A. Appendix

A.1. Proofs

Theorem A.1. *The expected target of PTD's forward view can be summarized using the operator*

$$\mathcal{T}^{\beta}\mathbf{v} = B(I - \gamma\mathcal{P}_{\pi}(I-B))^{-1}(r_{\pi} + \gamma\mathcal{P}_{\pi}B\mathbf{v}) + (I-B)\mathbf{v},$$

where B is the $|\mathcal{S}| \times |\mathcal{S}|$ diagonal matrix with $\beta(s)$ on its diagonal and r_{π} and $\mathcal{P}_{\pi}$ are the state reward vector and state-to-state transition matrix for policy π.

Proof. The expected target of Preferential TD in the vector form is given by,

$$\mathcal{T}^{\beta}\mathbf{v} = (I-B)v + B\Big(r_{\pi} + \gamma\mathcal{P}_{\pi}B\mathbf{v} + \gamma\mathcal{P}_{\pi}(I-B)r_{\pi} + \gamma^2\mathcal{P}_{\pi}(I-B)\mathcal{P}_{\pi}B\mathbf{v} + \dots\Big).$$

We can now express the reward terms and the value terms compactly by using the Neumann series expansion.

$$\begin{aligned}
\mathcal{T}^{\beta}\mathbf{v} &= (I-B)\mathbf{v} + B\Big(r_{\pi} + \gamma\mathcal{P}_{\pi}(I-B)r_{\pi} + (\gamma\mathcal{P}_{\pi}(I-B))^2 r_{\pi} + \dots \\
&\qquad + \gamma\mathcal{P}_{\pi}B\mathbf{v} + \gamma^2\mathcal{P}_{\pi}(I-B)\mathcal{P}_{\pi}B\mathbf{v} + \gamma^3(\mathcal{P}_{\pi}(I-B))^2\mathcal{P}_{\pi}B\mathbf{v} + \dots\Big), \\
&= (I-B)\mathbf{v} + B\Big((I - \gamma\mathcal{P}_{\pi}(I-B))^{-1}r_{\pi} + \gamma(I-\gamma\mathcal{P}_{\pi}(I-B))^{-1}\mathcal{P}_{\pi}B\mathbf{v}\Big), \\
&= B(I - \gamma\mathcal{P}_{\pi}(I-B))^{-1}(r_{\pi} + \gamma\mathcal{P}_{\pi}B\mathbf{v}) + (I-B)\mathbf{v}.
\end{aligned}$$

□

Lemma A.1. *The expected quantities* $\mathbf{A}$ *and* $\mathbf{b}$ *are given by* $\mathbf{A} = \Phi^T D^{\pi} B(I - \mathcal{P}_{\pi}^{\beta})$ *and* $\mathbf{b} = \Phi^T D^{\pi} B(I - \gamma\mathcal{P}_{\pi})^{-1} r_{\pi}$.

Proof.

$$\begin{aligned}
\mathbf{A} &= \lim_{t\to\infty} \mathbb{E}_{\pi}[\mathbf{A}(X_t)], \\
&= \lim_{t\to\infty} \mathbb{E}_{\pi}\Big[e_t\Big(\phi(s_t) - \gamma\phi(s_{t+1})\Big)^T\Big], \\
&= \sum_s d_{\pi}(s) \lim_{t\to\infty} \mathbb{E}_{\pi}\Big[e_t\Big(\phi(s_t) - \gamma\phi(s_{t+1})\Big)^T \Big| s_t = s\Big], \\
&= \sum_s d_{\pi}(s) \lim_{t\to\infty} \mathbb{E}_{\pi}\Big[\underbrace{\Big(\gamma(1-\beta(s_t))e_{t-1} + \beta(s_t)\phi(s_t)\Big)\Big(\phi(s_t) - \gamma\phi(s_{t+1})\Big)^T}_{\text{Independent due to conditioning on } s} \Big| s_t = s\Big], \\
&= \sum_s d_{\pi}(s) \lim_{t\to\infty} \mathbb{E}_{\pi}\Big[\gamma(1-\beta(s_t))e_{t-1} + \beta(s_t)\phi(s_t)\Big| s_t = s\Big]\ \mathbb{E}_{\pi}\Big[\Big(\phi(s_t) - \gamma\phi(s_{t+1})\Big)^T \Big| s_t = s\Big], \\
&= \sum_s e(s)\Big(\phi(s) - \gamma\sum_{s'} \mathcal{P}_{\pi}(s'|s)\phi(s')\Big)^T,
\end{aligned}$$

where $e(s) = d_{\pi}(s)\lim_{t\to\infty}\mathbb{E}_{\pi}[\gamma(1-\beta(s_t))e_{t-1} + \beta(s_t)\phi(s_t)|s_t = s]$, can be expanded as:

$$\begin{aligned}
e(s) &= d_{\pi}(s) \lim_{t\to\infty} \mathbb{E}_{\pi}[\gamma(1-\beta(s_t))e_{t-1} + \beta(s_t)\phi_t|s_t = s] \\
&= d_{\pi}(s)\beta(s)\phi(s) + \gamma(1-\beta(s))d_{\pi}(s) \lim_{t\to\infty} \mathbb{E}_{\pi}[e_{t-1}|s_t = s] \\
&= d_{\pi}(s)\beta(s)\phi(s) + \gamma(1-\beta(s))d_{\pi}(s) \lim_{t\to\infty} \sum_{\bar{s},\bar{a}} \mathbb{P}\{s_{t-1} = \bar{s}, a_{t-1} = \bar{a}|s_t = s\}\mathbb{E}_{\pi}[e_{t-1}|s_{t-1} = \bar{s}] \\
&= d_{\pi}(s)\beta(s)\phi(s) + \gamma(1-\beta(s))d_{\pi}(s) \sum_{\bar{s}} \frac{d_{\pi}(\bar{s})\mathcal{P}_{\pi}(s|\bar{s})}{d_{\pi}(s)} \lim_{t\to\infty} \mathbb{E}_{\pi}[e_{t-1}|s_{t-1} = \bar{s}] \\
&= d_{\pi}(s)\beta(s)\phi(s) + \gamma(1-\beta(s)) \sum_{\bar{s}} \mathcal{P}_{\pi}(s|\bar{s})e(\bar{s}).
\end{aligned}$$

We can express these quantities in a matrix as,

$$\begin{aligned}\mathbf{E}^T &= \Phi^T DB + \gamma \mathbf{E}^T \mathcal{P}_\pi (I - B),\\ &= \Phi^T DB + \gamma \Phi^T DB \mathcal{P}_\pi (I - B) + \gamma^2 \Phi^T DB (\mathcal{P}_\pi (I - B)))^2 + \dots,\\ &= \Phi^T DB (I - \gamma \mathcal{P}_\pi (I - B))^{-1}.\end{aligned}$$

We can perform a similar analysis on $\mathbf{b}$ and substitute $\mathbf{E}^T$ to get,

$$\mathbf{A} = \Phi^T DB (I - \gamma \mathcal{P}_\pi (I - B))^{-1} (I - \gamma \mathcal{P}_\pi)\Phi, \tag{7}$$

$$\mathbf{b} = \Phi^T DB (I - \gamma \mathcal{P}_\pi (I - B))^{-1} r_\pi. \tag{8}$$

These expressions can be simplified further by considering a new transition matrix $\mathcal{P}_\pi^\beta$ that accounts for the termination due to bootstrapping and discounted by γ. In other words, $\mathcal{P}_\pi^\beta$ is made up of $\mathcal{P}_\pi$, but the transitions are terminated according to B and continued according to $(I - B)$ at each step. This is a sub-stochastic matrix for $\gamma \in [0, 1)$ and a stochastic matrix when $\gamma = 1$. By definition,

$$\begin{aligned}\mathcal{P}_\pi^\beta &= \gamma \mathcal{P}_\pi B + \gamma \mathcal{P}_\pi (I - B) \gamma \mathcal{P}_\pi B + (\gamma \mathcal{P}_\pi (I - B))^2 \gamma \mathcal{P}_\pi B + \dots\\ &= \gamma \Big(\sum_{k=0}^{\infty} (\gamma \mathcal{P}_\pi (I - B))^k \Big) \mathcal{P}_\pi B \qquad (9)\\ &= \gamma (I - \gamma \mathcal{P}_\pi (I - B))^{-1} \mathcal{P}_\pi B \qquad (10)\\ &= (I - \gamma \mathcal{P}_\pi (I - B))^{-1} (-\gamma \mathcal{P}_\pi (I - B) + \gamma \mathcal{P}_\pi)\\ &= (I - \gamma \mathcal{P}_\pi (I - B))^{-1} (I - \gamma \mathcal{P}_\pi (I - B) + \gamma \mathcal{P}_\pi - I)\\ &= I - (I - \gamma \mathcal{P}_\pi (I - B))^{-1} (I - \gamma \mathcal{P}_\pi)\\ I - \mathcal{P}_\pi^\beta &= (I - \gamma \mathcal{P}_\pi (I - B))^{-1} (I - \gamma \mathcal{P}_\pi). \qquad (11)\end{aligned}$$

Therefore $\mathbf{A}$ and $\mathbf{b}$ are given by,

$$\mathbf{A} = \Phi^T DB (I - \mathcal{P}_\pi^\beta)\Phi, \tag{12}$$

$$\mathbf{b} = \Phi^T DB (I - \gamma \mathcal{P}_\pi (I - B))^{-1} r_\pi. \tag{13}$$

Under assumption 1, $\|\Phi\| \leq M$, where M is a constant. This implies each quantity in the above expression is bounded and well-defined. □

Theorem A.2. *The forward and the backward views of PTD are equivalent in expectation:*

$$\mathbf{b} - \mathbf{A}\mathbf{w} = \Phi^T D \Big(\mathcal{T}^\beta (\Phi \mathbf{w}) - \Phi \mathbf{w} \Big).$$

Proof. We have $\mathbf{A} = \Phi^T DB (I - \mathcal{P}_\pi^\beta)\Phi$ and $\mathbf{b} = \Phi^T DB (I - \gamma \mathcal{P}_\pi (I - B))^{-1} r_\pi$.

$$\begin{aligned}\mathbf{b} - \mathbf{A}\mathbf{w} &= \Phi^T DB (I - \gamma \mathcal{P}_\pi (I - B))^{-1} r_\pi - \Phi^T DB (I - \mathcal{P}_\pi^\beta)\Phi \mathbf{w},\\ &= \Phi^T DB (I - \gamma \mathcal{P}_\pi (I - B))^{-1} r_\pi + \Phi^T DB \mathcal{P}_\pi^\beta \Phi \mathbf{w} - \Phi^T DB \Phi \mathbf{w},\\ &= \Phi^T DB (I - \gamma \mathcal{P}_\pi (I - B))^{-1} r_\pi + \gamma \Phi^T DB (I - \gamma \mathcal{P}_\pi (I - B))^{-1} \mathcal{P}_\pi B \Phi \mathbf{w} - \Phi^T DB \Phi \mathbf{w},\\ &= \Phi^T D \Big(B (I - \gamma \mathcal{P}_\pi (I - B))^{-1} (r_\pi + \gamma \mathcal{P}_\pi B \Phi \mathbf{w}) - B \Phi \mathbf{w} \Big),\\ &= \Phi^T D \Big(B (I - \gamma \mathcal{P}_\pi (I - B))^{-1} (r_\pi + \gamma \mathcal{P}_\pi B \Phi \mathbf{w}) + (I - B)\Phi \mathbf{w} - \Phi \mathbf{w} \Big),\\ &= \Phi^T D \Big(\mathcal{T}^\beta (\Phi \mathbf{w}) - \Phi \mathbf{w} \Big).\end{aligned}$$

We used the definition of $\mathcal{P}_\pi^\beta$ (c.f equation 10) in the third step. □

Lemma A.2. *The column sums of* $\mathbf{A}$ *are positive.*

Proof. Let $\gamma = 1$. The column sums of the key matrix is given by $1^T DB(I - \mathcal{P}_\pi^\beta) = d_\pi B - d_\pi B \mathcal{P}_\pi^\beta$. We can expand $d_\pi B \mathcal{P}_\pi^\beta$ using equation 9,

$$\begin{aligned}
d_\pi B \mathcal{P}_\pi^\beta &= d_\pi B \Big(\sum_{k=0}^{\infty} (\mathcal{P}_\pi (I - B))^k \Big) \mathcal{P}_\pi B, \\
&= d_\pi \Big(B + B \sum_{k=1}^{\infty} (\mathcal{P}_\pi (I - B))^k \Big) \mathcal{P}_\pi B, \\
&= d_\pi \Big(I - (I - B) + B \sum_{k=1}^{\infty} (\mathcal{P}_\pi (I - B))^k \Big) \mathcal{P}_\pi B, \\
&= d_\pi \Big(I + \Big(- I + B \sum_{k=1}^{\infty} \mathcal{P}_\pi ((I - B)\mathcal{P}_\pi)^{k-1} \Big) (I - B) \Big) \mathcal{P}_\pi B, \\
&= d_\pi \Big(I + \underbrace{\Big(-I + B \sum_{k=0}^{\infty} \mathcal{P}_\pi ((I - B)\mathcal{P}_\pi)^{k} \Big)}_{=\mathbf{S}} (I - B) \Big) \mathcal{P}_\pi B, \\
&= \Big(d_\pi + d_\pi \mathbf{S} (I - B) \Big) \mathcal{P}_\pi B. \qquad (14)
\end{aligned}$$

Consider $d_\pi \mathbf{S}$,

$$\begin{aligned}
d_\pi \mathbf{S} &= d_\pi \Big(- I + B \sum_{k=0}^{\infty} \mathcal{P}_\pi ((I - B)\mathcal{P}_\pi)^k \Big), \\
&= d_\pi \Big(- I + B\mathcal{P}_\pi + B \sum_{k=1}^{\infty} \mathcal{P}_\pi ((I - B)\mathcal{P}_\pi)^k \Big), \\
&= d_\pi \Big(- I + \mathcal{P}_\pi - (I - B)\mathcal{P}_\pi + B \sum_{k=1}^{\infty} \mathcal{P}_\pi ((I - B)\mathcal{P}_\pi)^k \Big), \\
&= \underbrace{-d_\pi + d_\pi \mathcal{P}_\pi}_{=0,\ d_\pi \mathcal{P}_\pi = d_\pi} + d_\pi \Big(- (I - B)\mathcal{P}_\pi + B \sum_{k=1}^{\infty} \mathcal{P}_\pi ((I - B)\mathcal{P}_\pi)^k \Big), \\
&= d_\pi \Big(- I + B \sum_{k=1}^{\infty} \mathcal{P}_\pi ((I - B)\mathcal{P}_\pi)^{k-1} \Big) (I - B)\mathcal{P}_\pi, \\
&= d_\pi \Big(- I + B \sum_{k=0}^{\infty} \mathcal{P}_\pi ((I - B)\mathcal{P}_\pi)^{k} \Big) (I - B)\mathcal{P}_\pi, \\
&= d_\pi \mathbf{S} (I - B)\mathcal{P}_\pi.
\end{aligned}$$

We can recursively expand $d_\pi \mathbf{S}$ n times to get $d_\pi \mathbf{S} = d_\pi \mathbf{S}((I - B)\mathcal{P}_\pi)^n$. Notice that $(I - B)\mathcal{P}_\pi$ is a sub-stochastic matrix, therefore, the elements of the matrix keep getting smaller as n gets bigger. In fact, $\lim_{n\to\infty}((I - B)\mathcal{P}_\pi)^n = 0$. Therefore, $d_\pi \mathbf{S} = 0$ as $n \to \infty$. We can use this result in equation 14 to get,

$$d_\pi B \mathcal{P}_\pi^\beta = d_\pi \mathcal{P}_\pi B \ = d_\pi B. \qquad (15)$$

This implies, $d_\pi B(I - \mathcal{P}_\pi^\beta) = 0$ for $\gamma = 1$ and $d_\pi B(I - \mathcal{P}_\pi^\beta) > 0$ for $\gamma \in [0, 1)$ as $d_\pi B \mathcal{P}_\pi^\beta < d_\pi B$. Therefore, column sums are positive. □

Lemma A.3. *We have,* $\sum_{t=0}^{\infty} \|\mathbb{E}_\pi[A(X_t)|X_0] - \mathbf{A}\| \leq C_1$ *and* $\sum_{t=0}^{\infty} \|\mathbb{E}_\pi[b(X_t)|X_0] - \mathbf{b}\| \leq C_2$ *under assumption 2, where* C_1 *and* C_2 *are constants.*

Proof. By similar analysis to finding $\mathbf{A}$ (lemma 1), we have

$$\mathbb{E}_\pi[A(X_t)|X_0] = \Phi^T B D_t (I - \mathcal{P}_\pi^\beta)\Phi - \Phi^T B D_t \sum_{m=t+1}^{\infty} (\gamma \mathcal{P}_\pi (I - B))^m (I - \gamma \mathcal{P}_\pi)\Phi,$$

where $D_t(i,i) = \mathbb{P}(s_t = s_i|s_0)$. Now,

$$\begin{aligned}
\|\mathbb{E}_\pi[A(X_t)|X_0] - \mathbf{A}\| &= \left\|\Phi^T B(D_t - D)(I - \mathcal{P}_\pi^\beta)\Phi - \Phi^T B D_t \sum_{m=t+1}^{\infty} (\gamma \mathcal{P}_\pi (I - B))^m (I - \gamma \mathcal{P}_\pi)\Phi\right\|, \\
&\leq \left\|\Phi^T B(D_t - D)(I - \mathcal{P}_\pi^\beta)\Phi\right\| + \left\|\Phi^T B D_t \sum_{m=t+1}^{\infty} (\gamma \mathcal{P}_\pi (I - B))^m (I - \gamma \mathcal{P}_\pi)\Phi\right\|, \\
&\leq \underbrace{\left\|\Phi^T\right\| \|B\|}_{\text{constant}} \|D_t - D\| \underbrace{\left\|I - \mathcal{P}_\pi^\beta\right\| \|\Phi\|}_{\text{constant}} + \underbrace{\left\|\Phi^T\right\| \|B\| \|D_t\|}_{\text{constant}} \left\|\sum_{m=t+1}^{\infty} (\gamma \mathcal{P}_\pi (I - B))^m\right\| \underbrace{\|(I - \gamma \mathcal{P}_\pi)\| \|\Phi\|}_{\text{constant}}, \\
&\leq K_1 C \rho^t + K_2 \sum_{m=t+1}^{\infty} \gamma^m \underbrace{\|(\mathcal{P}_\pi (I - B))^m\|}_{\leq K_5, \forall m}, \\
&\leq K_1 C \rho^t + K_3 \sum_{m=t+1}^{\infty} \gamma^m, \\
&= K_4 \rho^t + \frac{\gamma^{t+1}}{1-\gamma} K_3,
\end{aligned}$$

where K_1, K_2, K_3, K_4, K_5 are constants, we have used the triangle inequality and $\|AB\| \leq \|A\| \cdot \|B\|$ properties. Now,

$$\begin{aligned}
\sum_{t=0}^{\infty} \|\mathbb{E}_\pi[A(X_t)|X_0] - \mathbf{A}\| &\leq \sum_{t=0}^{\infty} K_3 \rho^t + \frac{\gamma^{t+1}}{1-\gamma} K_2, \\
&= K_3 \frac{1}{1-\rho} + K_2 \frac{\gamma}{(1-\gamma)^2}, \\
&= C_1.
\end{aligned}$$

By a similar analysis, we get $\sum_{t=0}^{\infty} \|\mathbb{E}_\pi[b(X_t)|X_0] - \mathbf{b}\| \leq C_2$. Therefore, the bias due to sampling is contained. □

A.2. Standard results used in the proof

Corollary A.1. *(Varga, 1999) If* $\mathbf{A}$ *is a Hermitian* $n \times n$ *strictly diagonally dominant or irreducibly diagonally dominant matrix with positive real diagonal entries, then* $\mathbf{A}$ *is positive definite.*

Theorem A.3. *(Bertsekas & Tsitsiklis, 1996; Tsitsiklis & Van Roy, 1997) Consider an iterative algorithm of the form* $\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha_t(A(X_t)\mathbf{w}_t + b(X_t))$ *where,*

1. *the step-size sequence* α_t *satisfies assumption 3;*
2. X_t *is a Markov process with a unique invariant distribution, and there exists a mapping* h *from the states of the Markov process to the positive reals, satisfying the remaining conditions;*
3. $A(\cdot)$ *and* $b(\cdot)$ *are matrix and vector valued functions respectively, for which* $\mathbf{A} = \mathbb{E}_{d_\pi}[A(X_t)]$ *and* $\mathbf{b} = \mathbb{E}_{d_\pi}[b(X_t)]$ *are well defined and finite;*
4. *the matrix* $\mathbf{A}$ *is negative definite;*
5. *there exists constants* C *and* q *such that for all* X
   - $\sum_{t=0}^{\infty} \|\mathbb{E}_\pi[A(X_t)|X_0 = X] - \mathbf{A}\| \leq C(1 + h^q(X))$, *and*
   - $\sum_{t=0}^{\infty} \|\mathbb{E}_\pi[b(X_t)|X_0 = X] - \mathbf{b}\| \leq C(1 + h^q(X))$;
6. *for any* $q > 1$ *there exists a constant* μ_q *such that for all* X, t
   - $\mathbb{E}_\pi[h^q(X_t)|X_0 = X] \leq \mu_q(1 + h^q(X))$.

Then, $\mathbf{w}_t$ *converges to* $\mathbf{w}_\pi$*, with probability one, where* $\mathbf{w}_\pi$ *is the unique vector that satisfies* $\mathbf{A}\mathbf{w}_\pi + \mathbf{b} = 0$.

A.3. Counterexamples of TD(λ)

Example 2 (Ghiassian et al., 2017): Two state MDP with $\mathcal{P}_\pi = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, $\Phi = \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}$, $\lambda = \begin{bmatrix} 0 & 0.99 \end{bmatrix}$, $d_\pi = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}$, $\gamma = 0.95$. The key matrix of TD(λ) is given by, $\mathbf{A} = \begin{bmatrix} -0.46 & 0.15 \\ -0.77 & 0.07 \end{bmatrix}$, which is not positive definite. Therefore, the updates will not converge. The key matrix of Preferential TD for the same setup is $\mathbf{A} = \begin{bmatrix} 0.46 & 0.15 \\ 0.15 & 0.05 \end{bmatrix}$ which is positive definite.

A.4. Experiments

A.4.1. Tabular setting

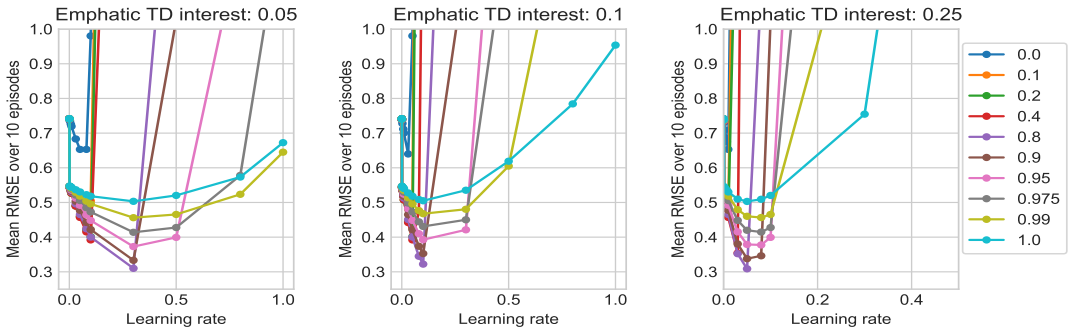


Figure A.1. Performance of Emphatic TD with a fixed interest value (shown in the title of the plot). Average RMSE of the first 10 episodes is plotted against learning rate for different choices of λ.

We analyzed the bias-variance trade-off of Emphatic TD with several choices of fixed interest values. ETD with a low interest value resulted in lower errors compared to high interest values. The difference in performance is due to the algorithm's sensitivity to learning rates when high interest is used. We present the root mean squared error obtained for various fixed interests in Figure A.1. Different curves in the plot correspond to various values of λ. We generated the plots by averaging the root mean squared error of the value function over 10 episodes of training across 25 seeds.

A.4.2. Linear setting

The error curves for the corridor lengths $\{10, 20\}$ are provided in Figure A.2. The observations made in the main paper holds true for these lengths as well.

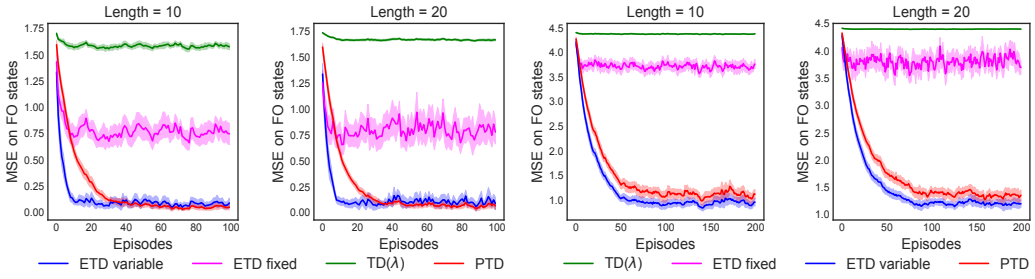


Figure A.2. The mean squared error of fully observable states' values is plotted as a function of episodes for various algorithms. The first two plots correspond to the results on Task 1 (left) and the next plots correspond to Task 2 (right). The corridor length is indicated in the title of the plot.

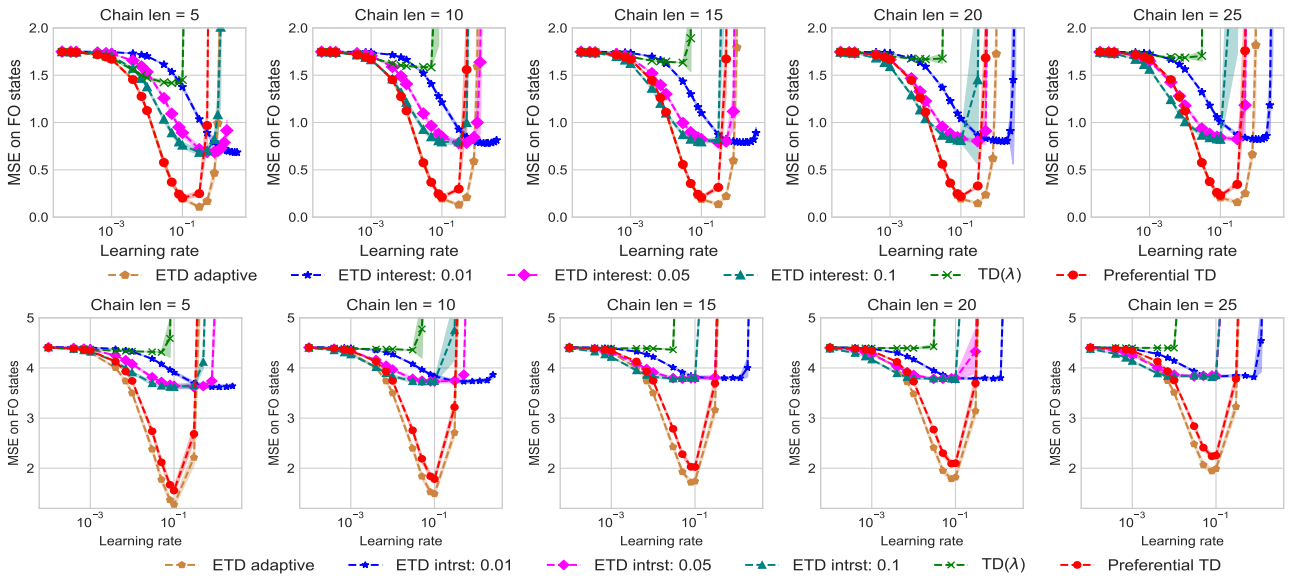


Figure A.3. The average mean squared error of the fully observable states' values is plotted against learning rate on Task 1 (top) and Task 2 (bottom).

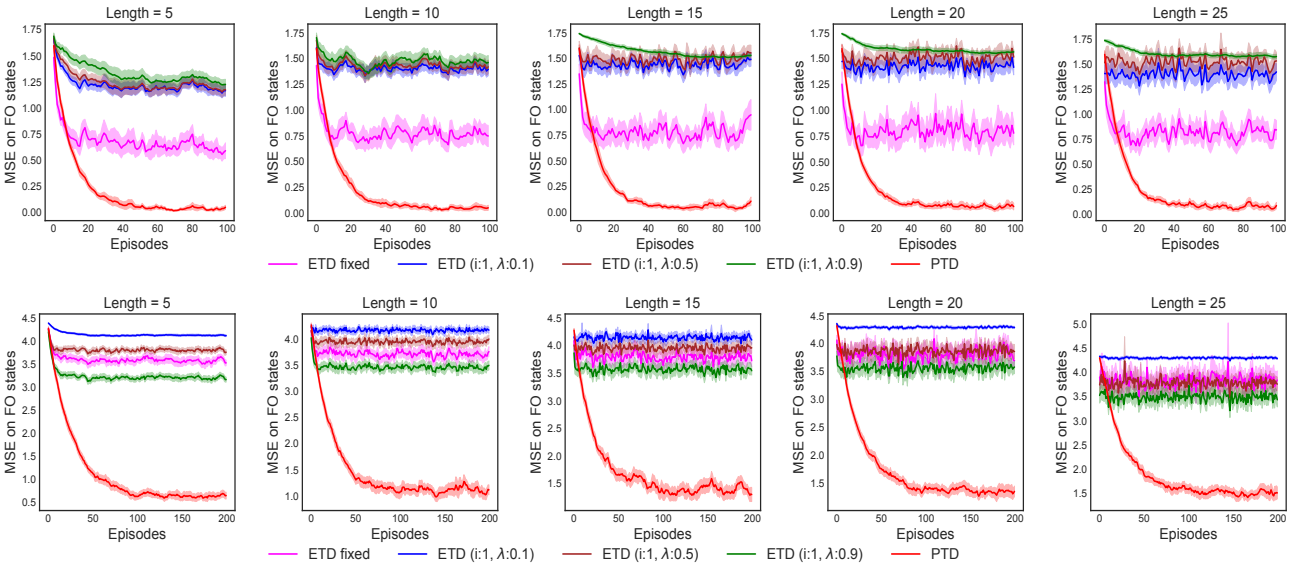


Figure A.4. The mean squared error of the fully observable states' values is plotted against episodes on Task 1 (top) and Task 2 (bottom).

Task 1: We chose the best learning rate from {1.2, 1.0, 0.8, 0.5, 0.3, 0.1, 8e-2, 5e-2, 3e-2, 1e-2, 7e-3, 4e-3, 1e-3, 7e-4, 4e-4, 1e-4, 7e-5, 4e-5}. We extended the range of search to include {10.0, 5.0, 4.0, 3.5, 3.0, 2.5, 2.0, 1.8, 1.5,} for ETD-fixed. We also ran ETD-fixed on 3 different interests ({0.01, 0.05, 0.1}). We ran all the algorithms for 100 episodes. We calculated the mean squared error of fully observable states' values and averaged it across 100 episodes and 25 seeds for the above mentioned learning rates. We chose the learning rate that resulted in the lowest error. The hyperparameter tuning plot is presented in Figure A.3. We ran all the algorithms on 25 different seeds and plotted the mean error with a confidence interval of 0.5 times the standard deviation.

Task 2: The learning rate for this task was selected from the set {0.8, 0.5, 0.3, 0.1, 8e-2, 5e-2, 3e-2, 1e-2, 7e-3, 4e-3, 1e-3, 7e-4, 4e-4, 1e-4}. We extended the range of search to include {2.5, 1.8, 1.2,} for ETD-fixed. We also ran ETD-fixed on 3 different interest values ({0.01, 0.05, 0.1}). We calculated the mean squared error of fully observable states' values and

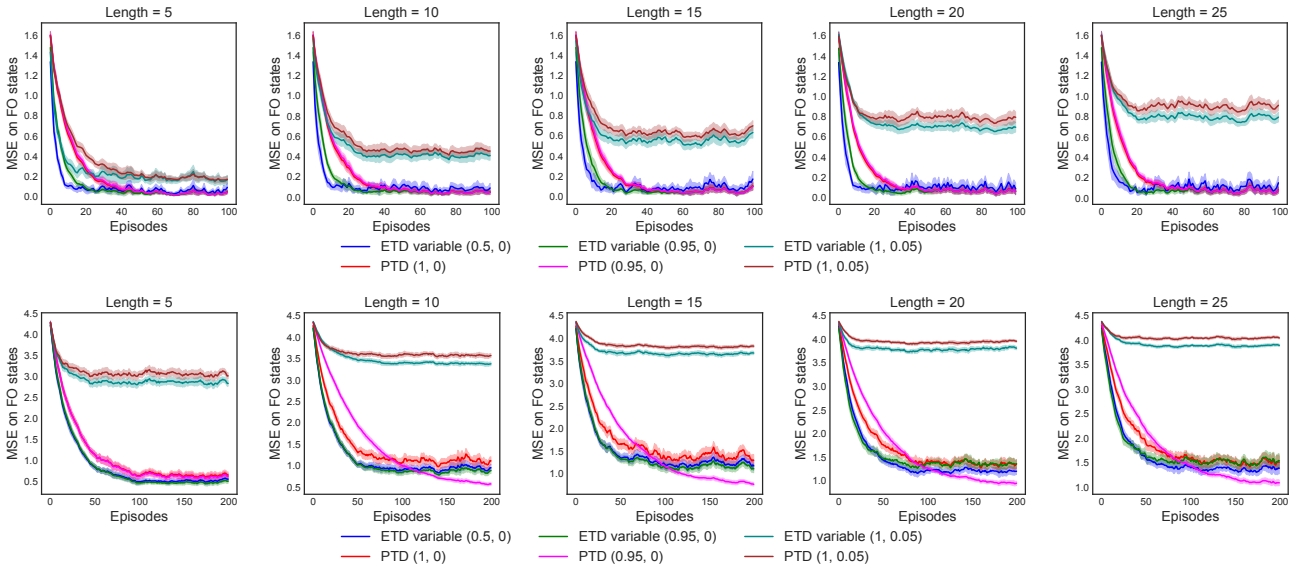


Figure A.5. Performance plot of PTD and ETD-variable for different choices of β (or i and λ) for FO and PO states on Task 1 (top) and Task 2 (bottom).

averaged it across 100 episodes and 25 seeds for the above mentioned learning rates. We choose the learning rate that resulted in the lowest error. The hyperparameter tuning plot is presented in Figure A.3. We ran all the algorithms on 25 different seeds and plotted the mean with a confidence interval of 0.5 times the standard deviation. We ran 200 episodes to generate the learning curves presented in Figure 3 and A.2 using the optimal learning rate.

ETD-fixed with $i = 1$ and a constant λ for all the states: In this experiment, we tested the performance of a different version of ETD-fixed. In this version, we set the interest and λ to constant values for all the states. The interest value was set to 1, and we experimented with three separate values of λ - $\{0.1, 0.5, 0.9\}$. Like the earlier version, this too performs poorly compared to PTD and ETD-variable. This is because, setting $i = 1$ results in updating all the states including those that are partially observable causing poor generalization. The results are presented in Figure A.4.

PTD and ETD-variable with non-extreme values: In this experiment, we tested PTD and ETD-variable with non-extreme preference and (interest, λ) values respectively. The performance is not affected when β (or i) of fully observable states is set to a value < 1 as the updates on partially observable states is still blocked. However, the performance drops significantly when β (or i) of partially observable states is set to non-zero values. The results are presented in Figure A.5.

A.4.3. Semi-linear setting

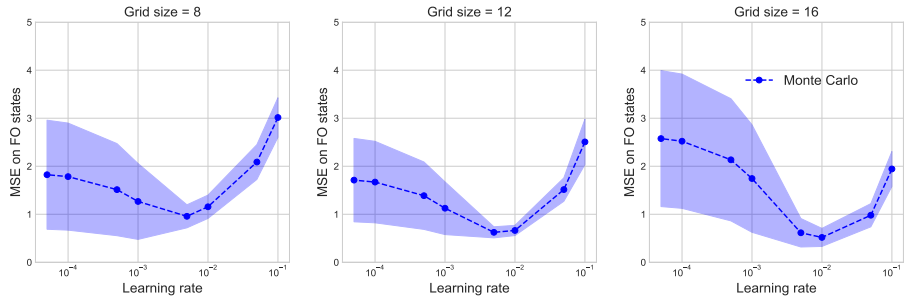


Figure A.6. Average mean squared error of the value function is plotted against learning rate for the feature net. Different plots correspond to various grid sizes.

Feature net: Feature net is a single-layered neural network with 32 neurons with ReLU non-linearity in the hidden layer. The network's input is a one-hot vector of size $n \times n$, where n is the size of the grid. The component corresponding to

the agent's location is set to 1 and the remaining bits are set to 0. The network is trained to minimize the mean squared error of Monte Carlo returns and its predictions for the states visited in a trajectory. We train the network using ADAM optimizer. We find the optimal learning rate from {1e-1, 5e-2, 1e-2, 5e-3, 1e-3, 5e-4, 1e-4, 5e-5} for every grid size through hyperparameter search. The networks are trained on 50 episodes on each grid size and the mean squared error of the value function across 50 episodes is used as a metric to pick the optimal learning rate. The hyperparameter tuning results are presented in Figure A.6. We run experiments on 25 seeds and use a confidence interval of 0.5 times the standard deviation for plotting.

Linear: A linear function approximator is used to estimate the value function of the fully observable states. The hidden layer output of a fully trained feature net is used as features for the input state. A one-hot vector (for fully observable states) or a Gaussian vector (each component is generated from $\mathcal{N}(0, 1)$ for partially observable states) is used as an input to the feature net to get features for the downstream task.

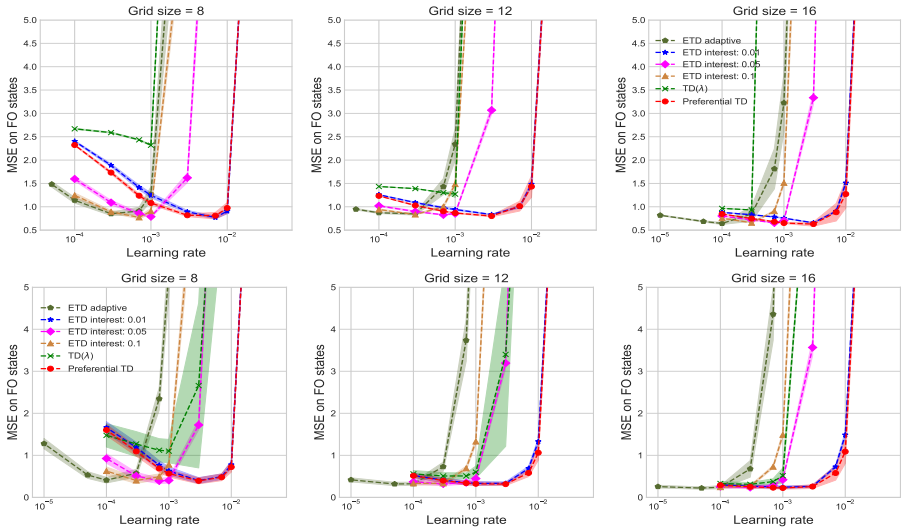
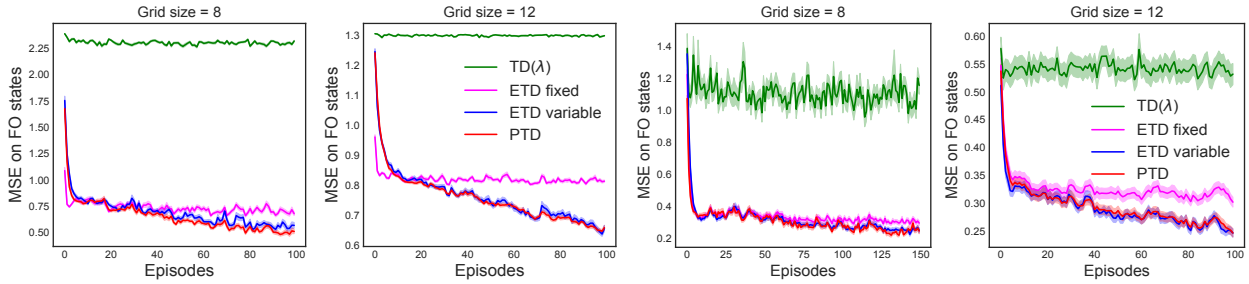


Figure A.7. The average mean squared error of fully observable states' values plotted against various learning rates for Task 1 (top) and Task 2 (bottom).



(a) Learning curves on Task 1

(b) Learning curves on Task 2

Figure A.8. The mean squared error of fully observable states' values plotted against episodes for various algorithms. The grid size is indicated in the title of the plot.

Task 1: We chose the optimal learning rate from {5e-2, 1e-2, 7e-3, 3e-3, 1e-3, 7e-4, 3e-4, 1e-4}. We extended the range of search to include {5e-5, 1e-5} for ETD-variable. We also ran ETD-fixed on 3 different interest values - ({0.01, 0.05, 0.1}). We ran all the algorithms for 50 episodes. We calculated the mean squared error of fully observable states' values over 50 episodes for the learning rates mentioned earlier and chose the best learning rate based on the average MSE across episodes and seeds. The hyperparameter tuning plot is presented in Figure A.7. We ran all the algorithms on 25 different seeds and plotted the mean MSE with a confidence interval of 0.5 times the standard deviation. We ran 100 episodes for 8×8 and

12×12 grids, 50 episodes for 16×16 grid to generate the learning curves presented in the main paper using the optimal learning rates.

Task 2: All the experimental details are same as task 1. The hyperparameter tuning results are presented in Figure A.7. Once the best learning rate was found for all the algorithms, we ran 150 episodes for 8×8 grid, 100 episodes for 12×12 and 16×16 grids to generate the learning curves presented in Figure 4 and A.8.

Additional results: The error curves for the grid sizes 8 and 12 are presented in Figure A.8. The observations made in the main paper hold true for these grids too. However, both ETD-variable and PTD perform similarly. The other two methods result in poor predictions.

A.4.4. Non-Linear setting

Forward view: For each algorithm and network combination, we find the best learning rate from {5e-2, 1e-2, 5e-3, 1e-3, 5e-4}. We included 1e-4 in the search space for grid task 2. The best learning rate was decided based on the area under the error curve (AUC) of MSE averaged across episodes and seeds. We used SGD optimizer to train the network. We consider 250 episodes and 25 seeds for all the tasks for hyperparameter tuning and use 50 seeds to plot the final results. The hyperparameter tuning results are presented in Figures A.9 and A.10.

Backward view: For each algorithm and network combination, we find the best learning rate from {0.1, 5e-2, 1e-2, 5e-3, 1e-3, 5e-4, 1e-4}. The best learning rate was decided using the process described in *Forward view*. Plotting details also remain the same as before. The hyperparameter tuning results are presented in Figures A.11 and A.12.

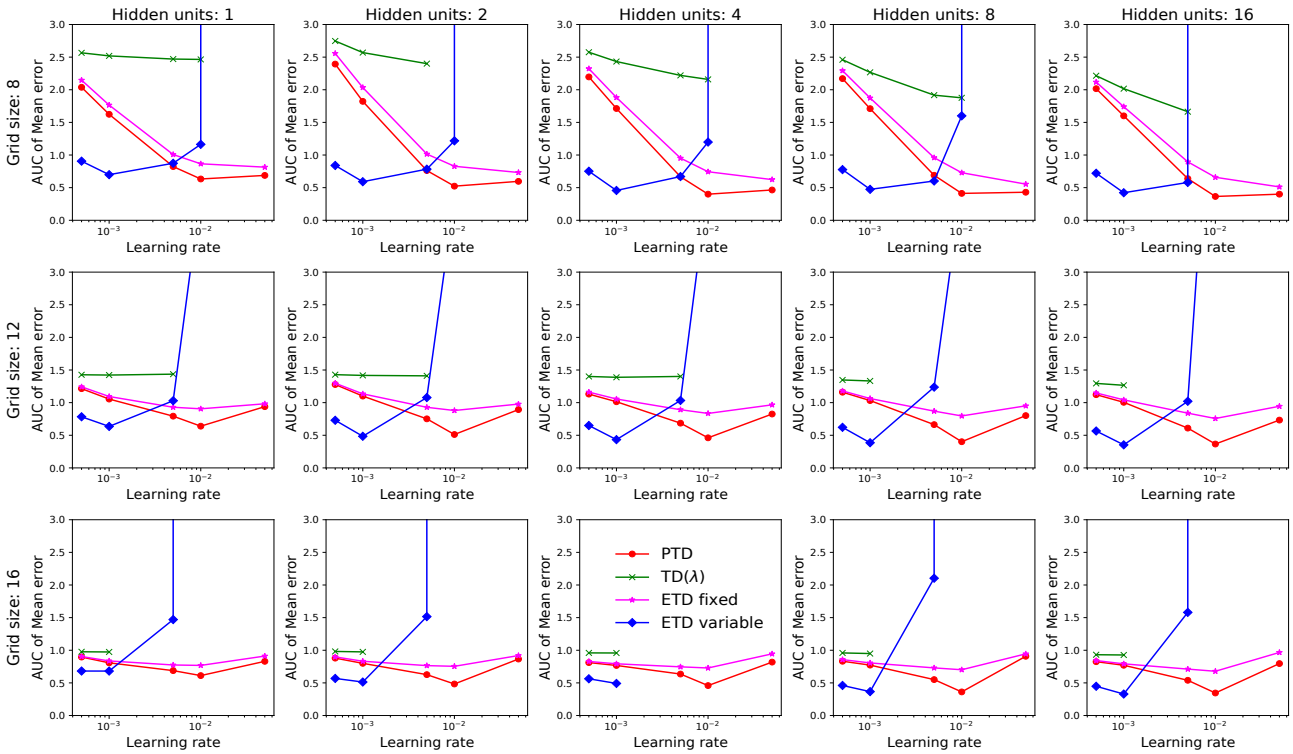


Figure A.9. Task 1 forward view hyperparameter tuning curves. AUC of mean error is plotted against learning rates.

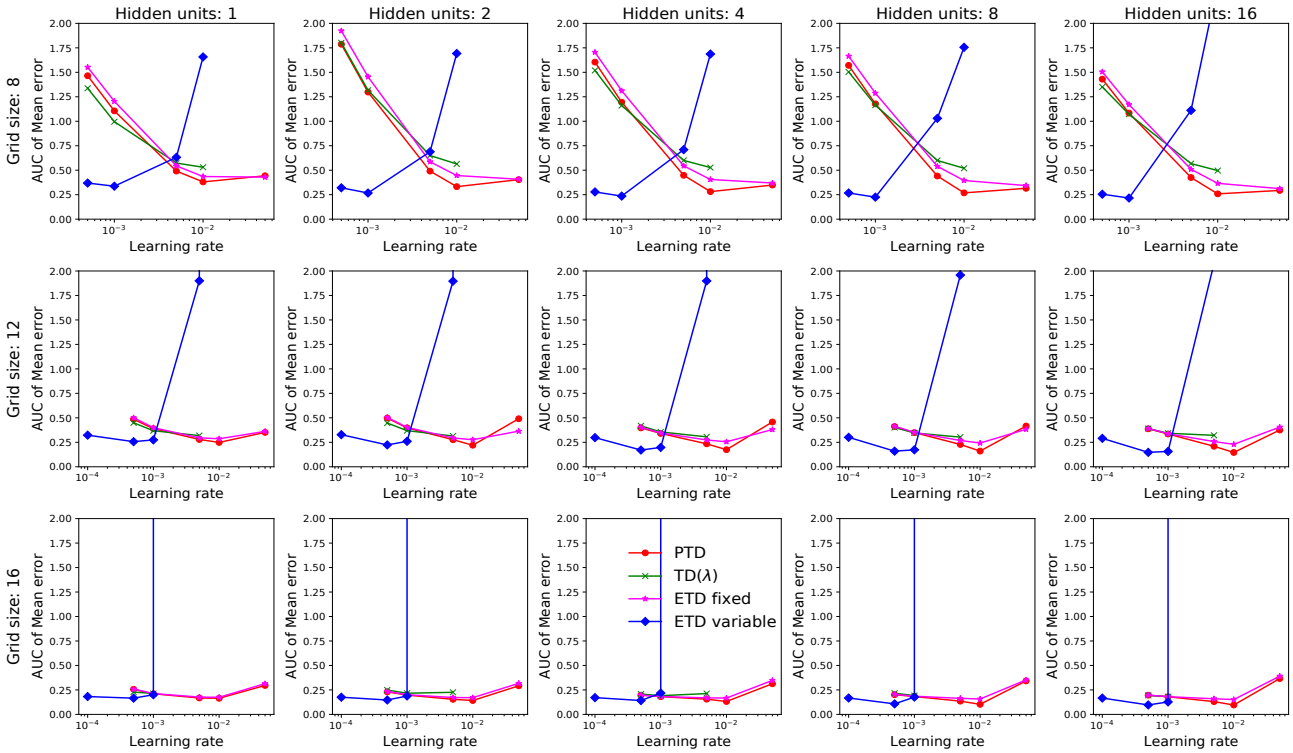


Figure A.10. Task 2 forward view hyperparameter tuning curves. AUC of mean error is plotted against learning rates.

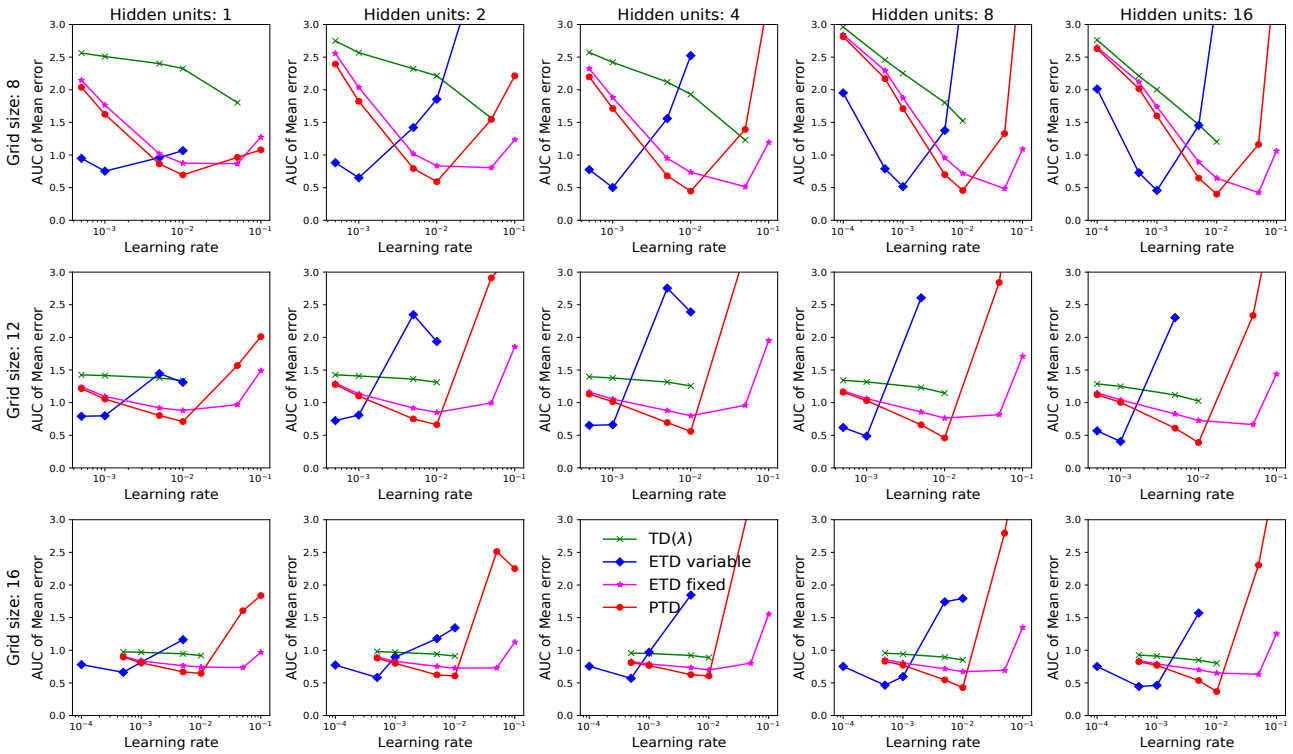


Figure A.11. Task 1 backward view hyperparameter tuning curves. AUC of mean error is plotted against learning rates.

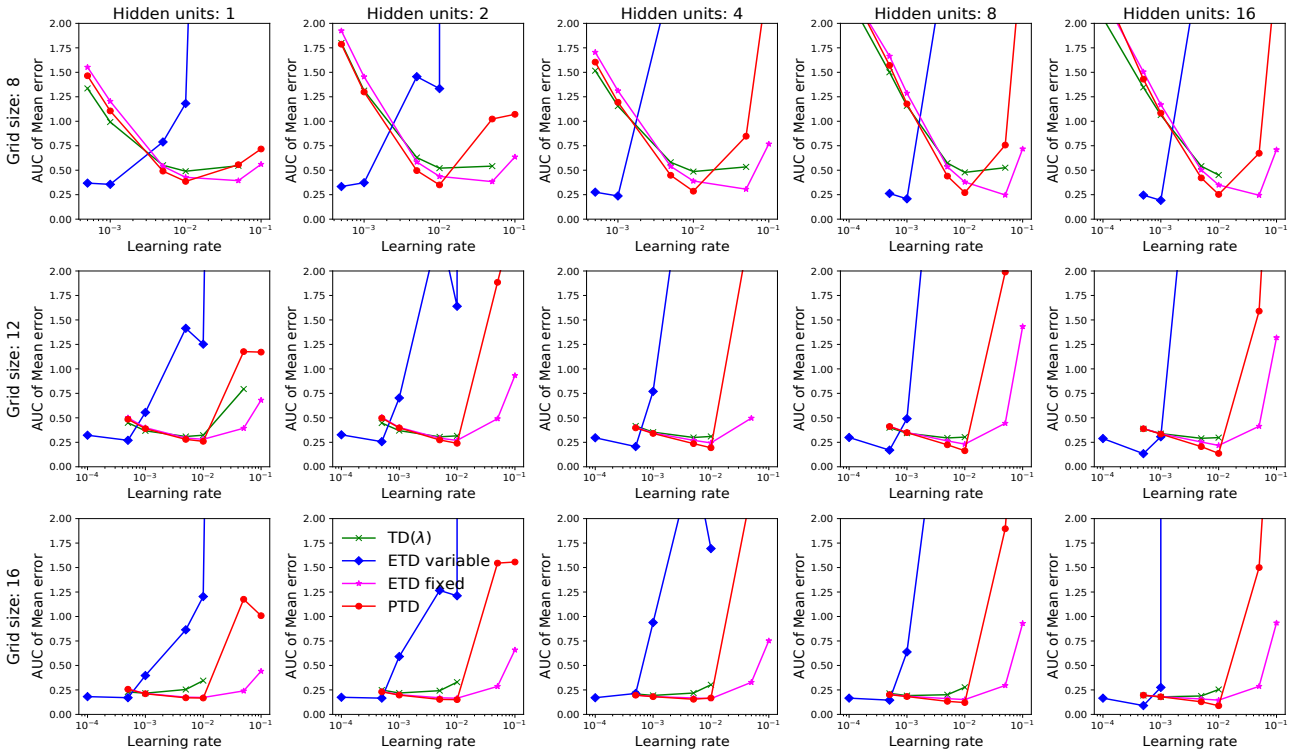


Figure A.12. Task 2 backward view hyperparameter tuning curves. AUC of mean error is plotted against learning rates.

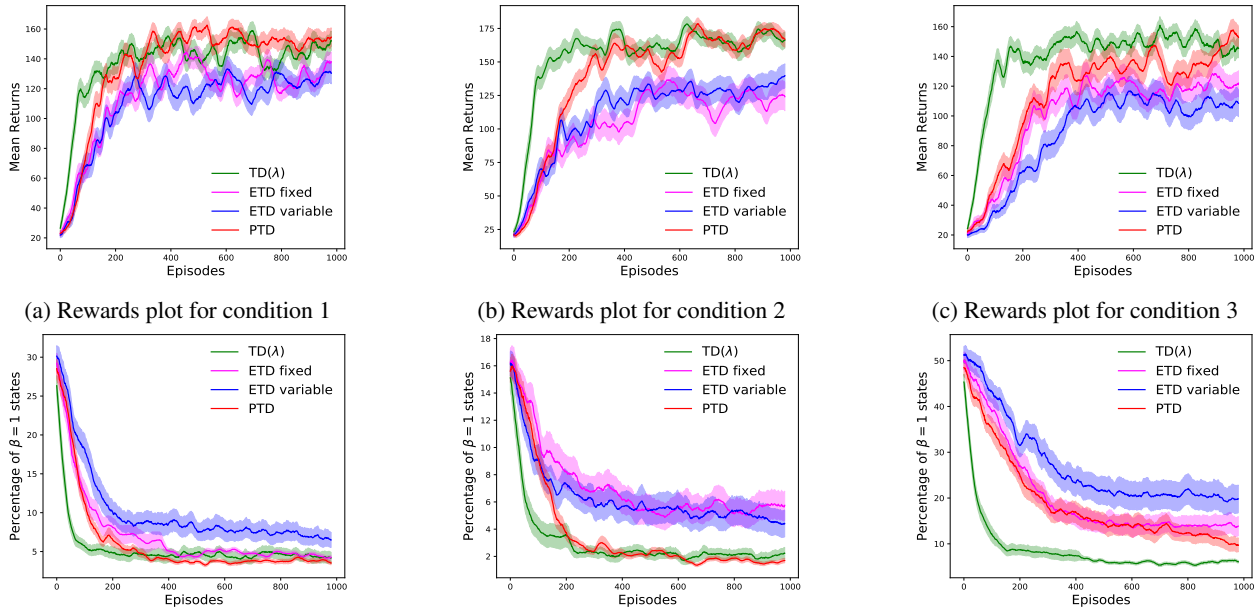
A.4.5. CONTROL - ACTOR CRITIC

Algorithm 2 Preferential TD: Actor-Critic

Input: γ,β, ϕ
Initialize: $\mathbf{w}_{act} = 0, \mathbf{w}_{cri} = 0$
Output: $\mathbf{w}_{act}, \mathbf{w}_{cri}$
for all episodes **do**
  for each step in an episode **do**
    Compute one-step TD error δ_t
    Update $\mathbf{w}_{cri}$ using PTD traces
  end for
  Compute forward view returns G_t^β for each state
  Update $\mathbf{w}_{act}$ using $\nabla_{\mathbf{w}_{act}} \log \pi(a|s)\beta(s)(G_t^\beta - \mathbf{w}_{cri}^T \phi(s))$ for every state
end for

Task description: We consider the cartpole task from the openAI gym (Brockman et al., 2016). The goal of the task is to learn a policy to balance the pole using the algorithm described in 2. We provide this experiment to demonstrate that Preferential TD works (1) when the environment is fully observable, and (2) in control setting.

Setup: We consider actor-critic algorithm for learning the optimal policy. The actor is a single layered neural network with 4 hidden units. The critic is a linear function approximator. The actor is updated at the end of each episode since it is a non-linear function approximator. We use forward view returns to compute the error for updating the actor. The critic is updated at each timestep using eligibility traces. We consider a non-linear approximator for actor as we were not able solve the task using a linear actor on the raw input. We consider the same 4 algorithms for experimentation. For each algorithm we picked the learning rate for the actor from {0.1, 5e-2, 1e-2, 5e-3, 1e-3} and for the critic from {1e-3, 5e-4, 1e-4, 5e-5, 1e-5, 5e-6, 1e-6}. The best learning rate was decided based on the average returns achieved per episode across 25 seeds and 500 episodes. The interest for both ETD-variable and ETD-fixed was set to 0.5. $\beta = 1$ (or $\lambda = 0$) for a state if the difference between the current cart position (or cart velocity) and the cart position (or cart velocity) of the previous $\beta = 1$



(a) Rewards plot for condition 1 (b) Rewards plot for condition 2 (c) Rewards plot for condition 3

(d) Beta plot for condition 1 (e) Beta plot for condition 2 (f) Beta plot for condition 3

Figure A.13. (*Top*) Return is plotted against episodes. (*Bottom*) Percentage of $\beta = 1$ (or $\lambda = 0$) states plotted against episodes. Different columns correspond to different thresholds used to set $\beta = 1$ (or $\lambda = 0$). The plots are smoothed using a moving average window of 20 episodes. The shaded region represents a confidence interval of 50% on 25 seeds.

(or $\lambda = 0$) state is greater than a threshold. A similar condition was also checked for pole angle with a different threshold. The thresholds for all 3 conditions are provided in table A.1. We experimented with three combinations of threshold values to set β value to 1. All the states in between two $\beta = 1$ ($\lambda = 0$) states were set to a constant value of 0.1 ($\lambda = 0.9$). We also experimented with $\{0, 0.5\}$ as intermediate values but their performance was relatively poor. All the results are averaged across 25 different seeds. We set the discount factor to 0.99.

Table A.1. Thresholds to determine $\beta = 1$ state.

Condition	Cart position	Cart velocity	Pole angle
1	0.5	0.5	0.05
2	1.0	1.0	0.1
3	0.3	0.3	0.03

Observations: The environment is fully observable and therefore we expect to observe the best performance when critic and actor are updated for all the states. TD(λ) updates the values and policy for all the states and therefore it achieves the best performance as seen in Figure A.13. ETD-fixed and ETD-variable achieves a similar performance. The performances are substantially poorer compared to the other methods. We suspect that it is because of high variance in the updates resulting from trajectory dependent emphasis. This observation is also validated from the beta plots (Figure A.13), where we observe significantly higher $\beta = 1$ states. The performances drop further for condition 3, where the percentage of $\beta = 1$ states are much higher compared to the other conditions. Besides, hyperparameter tuning was found to be challenging as these algorithms were sensitive to them. Preferential TD's performance was as good as TD(λ). However, PTD updates and bootstraps from a small percentage (roughly 5%) of states. PTD learning is not sharp at the beginning when the estimates are poor. But it catches up quickly and the overall performance is on par with TD(λ). We also observed a much lower variance across seeds in PTD compared to the other algorithms. This experiment validates our hypothesis that a similar performance can be achieved by bootstrapping and updating from a small set of states as opposed to all. Also, this experiment provides preliminary evidence for PTD's usage in fully observable setting and in settings beyond policy evaluation. Although further analysis, both theoretical and empirical, is required to make meaningful conclusions.