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 $\hat{P}$ of transition functions, L_1 bound $b \in (0,1]^{|\mathcal{X}||\mathcal{A}|}$, reward function $r \in [0,1]^{|\mathcal{X}||\mathcal{A}|}$.
Initialization: Set $w(x_L) = 0$.
For $l = L-1, L-2, \ldots, 0$

1. Let $k = |\mathcal{X}_{l+1}|$ and $(x_1^*, x_2^*, \ldots, x_k^*)$ be a sorting of the states in $\mathcal{X}_{l+1}$ such that $w(x_1^*) \geq w(x_2^*) \geq \cdots \geq w(x_k^*)$.
2. **For all** $(x,a) \in \mathcal{X}_l \times \mathcal{A}$
   (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** $x \in \mathcal{X}_l$
   (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 π^*.

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 $\mathcal{P} = \{\bar{P} : \|\bar{P} - \hat{P}\|_1 \leq b\}$. Let $S = \sum_{l=0}^{L-1} |\mathcal{X}_l||X_{l+1}|$ 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 $\bar{P}$ allowed by the given structure, and so it is $\mathcal{O}(S|\mathcal{A}|)$, which, in turn, is $\mathcal{O}(|\mathcal{A}||\mathcal{X}|^2)$.*

B The detailed bound

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

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

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

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

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

and $\delta = 1/T$ *gives*

$$V_T^* - \mathbb{E}\left[\sum_{t=1}^{T} v_t(\boldsymbol{\pi}_t)\right] \le 2\,L|\mathcal{X}||\mathcal{A}|\sqrt{T(e-1)\,\log_2\left(\frac{8T}{|\mathcal{X}||\mathcal{A}|}\right)\left(\ln\left(\frac{|\mathcal{X}||\mathcal{A}|}{L}\right)+1\right)}$$
$$+\left(\sqrt{2}+1\right)L|\mathcal{X}|\sqrt{T\,|\mathcal{A}|\ln\frac{T^2|\mathcal{X}||\mathcal{A}|}{L}} + L|\mathcal{X}|\sqrt{2\,T\,\ln(LT)} + 3\,L.$$

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

$$\sum_{l=0}^{L-1} \ln\left(|\mathcal{X}_l||\mathcal{A}|\right) \le L\ln\left(\frac{|\mathcal{X}||\mathcal{A}|}{L}\right).$$

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 $\hat{\mathbf{p}}$ over m distinct events from n samples:

$$\mathbb{P}\left[\|p - \hat{\mathbf{p}}\|_1 \ge \varepsilon\right] \le (2^m - 2)\exp\left(-\frac{n\varepsilon^2}{2}\right).$$

As now we have $|\mathcal{X}_{l+1}|$ distinct events, we get that setting

$$\varepsilon = \sqrt{\frac{4|\mathcal{X}_{l+1}|\ln\frac{T|\mathcal{X}||\mathcal{A}|}{\delta}}{n}}$$

for some fixed $n \in [1, 2, \ldots, t]$ yields

$$\mathbb{P}\left[\left\|\bar{\mathbf{P}}_i(\cdot|x,a) - P(\cdot|x,a)\right\|_1 \ge \sqrt{\frac{2|\mathcal{X}_{l+1}|\ln\frac{T|\mathcal{X}||\mathcal{A}|}{\delta}}{n}}\,\middle|\,\mathbf{N}_i(x,a) = n\right] \le \frac{\delta}{T^2|\mathcal{X}||\mathcal{A}|}.$$

Using the union bound for all possible values of $\mathbf{N}_i(x,a)$, all $(x,a) \in \mathcal{X}\times\mathcal{A}$, all $i = 1, 2, \ldots, \mathbf{K}_T$ (note that for the bound, we have used the very crude upper bound $T > \mathbf{K}_T$ 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

$$(\boldsymbol{\sigma}_t(\mathbf{Y}), \boldsymbol{\Gamma}_t(\mathbf{Y})) = \operatorname*{arg\,max}_{\pi\in\Pi, \bar{P}\in\mathcal{P}_{\mathbf{i}(t)}} \left\{W(R_{t-1} + \mathbf{Y}, \pi, \bar{P})\right\}$$

and

$$\mathbf{F}_t(\mathbf{Y}) = W(r_t, \boldsymbol{\sigma}_t(\mathbf{Y}), \boldsymbol{\Gamma}_t(\mathbf{Y})).$$

Clearly,

$$\tilde{\mathbf{v}}_t = \mathbf{F}_t(\mathbf{Y}_{\mathbf{i}(t)})$$

and

$$\hat{\mathbf{v}}_t = \mathbf{F}_t(\mathbf{Y}_{\mathbf{i}(t)} + r_t).$$

Now let f be the density function of $\mathbf{Y}_{\mathbf{i}(t)}$ and $\mathcal{F}_{\mathbf{i}(t)}$ denote the σ-algebra generated by all random variables before epoch $E_{\mathbf{i}(T)}$.[4] We have

$$\mathbb{E}\left[\hat{\mathbf{v}}_t\middle|\,\mathcal{F}_{\mathbf{i}(t-1)}\right] = \int_{\mathbb{R}^{|\mathcal{X}||\mathcal{A}|}} \mathbf{F}_t(y + r_t)f(y)dy = \int_{\mathbb{R}^{|\mathcal{X}||\mathcal{A}|}} \mathbf{F}_t(y)f(y - r_t)dy$$
$$\le \sup_{y,t}\frac{f(y-r_t)}{f(y)}\int_{\mathbb{R}^{|\mathcal{X}||\mathcal{A}|}} \mathbf{F}_t(y)f(y)dy \le \sup_{y,t}\frac{f(y-r_t)}{f(y)}\mathbb{E}\left[\tilde{\mathbf{v}}_t\middle|\,\mathcal{F}_{\mathbf{i}(t-1)}\right].$$

[4] Note that $\mathbf{Y}_{\mathbf{i}(t)}$ is generated independently from the history up to epoch $\mathbf{i}(t)$.

Since $f(y) = \eta \exp\left(-\eta \sum_{x,a} y(x,a)\right)$ for all $y \succeq 0$, we get

$$\sup_y \frac{f(y-r_t)}{f(y)} = \exp\left(\eta \sum_{x,a} r_t(x,a)\right) \leq \exp\left(\eta |\mathcal{X}||\mathcal{A}|\right).$$

Using $e^x \leq 1 + (e-1)x$ for $x \in [0,1]$, which holds by our assumption on η, we get

$$\mathbb{E}\left[\hat{\mathbf{v}}_t\right] \leq \mathbb{E}\left[\tilde{\mathbf{v}}_t\right] \left(1 + \eta(e-1)|\mathcal{X}||\mathcal{A}|\right).$$

Noticing that $\tilde{\mathbf{v}}_t \leq L$ gives the result. □

E Proof of Lemma 4

We prove the statement by induction on l. For $l = 1$ we have

$$\sum_{x_1} |\tilde{\boldsymbol{\mu}}_t(x_1) - \boldsymbol{\mu}_t(x_1)| = \sum_{x_1} \left| \tilde{\mathbf{P}}_t(x_1|x_0, \boldsymbol{\pi}_t(x_0)) - P(x_1|x_0, \boldsymbol{\pi}_t(x_0)) \right| \leq \mathbf{a}_t(x_0, , \boldsymbol{\pi}_t(x_0)),$$

proving the statement for this case. Now assume that the statement holds for some $l-1$. We have

$$\begin{aligned}
&\tilde{\boldsymbol{\mu}}_t(x_l) - \boldsymbol{\mu}_t(x_l) \\
&= \sum_{x_{l-1}} \left(\tilde{\mathbf{P}}_t(x_l|x_{l-1}, \boldsymbol{\pi}_t(x_{l-1}))\tilde{\boldsymbol{\mu}}_t(x_{l-1}) - P(x_l|x_{l-1}, \boldsymbol{\pi}_t(x_{l-1}))\boldsymbol{\mu}_t(x_{l-1}) \right) \\
&= \sum_{x_{l-1}} \left(\tilde{\mathbf{P}}_t(x_l|x_{l-1}, \boldsymbol{\pi}_t(x_{l-1})) \left(\tilde{\boldsymbol{\mu}}_t(x_{l-1}) - \boldsymbol{\mu}_t(x_{l-1})\right) + \left(\tilde{\mathbf{P}}_t(x_l|x_{l-1}, \boldsymbol{\pi}_t(x_{l-1})) - P(x_l|x_{l-1}, \boldsymbol{\pi}_t(x_{l-1})) \right) \boldsymbol{\mu}_t(x_{l-1}) \right),
\end{aligned}$$

and thus

$$\begin{aligned}
&\sum_{x_l} |\tilde{\boldsymbol{\mu}}_t(x_l) - \boldsymbol{\mu}_t(x_l)| \\
&\leq \sum_{x_l, x_{l-1}} \left(\tilde{\mathbf{P}}_t(x_l|x_{l-1}, \boldsymbol{\pi}_t(x_{l-1})) \left|\tilde{\boldsymbol{\mu}}_t(x_{l-1}) - \boldsymbol{\mu}_t(x_{l-1})\right| + \left| \tilde{\mathbf{P}}_t(x_l|x_{l-1}, \boldsymbol{\pi}_t(x_{l-1})) - P(x_l|x_{l-1}, \boldsymbol{\pi}_t(x_{l-1})) \right| \boldsymbol{\mu}_t(x_{l-1}) \right) \\
&= \sum_{x_{l-1}} |\tilde{\boldsymbol{\mu}}_t(x_{l-1}) - \boldsymbol{\mu}_t(x_{l-1})| + \sum_{x_{l-1}} \boldsymbol{\mu}_t(x_{l-1}) \sum_{x_l} \left| \tilde{\mathbf{P}}_t(x_l|x_{l-1}, \boldsymbol{\pi}_t(x_{l-1})) - P(x_l|x_{l-1}, \boldsymbol{\pi}_t(x_{l-1})) \right| \\
&\leq \sum_{k=0}^{l-2} \sum_{x_k \in \mathcal{X}_k} \boldsymbol{\mu}_t(x_k)\, \mathbf{a}_t(x_k, \boldsymbol{\pi}_t(x_k)) + \sum_{x_{l-1}} \boldsymbol{\mu}_t(x_{l-1}) \sum_{x_l} \mathbf{a}_t(x_{l-1}, \boldsymbol{\pi}_t(x_{l-1})),
\end{aligned}$$

proving the statement. □

F Proof of Lemma 5

We start by some arguments borrowed from Jaksch et al. (2010). Let $\mathbf{n}_i(x,a)$ be the number of times state-action pair (x,a) has been visited in epoch E_i. We have

$$\mathbf{N}_i(x,a) = \sum_{j=1}^{i-1} \mathbf{n}_i(x,a).$$

For simplicity, let $\mathbf{K}_T = m$ be the number of epochs. By Appendix C.3 of Jaksch et al. (2010), we have

$$\sum_{i=1}^{m} \frac{\mathbf{n}_i(x,a)}{\sqrt{\mathbf{N}_i(x,a)}} \leq \left(\sqrt{2}+1\right)\sqrt{\mathbf{N}_m(x,a)},$$

and by Jensen's inequality,

$$\sum_{x,a}\sum_{i=1}^{m}\frac{\mathbf{n}_i(x,a)}{\sqrt{\mathbf{N}_i(x,a)}} \le \left(\sqrt{2}+1\right)\sqrt{|\mathcal{X}||\mathcal{A}|T}.$$

Now fix an arbitrary $1 \le t \le T$. We have

$$\tilde{\mathbf{v}}_t = \sum_{l=0}^{L-1}\sum_{x\in\mathcal{X}_l}\tilde{\boldsymbol{\mu}}_t(x) r_t(x,\boldsymbol{\pi}_t(x))$$

and

$$v_t(\boldsymbol{\pi}_t) = \sum_{l=0}^{L-1}\sum_{x\in\mathcal{X}_l}\boldsymbol{\mu}_t(x) r_t(x,\boldsymbol{\pi}_t(x)),$$

thus

$$\tilde{\mathbf{v}}_t(\boldsymbol{\pi}_t) - v_t(\boldsymbol{\pi}_t) = \sum_{l=0}^{L-1}\sum_{x\in\mathcal{X}_l}\left(\tilde{\boldsymbol{\mu}}_t(x)-\boldsymbol{\mu}_t(x)\right) r_t(x,\boldsymbol{\pi}_t(x)) \le \sum_{l=0}^{L-1}\sum_{x\in\mathcal{X}_l}\left|\tilde{\boldsymbol{\mu}}_t(x)-\boldsymbol{\mu}_t(x)\right|.$$

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

Setting $\mathbf{a}_t(x,a) = \left\|\tilde{\mathbf{P}}_t(\cdot|x,a) - P(\cdot|x,a)\right\|_1$ for all $(x,a)\in\mathcal{X}\times\mathcal{A}$, the conditions of Lemma 4 are clearly satisfied, and so

$$\begin{aligned}\sum_{x\in\mathcal{X}_l}\left|\tilde{\boldsymbol{\mu}}_t(x)-\boldsymbol{\mu}_t(x)\right| &\le \sum_{k=0}^{l-1}\sum_{x_k\in\mathcal{X}_k}\boldsymbol{\mu}_t(x_k)\,\mathbf{a}_t\left(x_k,\boldsymbol{\pi}_t\left(x_k\right)\right)\\ &\le \sum_{k=0}^{l-1}\mathbf{a}_t\left(\mathbf{x}_k^{(t)},\mathbf{a}_k^{(t)}\right) + \sum_{k=0}^{l-1}\sum_{x_k\in\mathcal{X}_k}\left(\boldsymbol{\mu}_t(x_k) - \mathbb{I}_{\left\{\mathbf{x}_k^{(t)}=x_k\right\}}\right)\mathbf{a}_1\left(x_k,\boldsymbol{\pi}_t\left(x_k\right)\right).\end{aligned} \tag{9}$$

Now, by Lemma 1, we have with probability at least $1-\delta$ simultaneously for all k that

$$\begin{aligned}\sum_{t=1}^{T}\mathbf{a}_t\left(\mathbf{x}_k^{(t)},\mathbf{a}_k^{(t)}\right) &\le \sum_{t=1}^{T}\sqrt{\frac{2|\mathcal{X}_{k+1}|\ln\frac{T|\mathcal{X}||\mathcal{A}|}{\delta}}{\max\left\{1,\mathbf{N}_{\mathbf{i}(t)}\left(\mathbf{x}_k^{(t)},\mathbf{a}_k^{(t)}\right)\right\}}}\\ &\le \sum_{x_k,a_k}\sum_{i=1}^{m}\mathbf{n}_i(x_k,a_k)\sqrt{\frac{2|\mathcal{X}_{k+1}|\ln\frac{T|\mathcal{X}||\mathcal{A}|}{\delta}}{\max\left\{1,\mathbf{N}_{\mathbf{i}(t)}\left(x_k,a_k\right)\right\}}}\\ &\le \left(\sqrt{2}+1\right)\sqrt{2\,T\,|\mathcal{X}_k||\mathcal{X}_{k+1}||\mathcal{A}|\ln\frac{T|\mathcal{X}||\mathcal{A}|}{\delta}}.\end{aligned}$$

For the second term on the right hand side of (9), notice that $\left(\boldsymbol{\mu}_t(x_k) - \mathbb{I}_{\left\{\mathbf{x}_k^{(t)}=x_k\right\}}\right)$ form a martingale difference sequence with respect to $\{\mathbf{U}_t\}_{t=1}^{T}$ and thus by the Hoeffding–Azuma inequality and $\mathbf{a}_1 \le 2$, we have

$$\sum_{t=1}^{T}\left(\boldsymbol{\mu}_t(x_k) - \mathbb{I}_{\left\{\mathbf{x}_k^{(t)}=x_k\right\}}\right)\mathbf{a}_1\left(x_k,\boldsymbol{\pi}_t\left(x_k\right)\right) \le \sqrt{2\,T\,\ln\frac{L}{\delta}}$$

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 $l=1,\ldots,L$,

$$
\begin{aligned}
\sum_{t=1}^{T}\sum_{x\in\mathcal{X}_l}\left(\tilde{\boldsymbol{\mu}}_t(x)-\boldsymbol{\mu}_t(x)\right) &\le \sum_{k=0}^{l-1}\left(\sqrt{2}+1\right)\sqrt{T\,|\mathcal{X}_k||\mathcal{X}_{k+1}||\mathcal{A}|\ln\frac{T|\mathcal{X}||\mathcal{A}|}{\delta}}+\sum_{k=0}^{l-1}|\mathcal{X}_k|\sqrt{2\,T\,\ln\frac{L}{\delta}} \\
&\le \left(\sqrt{2}+1\right)L\sum_{k=0}^{L-1}\frac{1}{L}\sqrt{T\,|\mathcal{X}_k||\mathcal{X}_{k+1}||\mathcal{A}|\ln\frac{T|\mathcal{X}||\mathcal{A}|}{\delta}}+\sum_{k=0}^{l-1}|\mathcal{X}_k|\sqrt{2\,T\,\ln\frac{L}{\delta}} \\
&\le \left(\sqrt{2}+1\right)L\sqrt{T|\mathcal{A}|\left(\frac{|\mathcal{X}|}{L}\right)^2\ln\frac{T|\mathcal{X}||\mathcal{A}|}{\delta}}+|\mathcal{X}|\sqrt{2\,T\,\ln\frac{L}{\delta}} \\
&= \left(\sqrt{2}+1\right)|\mathcal{X}|\sqrt{T|\mathcal{A}|\,\ln\frac{T|\mathcal{X}||\mathcal{A}|}{\delta}}+|\mathcal{X}|\sqrt{2\,T\,\ln\frac{L}{\delta}} \qquad (10)
\end{aligned}
$$

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 $v_t(\boldsymbol{\pi}_t)-\tilde{\mathbf{v}}_t\le L$ and (10) holds with probability at least $1-2\delta$, finishes the proof. $\square$