Bridging the Gap Between Value and Policy Based Reinforcement Learning
(Supplementary Material)

A Pseudocode

Pseudocode for PCL is presented in Algorithm 1.

B Experimental Details

We describe the tasks we experimented on as well as details of the experimental setup.

B.1 Synthetic Tree

As an initial testbed, we developed a simple synthetic environment. The environment is defined by a binary decision tree of depth 20. For each training run, the reward on each edge is sampled uniformly from $[-1, 1]$ and subsequently normalized so that the maximal reward trajectory has total reward 20. We trained using a fully-parameterized model: for each node s in the decision tree there are two parameters to determine the logits of $\pi_\theta(-|s)$ and one parameter to determine $V_\phi(s)$. In the Q-learning and Unified PCL implementations only two parameters per node s are needed to determine the Q-values.

Algorithm 1 Path Consistency Learning

```
Input: Environment $ENV$, learning rates $\eta_\pi, \eta_v$, discount factor $\gamma$, rollout $d$, number of steps $N$,
replay buffer capacity $B$, prioritized replay hyperparameter $\alpha$.
function Gradients($s_{0:T}$)
    // We use $G(s_{t:t+d}, \pi_\theta)$ to denote a discounted sum of log-probabilities from $s_t$ to $s_{t+d}$.
    Compute $\Delta\theta = \sum_{t=0}^{T-d} C_{\theta,\phi}(s_{t:t+d}) \nabla_\theta G(s_{t:t+d}, \pi_\theta)$.
    Compute $\Delta\phi = \sum_{t=0}^{T-d} C_{\theta,\phi}(s_{t:t+d}) \big(\nabla_\phi V_\phi(s_t) - \gamma^d \nabla_\phi V_\phi(s_{t+d})\big)$.
    Return $\Delta\theta, \Delta\phi$
end function
Initialize $\theta, \phi$.
Initialize empty replay buffer $RB(\alpha)$.
for $i = 0$ to $N - 1$ do
    Sample $s_{0:T} \sim \pi_\theta(s_{0:})$ on $ENV$.
    $\Delta\theta, \Delta\phi =$ Gradients($s_{0:T}$).
    Update $\theta \leftarrow \theta + \eta_\pi \Delta\theta$.
    Update $\phi \leftarrow \phi + \eta_V \Delta\phi$.
    Input $s_{0:T}$ into $RB$ with priority $R^1(s_{0:T})$.
    If $|RB| > B$, remove episodes uniformly at random.
    Sample $s_{0:T}$ from $RB$.
    $\Delta\theta, \Delta\phi =$ Gradients($s_{0:T}$).
    Update $\theta \leftarrow \theta + \eta_\pi \Delta\theta$.
    Update $\phi \leftarrow \phi + \eta_v \Delta\phi$.
end for
```

B.2 Algorithmic Tasks

For more complex environments, we evaluated PCL, Unified PCL, and the two baselines on the algorithmic tasks from the OpenAI Gym library [4]. This library provides six tasks, in rough order of difficulty: Copy, DuplicatedInput, RepeatCopy, Reverse, ReversedAddition, and ReversedAddition3. In each of these tasks, an agent operates on a grid of characters or digits, observing one character or digit at a time. At each time step, the agent may move one step in any direction and optionally write a character or digit to output. A reward is received on each correct emission. The agent's goal for each task is:

- **Copy:** Copy a $1 \times n$ sequence of characters to output.
- **DuplicatedInput:** Deduplicate a $1 \times n$ sequence of characters.
- **RepeatCopy:** Copy a $1 \times n$ sequence of characters first in forward order, then reverse, and finally forward again.
- **Reverse:** Copy a $1 \times n$ sequence of characters in reverse order.
- **ReversedAddition:** Observe two ternary numbers in little-endian order via a $2 \times n$ grid and output their sum.
- **ReversedAddition3:** Observe three ternary numbers in little-endian order via a $3 \times n$ grid and output their sum.

These environments have an implicit curriculum associated with them. To observe the performance of our algorithm without curriculum, we also include a task "Hard ReversedAddition" which has the same goal as ReversedAddition but does not utilize curriculum.

For these environments, we parameterized the agent by a recurrent neural network with LSTM [5] cells of hidden dimension 128.

B.3 Implementation Details

For our hyperparameter search, we found it simple to parameterize the critic learning rate in terms of the actor learning rate as $\eta_v = C\eta_\pi$, where C is the *critic weight*.

For the Synthetic Tree environment we used a batch size of 10, rollout of $d = 3$, discount of $\gamma = 1.0$, and a replay buffer capacity of 10,000. We fixed the α parameter for PCL's replay buffer to 1 and used $\epsilon = 0.05$ for DQN. To find the optimal hyperparameters, we performed an extensive grid search over actor learning rate $\eta_\pi \in \{0.01, 0.05, 0.1\}$; critic weight $C \in \{0.1, 0.5, 1\}$; entropy regularizer $\tau \in \{0.005, 0.01, 0.025, 0.05, 0.1, 0.25, 0.5, 1.0\}$ for A3C, PCL, Unified PCL; and $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, $\beta \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ for DQN replay buffer parameters. We used standard gradient descent for optimization.

For the algorithmic tasks we used a batch size of 400, rollout of $d = 10$, a replay buffer of capacity 100,000, ran using distributed training with 4 workers, and fixed the actor learning rate η_π to 0.005, which we found to work well across all variants. To find the optimal hyperparameters, we performed an extensive grid search over discount $\gamma \in \{0.9, 1.0\}$, $\alpha \in \{0.1, 0.5\}$ for PCL's replay buffer; critic weight $C \in \{0.1, 1\}$; entropy regularizer $\tau \in \{0.005, 0.01, 0.025, 0.05, 0.1, 0.15\}$; $\alpha \in \{0.2, 0.4, 0.6, 0.8\}$, $\beta \in \{0.06, 0.2, 0.4, 0.5, 0.8\}$ for the prioritized DQN replay buffer; and also experimented with exploration rates $\epsilon \in \{0.05, 0.1\}$ and copy frequencies for the target DQN, $\{100, 200, 400, 600\}$. In these experiments, we used the Adam optimizer [6].

All experiments were implemented using Tensorflow [1].

C Proofs

In this section, we provide a general theoretical foundation for this work, including proofs of the main path consistency results. We first establish the basic results for a simple one-shot decision making setting. These initial results will be useful in the proof of the general infinite horizon setting.

Although the main paper expresses the main claims under an assumption of deterministic dynamics, this assumption is not necessary: we restricted attention to the deterministic case in the main body merely for clarity and ease of explanation. Given that in this appendix we provide the general foundations for this work, we consider the more general stochastic setting throughout the later sections.

In particular, for the general stochastic, infinite horizon setting, we introduce and discuss the entropy regularized expected return O_{ENT} and define a "softmax" operator $\mathcal{B}^*$ (analogous to the Bellman operator for hard-max Q-values). We then show the existence of a unique fixed point V^* of $\mathcal{B}^*$, by establishing that the softmax Bellman operator ($\mathcal{B}^*$) is a contraction under the infinity norm. We then relate V^* to the optimal value of the entropy regularized expected reward objective O_{ENT}, which we term $V^\dagger$. We are able to show that $V^* = V^\dagger$, as expected. Subsequently, we present a policy determined by V^* that satisfies $V^*(s) = O_{\text{ENT}}(s, \pi^*)$. Then given the characterization of π^* in terms of V^*, we establish the consistency property stated in Theorem 1 of the main text. Finally, we show that a consistent solution is optimal by satisfying the KKT conditions of the constrained optimization problem (establishing Theorem 4 of the main text).

C.1 Basic results for one-shot entropy regularized optimization

For $\tau > 0$ and any vector $\mathbf{q} \in \mathbb{R}^n$, $n < \infty$, define the scalar valued function F_τ (the "softmax") by

$$F_\tau(\mathbf{q}) = \tau \log \left(\sum_{a=1}^{n} e^{q_a/\tau} \right) \tag{1}$$

and define the vector valued function $\mathbf{f}_\tau$ (the "soft indmax") by

$$\mathbf{f}_\tau(\mathbf{q}) = \frac{e^{\mathbf{q}/\tau}}{\sum_{a=1}^{n} e^{q_a/\tau}} = e^{(\mathbf{q} - F_\tau(\mathbf{q}))/\tau}, \tag{2}$$

where the exponentiation is component-wise. It is easy to verify that $\mathbf{f}_\tau = \nabla F_\tau$. Note that $\mathbf{f}_\tau$ maps any real valued vector to a probability vector. We denote the probability simplex by $\Delta = \{\boldsymbol{\pi} : \boldsymbol{\pi} \geq 0, \mathbf{1} \cdot \boldsymbol{\pi} = 1\}$, and denote the entropy function by $H(\boldsymbol{\pi}) = -\boldsymbol{\pi} \cdot \log \boldsymbol{\pi}$.

Lemma 1.

$$\begin{aligned} F_\tau(\mathbf{q}) &= \max_{\boldsymbol{\pi} \in \Delta} \Big\{ \boldsymbol{\pi} \cdot \mathbf{q} + \tau H(\boldsymbol{\pi}) \Big\} \qquad (3) \\ &= \mathbf{f}_\tau(\mathbf{q}) \cdot \mathbf{q} + \tau H(\mathbf{f}_\tau(\mathbf{q})) \qquad (4) \end{aligned}$$

Proof. First consider the constrained optimization problem on the right hand side of (3). The Lagrangian is given by $L = \boldsymbol{\pi} \cdot (\mathbf{q} - \tau \log \boldsymbol{\pi}) + \lambda(1 - \mathbf{1} \cdot \boldsymbol{\pi})$, hence $\nabla L = \mathbf{q} - \tau \log \boldsymbol{\pi} - \tau - \lambda$. The KKT conditions for this optimization problems are the following system of $n + 1$ equations

$$\begin{aligned} \mathbf{1} \cdot \boldsymbol{\pi} &= 1 \qquad (5) \\ \tau \log \boldsymbol{\pi} &= \mathbf{q} - v \qquad (6) \end{aligned}$$

for the $n + 1$ unknowns, $\boldsymbol{\pi}$ and v, where $v = \lambda + \tau$. Note that for any v, satisfying (6) requires the unique assignment $\boldsymbol{\pi} = \exp((\mathbf{q} - v)/\tau)$, which also ensures $\boldsymbol{\pi} > 0$. To subsequently satisfy (5), the equation $1 = \sum_a \exp((q_a - v)/\tau) = e^{-v/\tau} \sum_a \exp(q_a/\tau)$ must be solved for v; since the right hand side is strictly decreasing in v, the solution is also unique and in this case given by $v = F_\tau(\mathbf{q})$. Therefore $\boldsymbol{\pi} = \mathbf{f}_\tau(\mathbf{q})$ and $v = F_\tau(\mathbf{q})$ provide the unique solution to the KKT conditions (5)-(6). Since the objective is strictly concave, $\boldsymbol{\pi}$ must be the unique global maximizer, establishing (4). It is then easy to show $F_\tau(\mathbf{q}) = \mathbf{f}_\tau(\mathbf{q}) \cdot \mathbf{q} + \tau H(\mathbf{f}_\tau(\mathbf{q}))$ by algebraic manipulation, which establishes (3). □

Corollary 2 (Optimality Implies Consistency). *If* $v^* = \max_{\boldsymbol{\pi} \in \Delta} \Big\{ \boldsymbol{\pi} \cdot \mathbf{q} + \tau H(\boldsymbol{\pi}) \Big\}$ *then*

$$v^* = q_a - \tau \log \pi_a^* \;\; \textit{for all } a, \tag{7}$$

where $\boldsymbol{\pi}^* = \mathbf{f}_\tau(\mathbf{q})$.

Proof. From Lemma 1 we know $v^* = F_\tau(\mathbf{q}) = \boldsymbol{\pi}^* \cdot (\mathbf{q} - \tau \log \boldsymbol{\pi}^*)$ where $\boldsymbol{\pi}^* = \mathbf{f}_\tau(\mathbf{q})$. From the definition of $\mathbf{f}_\tau$ it also follows that $\log \pi_a^* = (q_a - F_\tau(\mathbf{q}))/\tau$ for all a, hence $v^* = F_\tau(\mathbf{q}) = q_a - \tau \log \pi_a^*$ for all a. □

Corollary 3 (Consistency Implies Optimality)**.** *If $v \in \mathbb{R}$ and $\boldsymbol{\pi} \in \Delta$ jointly satisfy*

$$v = q_a - \tau \log \pi_a \;\; \textit{for all } a, \tag{8}$$

then $v = F_\tau(\mathbf{q})$ and $\boldsymbol{\pi} = \mathbf{f}_\tau(\mathbf{q})$; that is, $\boldsymbol{\pi}$ must be an optimizer for (3) *and v is its corresponding optimal value.*

Proof. Any v and $\boldsymbol{\pi} \in \Delta$ that jointly satisfy (8) must also satisfy the KKT conditions (5)-(6); hence $\boldsymbol{\pi}$ must be the unique maximizer for (3) and v its corresponding objective value. □

Although these results are elementary, they reveal a strong connection between optimal state values (v), optimal action values ($\mathbf{q}$) and optimal policies ($\boldsymbol{\pi}$) under the softmax operators. In particular, Lemma 1 states that, if $\mathbf{q}$ is an optimal action value at some current state, the optimal state value must be $v = F_\tau(\mathbf{q})$, which is simply the entropy regularized value of the optimal policy, $\boldsymbol{\pi} = \mathbf{f}_\tau(\mathbf{q})$, at the current state.

Corollaries 2 and 3 then make the stronger observation that this mutual consistency between the optimal state value, optimal action values and optimal policy probabilities must hold for *every* action, not just in expectation over actions sampled from $\boldsymbol{\pi}$; and furthermore that achieving mutual consistency in this form is *equivalent* to achieving optimality.

Below we will also need to make use of the following properties of F_τ.

Lemma 4. *For any vector* $\mathbf{q}$*,*

$$F_\tau(\mathbf{q}) = \sup_{\mathbf{p} \in \Delta} \Big\{ \mathbf{p} \cdot \mathbf{q} - \tau \mathbf{p} \cdot \log \mathbf{p} \Big\}. \tag{9}$$

Proof. Let F_τ^* denote the conjugate of F_τ, which is given by

$$F_\tau^*(\mathbf{p}) = \sup_{\mathbf{q}} \Big\{ \mathbf{q} \cdot \mathbf{p} - F_\tau(\mathbf{q}) \Big\} = \tau \mathbf{p} \cdot \log \mathbf{p} \tag{10}$$

for $\mathbf{p} \in \mathrm{dom}(F_\tau^*) = \Delta$. Since F_τ is closed and convex, we also have that $F_\tau = F_\tau^{**}$ [3, Section 4.2]; hence

$$F_\tau(\mathbf{q}) = \sup_{\mathbf{p} \in \Delta} \Big\{ \mathbf{q} \cdot \mathbf{p} - F_\tau^*(\mathbf{p}) \Big\}. \tag{11}$$

□

Lemma 5. *For any two vectors $\mathbf{q}^{(1)}$ and $\mathbf{q}^{(2)}$,*

$$F_\tau(\mathbf{q}^{(1)}) - F_\tau(\mathbf{q}^{(2)}) \le \max_a \Big\{ q_a^{(1)} - q_a^{(2)} \Big\}. \tag{12}$$

Proof. Observe that by Lemma 4

$$F_\tau(\mathbf{q}^{(1)}) - F_\tau(\mathbf{q}^{(2)}) = \sup_{\mathbf{p}^{(1)} \in \Delta} \Big\{ \mathbf{q}^{(1)} \cdot \mathbf{p}^{(1)} - F_\tau^*(\mathbf{p}^{(1)}) \Big\} - \sup_{\mathbf{p}^{(2)} \in \Delta} \Big\{ \mathbf{q}^{(2)} \cdot \mathbf{p}^{(2)} - F_\tau^*(\mathbf{p}^{(2)}) \Big\} \tag{13}$$

$$= \sup_{\mathbf{p}^{(1)} \in \Delta} \Big\{ \inf_{\mathbf{p}^{(2)} \in \Delta} \Big\{ \mathbf{q}^{(1)} \cdot \mathbf{p}^{(1)} - \mathbf{q}^{(2)} \cdot \mathbf{p}^{(2)} - (F_\tau^*(\mathbf{p}^{(1)}) - F_\tau^*(\mathbf{p}^{(2)})) \Big\} \Big\} \tag{14}$$

$$\le \sup_{\mathbf{p}^{(1)} \in \Delta} \Big\{ \mathbf{p}^{(1)} \cdot (\mathbf{q}^{(1)} - \mathbf{q}^{(2)}) \Big\} \quad \text{by choosing } \mathbf{p}^{(2)} = \mathbf{p}^{(1)} \tag{15}$$

$$\le \max_a \Big\{ q_a^{(1)} - q_a^{(2)} \Big\}. \tag{16}$$

□

Corollary 6. *F_τ is an ∞-norm contraction; that is, for any two vectors $\mathbf{q}^{(1)}$ and $\mathbf{q}^{(2)}$,*

$$\Big| F_\tau(\mathbf{q}^{(1)}) - F_\tau(\mathbf{q}^{(2)}) \Big| \le \| \mathbf{q}^{(1)} - \mathbf{q}^{(2)} \|_\infty \tag{17}$$

Proof. Immediate from Lemma 5. □

C.2 Background results for *on-policy* entropy regularized updates

Although the results in the main body of the paper are expressed in terms of deterministic problems, we will prove that all the desired properties hold for the more general **stochastic** case, where there is a stochastic transition $s, a \mapsto s'$ determined by the environment. Given the characterization for this general case, the application to the deterministic case is immediate. We continue to assume that the action space is finite, and that the state space is discrete.

For any policy π, define the entropy regularized expected return by

$$\tilde{V}^\pi(s_\ell) = O_{\text{ENT}}(s_\ell, \pi) = \mathbb{E}_{a_\ell s_{\ell+1} \ldots | s_\ell}\left[\sum_{i=0}^{\infty} \gamma^i \big(r(s_{\ell+i}, a_{\ell+i}) - \tau \log \pi(a_{\ell+i}|s_{\ell+i})\big)\right], \tag{18}$$

where the expectation is taken with respect to the policy π and with respect to the stochastic state transitions determined by the environment. We will find it convenient to also work with the on-policy Bellman operator defined by

$$(\mathcal{B}^\pi V)(s) = \mathbb{E}_{a,s'|s}\Big[r(s,a) - \tau \log \pi(a|s) + \gamma V(s')\Big] \tag{19}$$

$$= \mathbb{E}_{a|s}\Big[r(s,a) - \tau \log \pi(a|s) + \gamma \mathbb{E}_{s'|s,a}\big[V(s')\big]\Big] \tag{20}$$

$$= \pi(:|s) \cdot (Q(s,:) - \tau \log \pi(:|s)), \quad \text{where} \tag{21}$$

$$Q(s,a) = r(s,a) + \gamma \mathbb{E}_{s'|s,a}[V(s')] \tag{22}$$

for each state s and action a. Note that in (21) we are using $Q(s,:)$ to denote a vector values over choices of a for a given s, and $\pi(:|s)$ to denote the vector of conditional action probabilities specified by π at state s.

Lemma 7. *For any policy π and state s, $\tilde{V}^\pi(s)$ satisfies the recurrence*

$$\tilde{V}^\pi(s) = \mathbb{E}_{a|s}\Big[r(s,a) + \gamma \mathbb{E}_{s'|s,a}[\tilde{V}^\pi(s')] - \tau \log \pi(a|s)\Big] \tag{23}$$

$$= \pi(:|s) \cdot \Big(\tilde{Q}^\pi(s,:) - \tau \log \pi(:|s)\Big) \;\textit{where}\; \tilde{Q}^\pi(s,a) = r(s,a) + \gamma \mathbb{E}_{s'|s,a}[\tilde{V}^\pi(s')] \tag{24}$$

$$= (\mathcal{B}^\pi \tilde{V}^\pi)(s). \tag{25}$$

Moreover, $\mathcal{B}^\pi$ is a contraction mapping.

Proof. Consider an arbitrary state s_ℓ. By the definition of $\tilde{V}^\pi(s_\ell)$ in (18) we have

$$\tilde{V}^\pi(s_\ell) = \mathbb{E}_{a_\ell s_{\ell+1} \ldots | s_\ell}\left[\sum_{i=0}^{\infty} \gamma^i \big(r(s_{\ell+i}, a_{\ell+i}) - \tau \log \pi(a_{\ell+i}|s_{\ell+i})\big)\right] \tag{26}$$

$$= \mathbb{E}_{a_\ell s_{\ell+1} \ldots | s_\ell}\Bigg[r(s_\ell, a_\ell) - \tau \log \pi(a_\ell|s_\ell) \tag{27}$$

$$+ \gamma \sum_{j=0}^{\infty} \gamma^j \big(r(s_{\ell+1+j}, a_{\ell+1+j}) - \tau \log \pi(a_{\ell+1+j}|s_{\ell+1+j})\big)\Bigg]$$

$$= \mathbb{E}_{a_\ell|s_\ell}\Bigg[r(s_\ell, a_\ell) - \tau \log \pi(a_\ell|s_\ell) \tag{28}$$

$$+ \gamma \mathbb{E}_{s_{\ell+1} a_{\ell+1} \ldots | s_\ell, a_\ell}\left[\sum_{j=0}^{\infty} \gamma^j \big(r(s_{\ell+1+j}, a_{\ell+1+j}) - \tau \log \pi(a_{\ell+1+j}|s_{\ell+1+j})\big)\right]\Bigg]$$

$$= \mathbb{E}_{a_\ell|s_\ell}\Big[r(s_\ell, a_\ell) - \tau \log \pi(a_\ell|s_\ell) + \gamma \mathbb{E}_{s_{\ell+1}|s_\ell, a_\ell}[\tilde{V}^\pi(s_{\ell+1})]\Big] \tag{29}$$

$$= \pi(:|s_\ell) \cdot \Big(\tilde{Q}^\pi(s_\ell,:) - \tau \log \pi(:|s_\ell)\Big) \tag{30}$$

$$= (\mathcal{B}^\pi \tilde{V}^\pi)(s_\ell). \tag{31}$$

The fact that $\mathcal{B}^\pi$ is a contraction mapping follows directly from standard arguments about the on-policy Bellman operator [7]. □

Note that this lemma shows $\tilde{V}^\pi$ is a *fixed point* of the corresponding on-policy Bellman operator $\mathcal{B}^\pi$. Next, we characterize how quickly convergence to a fixed point is achieved by repeated application of ther $\mathcal{B}^\pi$ operator.

Lemma 8. *For any π and any V, for all states s_ℓ, and for all $k \geq 0$ it holds that:* $\big((\mathcal{B}^\pi)^k V\big)(s_\ell) - \tilde{V}^\pi(s_\ell) = \gamma^k \mathbb{E}_{a_\ell s_{\ell+1} \ldots s_{\ell+k}|s_\ell}\big[V(s_{\ell+k}) - \tilde{V}^\pi(s_{\ell+k})\big]$.

Proof. Consider an arbitrary state s_ℓ. We use an induction on k. For the base case, consider $k = 0$ and observe that the claim follows trivially, since $\big((\mathcal{B}^\pi)^0 V\big)(s_\ell) - \big((\mathcal{B}^\pi)^0 \tilde{V}^\pi\big)(s_\ell) = V(s_\ell) - \tilde{V}^\pi(s_\ell)$. For the induction hypothesis, assume the result holds for k. Then consider:

$$\big((\mathcal{B}^\pi)^{k+1} V\big)(s_\ell) - \big(\tilde{V}^\pi\big)(s_\ell)$$

$$= \big((\mathcal{B}^\pi)^{k+1} V\big)(s_\ell) - \big((\mathcal{B}^\pi)^{k+1} \tilde{V}^\pi\big)(s_\ell) \quad \text{(by Lemma 7)} \tag{32}$$

$$= \big(\mathcal{B}^\pi(\mathcal{B}^\pi)^k V\big)(s_\ell) - \big(\mathcal{B}^\pi(\mathcal{B}^\pi)^k \tilde{V}^\pi\big)(s_\ell) \tag{33}$$

$$\begin{aligned} &= \mathbb{E}_{a_\ell s_{\ell+1}|s_\ell}\Big[r(s_\ell, a_\ell) - \tau \log \pi(a_\ell|s_\ell) + \gamma(\mathcal{B}^\pi)^k V(s_{\ell+1})\Big] \\ &\quad - \mathbb{E}_{a_\ell s_{\ell+1}|s_\ell}\Big[r(s_\ell, a_\ell) - \tau \log \pi(a_\ell|s_\ell) + \gamma(\mathcal{B}^\pi)^k \tilde{V}^\pi(s_{\ell+1})\Big] \end{aligned} \tag{34}$$

$$= \gamma \mathbb{E}_{a_\ell s_{\ell+1}|s_\ell}\Big[(\mathcal{B}^\pi)^k V(s_{\ell+1}) - (\mathcal{B}^\pi)^k \tilde{V}^\pi(s_{\ell+1})\Big] \tag{35}$$

$$= \gamma \mathbb{E}_{a_\ell s_{\ell+1}|s_\ell}\Big[(\mathcal{B}^\pi)^k V(s_{\ell+1}) - \tilde{V}^\pi(s_{\ell+1})\Big] \quad \text{(by Lemma 7)} \tag{36}$$

$$= \gamma \mathbb{E}_{a_\ell s_{\ell+1}|s_\ell}\Big[\gamma^k \mathbb{E}_{a_{\ell+1} s_{\ell+2} \ldots s_{\ell+k+1}|s_{\ell+1}}\big[V(s_{\ell+k+1}) - \tilde{V}^\pi(s_{\ell+k+1})\big]\Big] \quad \text{(by IH)} \tag{37}$$

$$= \gamma^{k+1} \mathbb{E}_{a_\ell s_{\ell+1} \ldots s_{\ell+k+1}|s_\ell}\Big[V(s_{\ell+k+1}) - \tilde{V}^\pi(s_{\ell+k+1})\Big], \tag{38}$$

establishing the claim. □

Lemma 9. *For any π and any V, we have $\big\|(\mathcal{B}^\pi)^k V - \tilde{V}^\pi\big\|_\infty \leq \gamma^k \big\|V - \tilde{V}^\pi\big\|_\infty$.*

Proof. Let $p^{(k)}(s_{\ell+k}|s_\ell)$ denote the conditional distribution over the kth state, $s_{\ell+k}$, visited in a random walk starting from s_ℓ, which is induced by the environment and the policy π. Consider

$$\big\|(\mathcal{B}^\pi)^k V - \tilde{V}^\pi\big\|_\infty = \gamma^k \max_{s_\ell} \Big|\mathbb{E}_{a_\ell s_{\ell+1} \ldots s_{\ell+k}|s_\ell}\big[V(s_{\ell+k}) - \tilde{V}^\pi(s_{\ell+k})\big]\Big| \quad \text{(by Lemma 8)} \tag{39}$$

$$= \gamma^k \max_{s_\ell} \Big|\sum_{s_{\ell+k}} p^{(k)}(s_{\ell+k}|s_\ell)\big(V(s_{\ell+k}) - \tilde{V}^\pi(s_{\ell+k})\big)\Big| \tag{40}$$

$$= \gamma^k \max_{s_\ell} \Big|p^{(k)}(:|s_\ell) \cdot \big(V - \tilde{V}^\pi\big)\Big| \tag{41}$$

$$\leq \gamma^k \max_{s_\ell} \|p^{(k)}(:|s_\ell)\|_1 \; \|V - \tilde{V}^\pi\|_\infty \quad \text{(by Hölder's inequality)} \tag{42}$$

$$= \gamma^k \|V - \tilde{V}^\pi\|_\infty. \tag{43}$$

□

Corollary 10. *For any bounded V and any $\epsilon > 0$ there exists a k_0 such that $(\mathcal{B}^\pi)^k V \geq \tilde{V}^\pi - \epsilon$ for all $k \geq k_0$.*

Proof. By Lemma 9 we have $(\mathcal{B}^\pi)^k V \geq \tilde{V}^\pi - \gamma^k \big\|V - \tilde{V}^\pi\big\|_\infty$ for all $k \geq 0$. Therefore, for any $\epsilon > 0$ there exists a k_0 such that $\gamma^k \big\|V - \tilde{V}^\pi\big\|_\infty < \epsilon$ for all $k \geq k_0$, since V is assumed bounded. □

Thus, any value function will converge to $\tilde{V}^\pi$ via repeated application of on-policy backups $\mathcal{B}^\pi$. Below we will also need to make use of the following monotonicity property of the on-policy Bellman operator.

Lemma 11. *For any π, if $V^{(1)} \geq V^{(2)}$ then $\mathcal{B}^\pi V^{(1)} \geq \mathcal{B}^\pi V^{(2)}$.*

Proof. Assume $V^{(1)} \geq V^{(2)}$ and note that for any state s_ℓ

$$\begin{aligned} (\mathcal{B}^\pi V^{(2)})(s_\ell) - (\mathcal{B}^\pi V^{(1)})(s_\ell) &= \gamma \mathbb{E}_{a_\ell s_{\ell+1}|s_\ell}\big[V^{(2)}(s_{\ell+1}) - V^{(1)}(s_{\ell+1})\big] && (44) \\ &\leq 0 \quad \text{since it was assumed that } V^{(2)} \leq V^{(1)}. && (45) \end{aligned}$$

□

C.3 Proof of main optimality claims for *off-policy* softmax updates

Define the optimal value function by

$$V^\dagger(s) = \max_\pi O_{\text{ENT}}(s, \pi) = \max_\pi \tilde{V}^\pi(s) \text{ for all } s. \tag{46}$$

For $\tau > 0$, define the softmax Bellman operator $\mathcal{B}^*$ by

$$\begin{aligned} (\mathcal{B}^* V)(s) &= \tau \log \sum_a \exp\Big(\big(r(s,a) + \gamma \mathbb{E}_{s'|s,a}[V(s')]\big)/\tau\Big) && (47) \\ &= F_\tau(Q(s,:)) \quad \text{where} \quad Q(s,a) = r(s,a) + \gamma \mathbb{E}_{s'|s,a}[V(s')] \quad \text{for all } a. && (48) \end{aligned}$$

Lemma 12. *For $\gamma < 1$, the fixed point of the softmax Bellman operator, $V^* = \mathcal{B}^* V^*$, exists and is unique.*

Proof. First observe that the softmax Bellman operator is a contraction in the infinity norm. That is, consider two value functions, $V^{(1)}$ and $V^{(2)}$, and let $p(s'|s,a)$ denote the state transition probability function determined by the environment. We then have

$$\begin{aligned} \Big\|\mathcal{B}^* V^{(1)} - \mathcal{B}^* V^{(2)}\Big\|_\infty &= \max_s \Big|(\mathcal{B}^* V^{(1)})(s) - (\mathcal{B}^* V^{(2)})(s)\Big| && (49) \\ &= \max_s \Big|F_\tau\big(Q^{(1)}(s,:)\big) - F_\tau\big(Q^{(2)}(s,:)\big)\Big| && (50) \\ &\leq \max_s \max_a \Big|Q^{(1)}(s,a) - Q^{(2)}(s,a)\Big| \quad \text{(by Corollary 6)} && (51) \\ &= \gamma \max_s \max_a \Big|\mathbb{E}_{s'|s,a}\big[V^{(1)}(s') - V^{(2)}(s')\big]\Big| && (52) \\ &= \gamma \max_s \max_a \Big|p(:|s,a) \cdot \big(V^{(1)} - V^{(2)}\big)\Big| && (53) \\ &\leq \gamma \max_s \max_a \|p(:|s,a)\|_1 \; \|V^{(1)} - V^{(2)}\|_\infty \quad \text{(Hölder's inequality)} && (54) \\ &= \gamma \|V^{(1)} - V^{(2)}\|_\infty < \|V^{(1)} - V^{(2)}\|_\infty \text{ if } \gamma < 1. && (55) \end{aligned}$$

The existence and uniqueness of V^* then follows from the contraction map fixed-point theorem [2]. □

Lemma 13. *For any π, if $V \geq \mathcal{B}^* V$ then $V \geq (\mathcal{B}^\pi)^k V$ for all k.*

Proof. Observe for any s that the assumption implies

$$\begin{aligned} V(s) &\geq (\mathcal{B}^* V)(s) && (56) \\ &= \max_{\tilde{\pi}(:|s) \in \Delta} \sum_a \tilde{\pi}(a|s)\Big(r(s,a) + \gamma \mathbb{E}_{s'|s,a}[V(s')] - \tau \log \tilde{\pi}(a|s)\Big) && (57) \\ &\geq \sum_a \pi(a|s)\Big(r(s,a) + \gamma \mathbb{E}_{s'|s,a}[V(s')] - \tau \log \pi(a|s)\Big) && (58) \\ &= (\mathcal{B}^\pi V)(s). && (59) \end{aligned}$$

The result then follows by the monotonicity of $\mathcal{B}^\pi$ (Lemma 11). □

Corollary 14. *For any π, if V is bounded and $V \geq \mathcal{B}^* V$, then $V \geq \tilde{V}^\pi$.*

Proof. Consider an arbitrary policy π. If $V \geq \mathcal{B}^* V$, then by Corollary 14 we have $V \geq (\mathcal{B}^\pi)^k V$ for all k. Then by Corollary 10, for any $\epsilon > 0$ there exists a k_0 such that $V \geq (\mathcal{B}^\pi)^k V \geq \tilde{V}^\pi - \epsilon$ for all $k \geq k_0$ since V is bounded; hence $V \geq \tilde{V}^\pi - \epsilon$ for all $\epsilon > 0$. We conclude that $V \geq \tilde{V}^\pi$. □

Next, given the existence of V^*, we define a specific policy π^* as follows

$$\pi^*(:|s) = \mathbf{f}_\tau\big(Q^*(s,:)\big), \quad \text{where} \tag{60}$$

$$Q^*(s,a) = r(s,a) + \gamma\mathbb{E}_{s'|s,a}[V^*(s')]. \tag{61}$$

Note that we are simply defining π^* at this stage and have not as yet proved it has any particular properties; but we will see shortly that it is, in fact, an optimal policy.

Lemma 15. *$V^* = \tilde{V}^{\pi^*}$; that is, for π^* defined in* (60)*, V^* gives its entropy regularized expected return from any state.*

Proof. We establish the claim by showing $\mathcal{B}^*\tilde{V}^{\pi^*} = \tilde{V}^{\pi^*}$. In particular, for an arbitrary state s consider

$$\begin{aligned}
(\mathcal{B}^*\tilde{V}^{\pi^*})(s) &= F_\tau\big(\tilde{Q}^{\pi^*}(s,:)\big) && \text{by (48)} && (62)\\
&= \pi^*(:|s)\cdot\big(\tilde{Q}^{\pi^*}(s,:) - \tau\log\pi^*(:|s)\big) && \text{by Lemma 1} && (63)\\
&= \tilde{V}^{\pi^*}(s) && \text{by Lemma 7.} && (64)
\end{aligned}$$

□

Theorem 16. *The fixed point of the softmax Bellman operator is the optimal value function: $V^* = V^\dagger$.*

Proof. Since $V^* \geq \mathcal{B}^*V^*$ (in fact, $V^* = \mathcal{B}^*V^*$) we have $V^* \geq \tilde{V}^\pi$ for all π by Corollary 14, hence $V^* \geq V^\dagger$. Next observe that by Lemma 15 we have $V^\dagger \geq \tilde{V}^{\pi^*} = V^*$. Finally, by Lemma 12, we know that the fixed point $V^* = \mathcal{B}^*V^*$ is unique, hence $V^\dagger = V^*$. □

Corollary 17 (Optimality Implies Consistency)**.** *The optimal state value function V^* and optimal policy π^* satisfy $V^*(s) = r(s,a) + \gamma\mathbb{E}_{s'|s,a}[V^*(s')] - \tau\log\pi^*(a|s)$ for every state s and action a.*

Proof. First note that

$$\begin{aligned}
Q^*(s,a) &= r(s,a) + \gamma\mathbb{E}_{s'|s,a}[V^*(s')] && \text{by (61)} && (65)\\
&= r(s,a) + \gamma\mathbb{E}_{s'|s,a}[V^{\pi^*}(s')] && \text{by Lemma 15} && (66)\\
&= Q^{\pi^*}(s,a) && \text{by (22).} && (67)
\end{aligned}$$

Then observe that for any state s,

$$\begin{aligned}
V^*(s) &= F_\tau\big(Q^*(s,:)\big) && \text{by (48)} && (68)\\
&= F_\tau\big(Q^{\pi^*}(s,:)\big) && \text{from above} && (69)\\
&= \pi^*(:|s)\cdot\big(Q^{\pi^*}(s,:) - \tau\log\pi^*(:|s)\big) && \text{by Lemma 1} && (70)\\
&= Q^{\pi^*}(s,a) - \tau\log\pi^*(a|s) \text{ for all } a && \text{by Corollary 2} && (71)\\
&= Q^*(s,a) - \tau\log\pi^*(a|s) \text{ for all } a && \text{from above.} && (72)
\end{aligned}$$

□

Corollary 18 (Consistency Implies Optimality)**.** *If V and π satisfy, for all s and a: $V(s) = r(s,a) + \gamma\mathbb{E}_{s'|s,a}[V(s')] - \tau\log\pi(a|s)$; then $V = V^*$ and $\pi = \pi^*$.*

Proof. We will show that satisfying the constraint for every s and a implies $\mathcal{B}^*V = V$; it will then immediately follow that $V = V^*$ and $\pi = \pi^*$ by Lemma 12. Let $Q(s,a) = r(s,a) + \gamma\mathbb{E}_{s'|s,a}[V(s')]$. Consider an arbitrary state s, and observe that

$$\begin{aligned}
(\mathcal{B}^*V)(s) &= F_\tau\big(Q(s,:)\big) \quad \text{(by (48))} && (73)\\
&= \max_{\boldsymbol{\pi}\in\Delta}\Big\{\boldsymbol{\pi}\cdot\big(Q(s,:) - \tau\log\boldsymbol{\pi}\big)\Big\} \quad \text{(by Lemma 1)} && (74)\\
&= Q(s,a) - \tau\log\pi(a|s) \text{ for all } a \quad \text{(by Corollary 3)} && (75)\\
&= r(s,a) + \gamma\mathbb{E}_{s'|s,a}[V(s')] - \tau\log\pi(a|s) \text{ for all } a \quad \text{(by definition of } Q \text{ above)} && (76)\\
&= V(s) \quad \text{(by the consistency assumption on } V \text{ and } \pi\text{).} && (77)
\end{aligned}$$

□

C.4 Proof of Theorem 1 from Main Text

Note: Theorem 1 from the main body was stated under an assumption of deterministic dynamics. We used this assumption in the main body merely to keep presentation simple and understandable. The development given in this appendix considers the more general case of a stochastic environment. We give the proof here for the more general setting; the result stated in Theorem 1 follows as a special case.

Proof. Assuming a stochastic environment, as developed in this appendix, we will establish that the optimal policy and state value function, π^* and V^* respectively, satisfy

$$V^*(s) = -\tau \log \pi^*(a|s) + r(s,a) + \gamma \mathbb{E}_{s'|s,a}[V^*(s')] \tag{78}$$

for all s and a. Theorem 1 will then follow as a special case.

Consider the policy π^* defined in (60). From Corollary 15 we know that $\tilde{V}^{\pi^*} = V^*$ and from Theorem 16 we know $V^* = V^\dagger$, hence $\tilde{V}^{\pi^*} = V^\dagger$; that is, π^* is the optimizer of $O_{\text{ENT}}(s,\pi)$ for any state s (including s_0). Therefore, this must be the same π^* as considered in the premise. The assertion (78) then follows directly from Corollary 17. □

C.5 Proof of Corollary 2 from Main Text

Note: We consider the more general case of a stochastic environment as developed in this appendix. First note that the consistency property for the stochastic case (78) can be rewritten as

$$\mathbb{E}_{s'|s,a}\big[-V^*(s) + \gamma V^*(s') + r(s,a) - \tau \log \pi^*(a|s)\big] = 0 \tag{79}$$

for all s and a. For a stochastic environment, the generalized version of (13) in Corollary 2 can then be expressed as

$$\mathbb{E}_{s_2\ldots s_t|s_1,a_1\ldots a_{t-1}}\left[-V^*(s_1) + \gamma^{t-1}V^*(s_t) + \sum_{i=1}^{t-1}\gamma^{i-1}\big(r(s_i,a_i) - \tau\log\pi^*(a_i|s_i)\big)\right] = 0 \tag{80}$$

for all states s_1 and action sequences $a_1 \ldots a_{t-1}$. We now show that (79) implies (80).

Proof. Observe that by (79) we have

$$0 = \mathbb{E}_{s_2\ldots s_t|s_1,a_1\ldots a_{t-1}}\left[\sum_{i=1}^{t-1}\gamma^{i-1}\Big(-V^*(s_i) + \gamma V^*(s_{i+1}) + r(s_i,a_i) - \tau\log\pi^*(a_i|s_i)\Big)\right] \tag{81}$$

$$= \mathbb{E}_{s_2\ldots s_t|s_1,a_1\ldots a_{t-1}}\left[\sum_{i=1}^{t-1}\gamma^{i-1}\Big(-V^*(s_i) + \gamma V^*(s_{i+1})\Big) + \sum_{i=1}^{t-1}\gamma^{i-1}\Big(r(s_i,a_i) - \tau\log\pi^*(a_i|s_i)\Big)\right] \tag{82}$$

$$= \mathbb{E}_{s_2\ldots s_t|s_1,a_1\ldots a_{t-1}}\left[-V^*(s_1) + \gamma^{t-1}V^*(s_t) + \sum_{i=1}^{t-1}\gamma^{i-1}\Big(r(s_i,a_i) - \tau\log\pi^*(a_i|s_i)\Big)\right] \tag{83}$$

by a telescopic sum on the first term, which yields the result. □

C.6 Proof of Theorem 3 from Main Text

Note: Again, we consider the more general case of a stochastic environment. The consistency property in this setting is given by (78) above.

Proof. Consider a policy π_θ and value function V_ϕ that satisfy the general consistency property for a stochastic environment: $V_\phi(s) = -\tau\log\pi_\theta(a|s) + r(s,a) + \gamma\mathbb{E}_{s'|s,a}[V_\phi(s')]$ for all s and a. Then by Corollary 18, we must have $V_\phi = V^*$ and $\pi_\theta = \pi^*$. Theorem 3 follows as a special case when the environment is deterministic. □

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. *arXiv:1605.08695*, 2016.

[2] D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 2. Athena Scientific, 1995.

[3] J. Borwein and A. Lewis. *Convex Analysis and Nonlinear Optimization*. Springer, 2000.

[4] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. OpenAI Gym. arXiv:1606.01540, 2016.

[5] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 1997.

[6] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[7] J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5), 1997.