s,a)=R(z,a)$. Thus, we capture the dynamics of the original system.

3 Additional Experimental Details

All algorithms were tabular variants of REINFORCE [3] (sometimes also known as "vanilla policy gradient") with no discount factor. Each policy was implemented using softmax action selection. Each observation or belief state was represented as a string, which was mapped to a vector (containing one element for each action) which was then passed through a softmax function to compute the action probabilities. The vector was always initialized to zero, i.e., a uniform policy. At the end of each episode, the policy parameters were update by

$$\theta_{x_t,i+1} = \theta_{x_t,i} + \alpha_i(g_t - b_t(h))\frac{\partial \ln \pi_\theta(x_t, a_t)}{\partial \theta_i}, \tag{8}$$

where x_t is the observation or belief state at time t, i is the episode, $\theta_{x_t,i}$ are the policy parameters for x_t during episode i, α_i is the learning rate during episode i, g_t is the return, h is the entire history of the episode, and b_t is the baseline. Both the belief state were computed analytically from a known model. The baselines for the posterior models were updated according to Equation 14 in the main text. The baselines for the standard models were updated for all t according to

$$\hat{v}_{i+1}(x_t) = \hat{v}_i(x_t) + \beta_i(g_t - \hat{v}_i(x_t)), \tag{9}$$

where β_i is the critic learning rate at episode i and the other terms are as defined above.

The latent state is $Z \in \{0, 1, 2, 3\}$, corresponding to each gridworld, and is constant for all timesteps during an episode. The prior distribution is computed at time t by computing which gridwords are compatible with the history H_t using a known model and assigning the remaining gridworlds equal probability. The posterior distribution is computed similarly, except using the complete history H.

The actor learning rate was 0.01 and the critic learning rate was 0.05 for all agents. These were annealed linearly to 0 over 20,000 episodes. Each value function was pretrained for 5000 episodes on a uniform random policy.

The environment used is illustrated in the main text. The agent always starts in the top left square, $\{0, 0\}$, and the episode terminates upon reaching the bottom right square, $\{4, 4\}$. At the beginning of each episode, the agent is placed randomly in one of the four gridworlds with equal probability. Transitions are fully deterministic and the agent can move up, down, left, or right, moving one square at at time. If the agent attempts to move into a wall, the movement fails and the timestep advances. If the agent fails to reach the goal within 30, the episode terminates and the agent receives a punishment of -10.

References

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

[2] Vijay R Konda, John N Tsitsiklis, et al. Convergence rate of linear two-time-scale stochastic approximation. *Annals of Applied Probability*, 14(2):796–819, 2004.

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