A Extended dynamic programming: technical details

The extended dynamic programming algorithm is given by Algorithm 2.

**Algorithm 2** Extended dynamic programming for finding an optimistic policy and transition model for a given confidence set of transition functions and given rewards.

**Input:** empirical estimate $P$ of transition functions, $L_1$ bound $b \in (0,1]^{|X||A|}$, reward function $r \in [0,1]^{|X||A|}$.

**Initialization:** Set $w(x_L) = 0$.

For $l = L - 1, L - 2, \ldots, 0$

1. Let $k = |X_{l+1}|$ and $(x_1^*, x_2^*, \ldots, x_k^*)$ be a sorting of the states in $X_{l+1}$ such that $w(x_1^*) \geq w(x_2^*) \geq \cdots \geq w(x_k^*)$.

2. **For all** $(x, a) \in X_l \times A$
   
   (a) $P^*(x_1^*|x, a) = \min \left\{ \hat{P}(x_1^*|x, a) + b(x, a)/2, 1 \right\}$
   
   (b) $P^*(x_i^*|x, a) = \hat{P}(x_i^*|x, a)$ for all $i = 2, 3, \ldots, k$.
   
   (c) Set $j = k$.
   
   (d) **While** $\sum_i P^*(x_i^*|x, a) > 1$ **do**
       
       i. Set $P^*(x_j^*|x, a) = \max \left\{ 0, 1 - \sum_{i \neq j} P^*(x_i^*|x, a) \right\}$
       
       ii. Set $j = j - 1$.

3. **For all** $x \in X_l$

   (a) Let $w(x) = \max_a \{ r(x, a) + \sum_{x'} P^*(x'|x, a)w(x') \}$.
   
   (b) Let $\pi^*(x) = \arg \max_a \{ r(x, a) + \sum_{x'} P^*(x'|x, a)w(x') \}$.

**Return:** optimistic transition function $P^*$, optimistic policy $\pi^*$.

The next lemma, which can be obtained by a straightforward modification of the proof of Theorem 7 of Jaksch et al. (2010), shows that Algorithm 2 efficiently solves the desired minimization problem.

**Lemma 6.** Algorithm 2 solves the maximization problem (5) for $\mathcal{P} = \{ \hat{P} : \| \hat{P} - \hat{P} \|_1 \leq b \}$. Let $S = \sum_{l=1}^{L-1} |X_l||X_{l+1}|$ denote the maximum number of possible transitions in the given model. The time and space complexity of Algorithm 2 is the number of possible non-zero elements of $\hat{P}$ allowed by the given structure, and so it is $O(S|A|)$, which, in turn, is $O(|A||X|^2)$.

B The detailed bound

Theorem 1 is a simplified version of the following, more detailed statement.

**Theorem 2.** Assume $\eta \leq (|X||A|)^{-1}$ and $T \geq |X||A|$. Then the expected regret of FPOP can be bounded as

$$V_T^* - \mathbb{E} \left[ \sum_{t=1}^T v_t(\pi_t) \right] \leq L|X||A| \log_2 \left( \frac{8T}{|X||A|} \right) \ln \frac{|X||A|}{L} + \frac{1}{\eta} + \frac{\ln T}{\delta} + L|X| \ln \frac{T|X||A|}{\delta L} + L|X| \sqrt{2T \ln \frac{L}{\delta} + 3\delta T L}.$$

In particular, assuming $T \geq (|X||A|)^2$, setting

$$\eta = \sqrt{\log_2 \left( \frac{8T}{|X||A|} \right) \ln \frac{|X||A|}{T(e-1)} + 1},$$

$$\delta = \frac{1}{\sqrt{T} \ln \frac{T|X||A|}{\delta L}}.$$

The adversarial stochastic shortest path problem with unknown transition probabilities
and \( \delta = 1/T \) gives
\[
V_T^* - \mathbb{E} \left[ \sum_{t=1}^{T} v_t(\pi_t) \right] \leq 2L|X||A|\sqrt{T(e-1) \log_2 \left( \frac{8T}{|X||A|} \right) \left( \ln \left( \frac{|X||A|}{L} \right) + 1 \right)} + \left( \sqrt{2} + 1 \right) L|X|\sqrt{T|A| \ln \frac{T^2|X||A|}{L}} + L|X|\sqrt{2T \ln(2T) + 3L}.
\]

The theorem can be obtained by a trivial combination of Lemmas 2, 3, and 5. The only complication is that in the last term of Lemma 2 we apply the bound
\[
\sum_{l=0}^{L-1} \ln(|X_l||A|) \leq L \ln \left( \frac{|X||A|}{L} \right).
\]

### C Proof of Lemma 1

Let us fix an arbitrary \( x \in \mathcal{X} \) and let \( l = l_x \). The statement follows from the following inequality due to Weissman et al. (2003) concerning the distance of a true discrete distribution \( p \) and the empirical distribution \( \hat{p} \) over \( m \) distinct events from \( n \) samples:
\[
\mathbb{P} \left( \|p - \hat{p}\|_1 \geq \varepsilon \right) \leq (2^m - 2) \exp \left( -\frac{n\varepsilon^2}{2} \right).
\]

As now we have \( |X_{l+1}| \) distinct events, we get that setting
\[
\varepsilon = \sqrt{\frac{4|X_{l+1}| \ln \frac{T|X||A|}{\delta}}{n}}
\]
for some fixed \( n \in [1, 2, \ldots, t] \) yields
\[
\mathbb{P} \left( \left\| \mathbb{P}_i(\cdot|x,a) - P(\cdot|x,a) \right\|_1 \geq \sqrt{\frac{2|X_{l+1}| \ln \frac{T|X||A|}{\delta}}{n}} \left| N_i(x,a) = n \right. \right) \leq \frac{\delta}{T^2|X||A|}.
\]

Using the union bound for all possible values of \( N_i(x,a) \), all \( (x,a) \in \mathcal{X} \times \mathcal{A} \), all \( i = 1, 2, \ldots, K_T \) (note that for the bound, we have used the very crude upper bound \( T > K_T \) for simplicity) and the fact that the confidence intervals trivially hold when there are no observations with probability 1, we get the statement of the lemma.

### D Proof of Lemma 3

Let
\[
(\sigma_t(Y), \Gamma_t(Y)) = \arg \max_{\pi \in \Pi, \bar{P} \in P_t(i)} \left\{ W(R_{t-1} + Y, \pi, \bar{P}) \right\}
\]
and
\[
F_t(Y) = W(r_t, \sigma_t(Y), \Gamma_t(Y)).
\]
Clearly,
\[
\tilde{v}_t = F_t(Y_{i(t)})
\]
and
\[
\tilde{v}_t = F_t(Y_{i(t)} + r_t).
\]

Now let \( f \) be the density function of \( Y_{i(t)} \) and \( F_{i(t)} \) denote the \( \sigma \)-algebra generated by all random variables before epoch \( E_{i(t)} \).

\footnote{Note that \( Y_{i(t)} \) is generated independently from the history up to epoch \( i(t) \).}

We have
\[
\mathbb{E} \left[ \tilde{v}_t | F_{i(t-1)} \right] = \int_{\mathcal{R}^2} F_t(y + r_t) f(y) dy = \int_{\mathcal{R}^2} F_t(y) f(y - r_t) dy \leq \sup_{y,t} \frac{f(y - r_t)}{f(y)} \int_{\mathcal{R}^2} F_t(y) f(y) dy \leq \sup_{y,t} \frac{f(y - r_t)}{f(y)} \mathbb{E} \left[ \tilde{v}_t | F_{i(t-1)} \right].
\]
Since \( f(y) = \eta \exp \left( -\eta \sum x,a y(x, a) \right) \) for all \( y \geq 0 \), we get
\[
\sup_y \frac{f(y - r_l)}{f(y)} = \exp \left( \eta \sum x,a r_l(x, a) \right) \leq \exp (\eta |\mathcal{X}| |\mathcal{A}|).
\]
Using \( e^x \leq 1 + (e - 1)x \) for \( x \in [0, 1] \), which holds by our assumption on \( \eta \), we get
\[
\mathbb{E} [\tilde{v}_t] \leq \mathbb{E} [\tilde{v}_t] (1 + \eta(e - 1)|\mathcal{X}| |\mathcal{A}|).
\]
Noticing that \( \tilde{v}_t \leq L \) gives the result. \( \square \)

E Proof of Lemma 4

We prove the statement by induction on \( l \). For \( l = 1 \) we have
\[
\sum x_1 |\tilde{\mu}_l(x_1) - \mu_l(x_1)| = \sum x_1 |\tilde{P}_l(x_1|x_0, \pi_l(x_0)) - P(x_1|x_0, \pi_l(x_0))| \leq a_t(x_0, \pi_l(x_0)),
\]
proving the statement for this case. Now assume that the statement holds for some \( l - 1 \). We have
\[
\tilde{\mu}_l(x_l) - \mu_l(x_l)
= \sum x_{l-1} \left( \tilde{P}_l(x_l|x_{l-1}, \pi_l(x_{l-1}))\tilde{\mu}_l(x_{l-1}) - P(x_l|x_{l-1}, \pi_l(x_{l-1}))\mu_l(x_{l-1}) \right)
\]
\[
= \sum x_{l-1} \left( \tilde{P}_l(x_l|x_{l-1}, \pi_l(x_{l-1})) (\tilde{\mu}_l(x_{l-1}) - \mu_l(x_{l-1})) + (\tilde{P}_l(x_l|x_{l-1}, \pi_l(x_{l-1})) - P(x_l|x_{l-1}, \pi_l(x_{l-1})) \right) \mu_l(x_{l-1})
\]
and thus
\[
\sum x_l |\tilde{\mu}_l(x_l) - \mu_l(x_l)|
\leq \sum x_{l-1} \left( \tilde{P}_l(x_l|x_{l-1}, \pi_l(x_{l-1}))|\tilde{\mu}_l(x_{l-1}) - \mu_l(x_{l-1})| + |\tilde{P}_l(x_l|x_{l-1}, \pi_l(x_{l-1})) - P(x_l|x_{l-1}, \pi_l(x_{l-1}))| \mu_l(x_{l-1}) \right)
\]
\[
= \sum_{x_{l-1}} |\tilde{\mu}_l(x_{l-1}) - \mu_l(x_{l-1})| + \sum_{x_{l-1}} \mu_l(x_{l-1}) \sum x_l \left| \tilde{P}_l(x_l|x_{l-1}, \pi_l(x_{l-1})) - P(x_l|x_{l-1}, \pi_l(x_{l-1})) \right|
\]
\[
\leq \sum_{k=0}^{l-2} \sum_{x_k \in X_k} \mu_l(x_k) a_t(x_k, \pi_l(x_k)) + \sum_{x_{l-1}} \mu_l(x_{l-1}) \sum_{x_l} a_t(x_{l-1}, \pi_l(x_{l-1}))
\]
proving the statement. \( \square \)

F Proof of Lemma 5

We start by some arguments borrowed from Jaksch et al. (2010). Let \( n_t(x, a) \) be the number of times state-action pair \( (x, a) \) has been visited in epoch \( E_t \). We have
\[
N_t(x, a) = \sum_{j=1}^{t-1} n_t(x, a).
\]
For simplicity, let \( K_T = m \) be the number of epochs. By Appendix C.3 of Jaksch et al. (2010), we have
\[
\sum_{i=1}^{m} \frac{n_i(x, a)}{\sqrt{N_i(x, a)}} \leq \left( \sqrt{2} + 1 \right) \sqrt{N_m(x, a)},
\]
and by Jensen’s inequality,
\[
\sum_{x,a} \frac{m_t(x,a)}{N_t(x,a)} \leq \left(\sqrt{2} + 1\right) \sqrt{|X||A|^T}.
\]

Now fix an arbitrary \(1 \leq t \leq T\). We have
\[
\tilde{v}_t = \sum_{l=0}^{L-1} \sum_{x \in \mathcal{X}_t} \mu_t(x)r_t(x, \pi_t(x))
\]
and
\[
v_t(\pi_t) = \sum_{l=0}^{L-1} \sum_{x \in \mathcal{X}_t} \mu_t(x)r_t(x, \pi_t(x)),
\]
thus
\[
\tilde{v}_t(\pi_t) - v_t(\pi_t) = \sum_{l=0}^{L-1} \sum_{x \in \mathcal{X}_t} (\tilde{\mu}_t(x) - \mu_t(x)) r_t(x, \pi_t(x)) \leq \sum_{l=0}^{L-1} \sum_{x \in \mathcal{X}_t} |\tilde{\mu}_t(x) - \mu_t(x)|.
\]

That is, we need to bound \(\sum_{t=1}^{T} \sum_{x \in \mathcal{X}_t} |\tilde{\mu}_t(x) - \mu_t(x)|\).

Setting \(a_t(x, a) = \left\| \bar{P}_t(\cdot|x, a) - P(\cdot|x, a) \right\|_1\) for all \((x, a) \in \mathcal{X} \times \mathcal{A}\), the conditions of Lemma 4 are clearly satisfied, and so
\[
\sum_{x \in \mathcal{X}_t} |\tilde{\mu}_t(x) - \mu_t(x)| \leq \sum_{k=0}^{t-1} \sum_{x_k \in \mathcal{X}_k} \mu_t(x_k) a_t(x_k, \pi_t(x_k))
\]
\[
\leq \sum_{k=0}^{t-1} a_t(x_k^{(t)}, a_k^{(t)}) + \sum_{k=0}^{t-1} \sum_{x_k \in \mathcal{X}_k} \left(\mu_t(x_k) - \mathbb{1}_{\{x_k^{(t)} = x_k\}}\right) a_1(x_k, \pi_t(x_k)) \tag{9}
\]

Now, by Lemma 1, we have with probability at least \(1 - \delta\) simultaneously for all \(k\) that
\[
\sum_{t=1}^{T} a_t(x_k^{(t)}, a_k^{(t)}) \leq \sum_{t=1}^{T} \frac{2|X_{k+1}| \ln \frac{T|X||A|}{\delta}}{\max \left\{1, N_{i(t)}(x_k^{(t)}, a_k^{(t)})\right\}}
\]
\[
\leq \sum_{x_k,a_k} m_n(x_k,a_k) \frac{2|X_{k+1}| \ln \frac{T|X||A|}{\delta}}{\max \left\{1, N_{i(t)}(x_k,a_k)\right\}}
\]
\[
\leq \left(\sqrt{2} + 1\right) \sqrt{2T|X_k||X_{k+1}||A| \ln \frac{T|X||A|}{\delta}}.
\]

For the second term on the right hand side of (9), notice that \(\left(\mu_t(x_k) - \mathbb{1}_{\{x_k^{(t)} = x_k\}}\right)\) form a martingale difference sequence with respect to \(\{U_t\}_{t=1}^{T}\) and thus by the Hoeffding–Azuma inequality and \(a_1 \leq 2\), we have
\[
\sum_{t=1}^{T} \left(\mu_t(x_k) - \mathbb{1}_{\{x_k^{(t)} = x_k\}}\right) a_1(x_k, \pi_t(x_k)) \leq \sqrt{2T \ln \frac{L}{\delta}}
\]
with probability at least \(1 - \delta/L\). Putting everything together, the union bound implies that we have, with
probability at least $1 - 2\delta$ simultaneously for all $l = 1, \ldots, L$,
\begin{align*}
\sum_{l=1}^{T} \sum_{x \in \mathcal{X}_l} (\tilde{\mu}_l(x) - \mu_l(x)) &\leq \sum_{k=0}^{l-1} \left( \sqrt{2} + 1 \right) \sqrt{T|\mathcal{X}_k||\mathcal{X}_{k+1}||\mathcal{A}| \ln \frac{T|\mathcal{X}| |\mathcal{A}|}{\delta}} + \sum_{k=0}^{l-1} |\mathcal{X}_k| \sqrt{2T \ln \frac{L}{\delta}} \\
&\leq \left( \sqrt{2} + 1 \right) L \sum_{k=0}^{L-1} \frac{1}{L} \sqrt{T|\mathcal{X}_k||\mathcal{X}_{k+1}||\mathcal{A}| \ln \frac{T|\mathcal{X}| |\mathcal{A}|}{\delta}} + \sum_{k=0}^{l-1} |\mathcal{X}_k| \sqrt{2T \ln \frac{L}{\delta}} \\
&\leq \left( \sqrt{2} + 1 \right) L \sqrt{T|\mathcal{A}| \left( \frac{|\mathcal{X}|}{L} \right)^2 \ln \frac{T|\mathcal{X}| |\mathcal{A}|}{\delta}} + |\mathcal{X}| \sqrt{2T \ln \frac{L}{\delta}} \\
&= \left( \sqrt{2} + 1 \right) |\mathcal{X}| \sqrt{T|\mathcal{A}| \ln \frac{T|\mathcal{X}| |\mathcal{A}|}{\delta}} + |\mathcal{X}| \sqrt{2T \ln \frac{L}{\delta}} \\
&\leq \left( \sqrt{2} + 1 \right) |\mathcal{X}| \sqrt{T|\mathcal{A}| |\mathcal{X}| \ln \frac{T|\mathcal{X}| |\mathcal{A}|}{\delta}} + |\mathcal{X}| \sqrt{2T \ln \frac{L}{\delta}}.
\end{align*}

(10)

where in the last step we used Jensen’s inequality for the concave function $f(x, y) = \sqrt{xy(a + \ln x)}$ with parameter $a > 0$ and the fact that $\sum_{k=0}^{L-1} |\mathcal{X}_k| = |\mathcal{X}| - 1 < |\mathcal{X}|$.

Summing up for all $l = 0, 1, \ldots, L - 1$ and taking expectation, using that $v_t(\pi_t) - \tilde{v}_t \leq L$ and (10) holds with probability at least $1 - 2\delta$, finishes the proof. \qed