Supplementary material

Here we prove the theorems and derive the recursive relationships stated but not proved in the main text. First we prove the recursions in the time horizon t for the forward-view errors used by PTD(λ) and PQ(λ). We then prove a recursion in k for PTD (Lemma 1) and use it to prove Theorem 3 for PTD. Next we prove an analogous recursion in k (Lemma 2) and theorem (Theorem 5) for PQ and for action values. Finally, we provide some further detail on a key step in the derivations of the update of the provisional weights, $\boldsymbol{u}_t$, for both algorithms.

S.1 Derivation of Equation (11), the PTD recursion in t

From (6), for $k < t$, we immediately have

$$\begin{aligned}
\delta^{\lambda\rho}_{k,t+1} &= \rho_k \sum_{i=k+1}^{t} C_k^{i-1}\Big[(1-\gamma_i)\epsilon_k^i + \gamma_i(1-\lambda_i)\bar{\delta}_k^i\Big] + \rho_k C_k^t \Big[(1-\gamma_{t+1})\epsilon_k^{t+1} + \gamma_{t+1}\bar{\delta}_k^{t+1}\Big] \\
&= \rho_k \sum_{i=k+1}^{t-1} C_k^{i-1}\Big[(1-\gamma_i)\epsilon_k^i + \gamma_i(1-\lambda_i)\bar{\delta}_k^i\Big] + \rho_k C_k^{t-1}\Big[(1-\gamma_t)\epsilon_k^t + \gamma_t(1-\lambda_t)\bar{\delta}_k^t\Big] \\
&\qquad\qquad + \rho_k C_k^t \Big[(1-\gamma_{t+1})\epsilon_k^{t+1} + \gamma_{t+1}\bar{\delta}_k^{t+1}\Big] \\
&= \underbrace{\rho_k \sum_{i=k+1}^{t-1} C_k^{i-1}\Big[(1-\gamma_i)\epsilon_k^i + \gamma_i(1-\lambda_i)\bar{\delta}_k^i\Big] + \rho_k C_k^{t-1}\Big[(1-\gamma_t)\epsilon_k^t + \gamma_t\bar{\delta}_k^t\Big]}_{\delta^{\lambda\rho}_{k,t}} \\
&\qquad\qquad - \rho_k C_k^{t-1}\gamma_t\lambda_t\bar{\delta}_k^t \;\; + \rho_k C_k^t \Big[(1-\gamma_{t+1})\epsilon_k^{t+1} + \gamma_{t+1}\bar{\delta}_k^{t+1}\Big] \\
&= \delta^{\lambda\rho}_{k,t} - \rho_k C_k^{t-1}\gamma_t\lambda_t\bar{\delta}_k^t + \rho_k C_k^t\Big[(1-\gamma_{t+1})\epsilon_k^{t+1} + \gamma_{t+1}\bar{\delta}_k^{t+1}\Big]. \qquad (28)
\end{aligned}$$

Although this is already a recursion of the desired form, expressing $\delta^{\lambda\rho}_{k,t+1}$ in terms of $\delta^{\lambda\rho}_{k,t}$, we are not done yet. The recursion can be simplified further by noting that

$$\begin{aligned}
(1-\gamma_{t+1})\epsilon_k^{t+1} + \gamma_{t+1}\bar{\delta}_k^{t+1} &= (1-\gamma_{t+1})\left(\sum_{i=k+1}^{t+1} R_i - \boldsymbol{\theta}^\top\boldsymbol{\phi}_k\right) + \gamma_{t+1}\left(\sum_{i=k+1}^{t+1} R_i + \boldsymbol{\theta}^\top\boldsymbol{\phi}_{t+1} - \boldsymbol{\theta}^\top\boldsymbol{\phi}_k\right) \\
&= \sum_{i=k+1}^{t+1} R_i - \boldsymbol{\theta}^\top\boldsymbol{\phi}_k + \gamma_{t+1}\boldsymbol{\theta}^\top\boldsymbol{\phi}_{t+1} \\
&= \sum_{i=k+1}^{t} R_i - \boldsymbol{\theta}^\top\boldsymbol{\phi}_k + R_{t+1} + \gamma_{t+1}\boldsymbol{\theta}^\top\boldsymbol{\phi}_{t+1} - \boldsymbol{\theta}^\top\boldsymbol{\phi}_t + \boldsymbol{\theta}^\top\boldsymbol{\phi}_t \\
&= \sum_{i=k+1}^{t} R_i - \boldsymbol{\theta}^\top\boldsymbol{\phi}_k + \delta_t + \boldsymbol{\theta}^\top\boldsymbol{\phi}_t \\
&= \bar{\delta}_k^t + \delta_t\,.
\end{aligned}$$

Substituting this in (28), we obtain our final recursion:

$$\begin{aligned}
\delta^{\lambda\rho}_{k,t+1} &= \delta^{\lambda\rho}_{k,t} - \rho_k C_k^{t-1}\gamma_t\lambda_t\bar{\delta}_k^t + \rho_k C_k^t\left(\bar{\delta}_k^t + \delta_t\right) \\
&= \delta^{\lambda\rho}_{k,t} - \rho_k C_k^{t-1}\gamma_t\lambda_t\bar{\delta}_k^t + \rho_k C_k^{t-1}\gamma_t\lambda_t\rho_t\bar{\delta}_k^t + \rho_k C_k^t\delta_t \\
&= \delta^{\lambda\rho}_{k,t} + \rho_k C_k^t\delta_t + (\rho_t - 1)\gamma_t\lambda_t\rho_k C_k^{t-1}\bar{\delta}_k^t. \qquad (11)
\end{aligned}$$

S.2 Derivation of Equation (24), the PQ recursion in t

The first steps of this derivation are directly analogous to those in the previous section leading to (28), except here using the definitions for the action-value case in Section 5. We do not repeat these steps here. In this case they lead to

$$\delta^{\lambda\rho}_{k,t+1} = \delta^{\lambda\rho}_{k,t} - C^{t-1}_k \gamma_t \lambda_t \bar{\delta}^t_k + C^t_k \Big[(1-\gamma_{t+1})\epsilon^{t+1}_k + \gamma_{t+1} \bar{\delta}^{t+1}_k \Big], \tag{29}$$

for all $k < t$. Note that, compared to (28), ρ_k is absent.

Again, this recursion can be simplified. Using (20–22) we get

$$\begin{aligned}
(1-\gamma_{t+1})\epsilon^{t+1}_k + \gamma_{t+1}\bar{\delta}^{t+1}_k &= (1-\gamma_{t+1}) \left(\sum_{i=k+1}^{t+1} R_i - \boldsymbol{\theta}^\top \boldsymbol{\phi}_k \right) + \gamma_{t+1} \left(\sum_{i=k+1}^{t+1} R_i + \boldsymbol{\theta}^\top \bar{\boldsymbol{\phi}}^\pi_{t+1} - \boldsymbol{\theta}^\top \boldsymbol{\phi}_k \right) \\
&= \sum_{i=k+1}^{t+1} R_i - \boldsymbol{\theta}^\top \boldsymbol{\phi}_k + \gamma_{t+1} \boldsymbol{\theta}^\top \bar{\boldsymbol{\phi}}^\pi_{t+1} \\
&= \sum_{i=k+1}^{t} R_i - \boldsymbol{\theta}^\top \boldsymbol{\phi}_k + R_{t+1} + \gamma_{t+1} \boldsymbol{\theta}^\top \bar{\boldsymbol{\phi}}^\pi_{t+1} - \boldsymbol{\theta}^\top \boldsymbol{\phi}_t + \boldsymbol{\theta}^\top \boldsymbol{\phi}_t \\
&= \epsilon^t_k + \delta_t + \boldsymbol{\theta}^\top \boldsymbol{\phi}_t \\
&= \bar{\delta}^t_k - \boldsymbol{\theta}^\top \bar{\boldsymbol{\phi}}^\pi_t + \delta_t + \boldsymbol{\theta}^\top \boldsymbol{\phi}_t \\
&= \bar{\delta}^t_k + \delta_t + \boldsymbol{\theta}^\top (\boldsymbol{\phi}_t - \bar{\boldsymbol{\phi}}^\pi_t)\,.
\end{aligned}$$

Using this in (29), we obtain our final recursion:

$$\begin{aligned}
\delta^{\lambda\rho}_{k,t+1} &= \delta^{\lambda\rho}_{k,t} - C^{t-1}_k \gamma_t \lambda_t \bar{\delta}^t_k + C^t_k \left(\bar{\delta}^t_k + \delta_t + \boldsymbol{\theta}^\top (\boldsymbol{\phi}_t - \bar{\boldsymbol{\phi}}^\pi_t) \right) \\
&= \delta^{\lambda\rho}_{k,t} - C^{t-1}_k \gamma_t \lambda_t \bar{\delta}^t_k + C^{t-1}_k \gamma_t \lambda_t \rho_t \bar{\delta}^t_k + C^t_k \delta_t + C^t_k \boldsymbol{\theta}^\top (\boldsymbol{\phi}_t - \bar{\boldsymbol{\phi}}^\pi_t) \\
&= \delta^{\lambda\rho}_{k,t} + C^t_k \delta_t + C^t_k \boldsymbol{\theta}^\top (\boldsymbol{\phi}_t - \bar{\boldsymbol{\phi}}^\pi_t) + (\rho_t - 1)\gamma_t \lambda_t C^{t-1}_k \bar{\delta}^t_k\,.
\end{aligned} \tag{24}$$

S.3 Lemma 1: PTD recursion in k

The following lemma, used in proving Theorem 3 in the next section, shows how $\delta^{\lambda\rho}_{k,t}$ depends on $\delta^{\lambda\rho}_{k+1,t}$. All definitions are from Sections 2–4 (the state-value or PTD case).

Lemma 1 (PTD error recursion in k)**.** *For all $k < t-1$,*

$$\delta^{\lambda\rho}_{k,t} = \rho_k \left(\delta_k + \left(D^t_k - 1 \right) \bar{\delta}^{k+1}_k + \gamma_{k+1}\lambda_{k+1}\delta^{\lambda\rho}_{k+1,t} \right), \tag{30}$$

where

$$D^t_k = \sum_{i=k+1}^{t-1} C^{i-1}_k (1-\gamma_i\lambda_i) + C^{t-1}_k\,. \tag{31}$$

Proof. First note that from definitions (3) and (4) it is clear that

$$\bar{\delta}_k^i = \epsilon_k^i + \boldsymbol{\theta}^\top \boldsymbol{\phi}_i \tag{32}$$

and

$$\begin{aligned}\epsilon_k^i &= R_{k+1} + R_{k+2} + \cdots + R_i - \boldsymbol{\theta}^\top \boldsymbol{\phi}_k && (3)\\ &= R_{k+1} + R_{k+2} + \cdots + R_i - \boldsymbol{\theta}^\top \boldsymbol{\phi}_{k+1} + \boldsymbol{\theta}^\top \boldsymbol{\phi}_{k+1} - \boldsymbol{\theta}^\top \boldsymbol{\phi}_k \\ &= R_{k+1} + \epsilon_{k+1}^i + \boldsymbol{\theta}^\top \boldsymbol{\phi}_{k+1} - \boldsymbol{\theta}^\top \boldsymbol{\phi}_k \\ &= \bar{\delta}_k^{k+1} + \epsilon_{k+1}^i. && (33)\end{aligned}$$

Using these, the lemma can be directly derived:

$$\begin{aligned}
\delta_{k,t}^{\lambda\rho} &= \rho_k \left(\sum_{i=k+1}^{t-1} C_k^{i-1} \Big[(1-\gamma_i)\epsilon_k^i + \gamma_i(1-\lambda_i)\bar{\delta}_k^i \Big] + C_k^{t-1} \Big[(1-\gamma_t)\epsilon_k^t + \gamma_t \bar{\delta}_k^t \Big] \right) && \text{(as in (6))}\\
&= \rho_k \left(\sum_{i=k+1}^{t-1} C_k^{i-1} \Big[(1-\gamma_i)\epsilon_k^i + \gamma_i(1-\lambda_i) \left(\epsilon_k^i + \boldsymbol{\theta}^\top \boldsymbol{\phi}_i \right) \Big] + C_k^{t-1} \Big[(1-\gamma_t)\epsilon_k^t + \gamma_t \left(\epsilon_k^t + \boldsymbol{\theta}^\top \boldsymbol{\phi}_t \right) \Big] \right) && \text{(using (32))}\\
&= \rho_k \left(\sum_{i=k+1}^{t-1} C_k^{i-1} \Big[(1-\gamma_i\lambda_i)\epsilon_k^i + \gamma_i(1-\lambda_i)\boldsymbol{\theta}^\top \boldsymbol{\phi}_i \Big] + C_k^{t-1} \Big[\epsilon_k^t + \gamma_t \boldsymbol{\theta}^\top \boldsymbol{\phi}_t \Big] \right) && (34)\\
&= \rho_k \Bigg(\sum_{i=k+2}^{t-1} C_k^{i-1} \Big[(1-\gamma_i\lambda_i)\epsilon_k^i + \gamma_i(1-\lambda_i)\boldsymbol{\theta}^\top \boldsymbol{\phi}_i \Big] + C_k^{t-1} \Big[\epsilon_k^t + \gamma_t \boldsymbol{\theta}^\top \boldsymbol{\phi}_t \Big] \\
&\qquad + C_k^k \Big[(1-\gamma_{k+1}\lambda_{k+1})\epsilon_k^{k+1} + \gamma_{k+1}(1-\lambda_{k+1})\boldsymbol{\theta}^\top \boldsymbol{\phi}_{k+1} \Big] \Bigg)\\
&= \rho_k \Bigg(\sum_{i=k+2}^{t-1} C_k^{i-1} \Big[(1-\gamma_i\lambda_i) \left(\bar{\delta}_k^{k+1} + \epsilon_{k+1}^i \right) + \gamma_i(1-\lambda_i)\boldsymbol{\theta}^\top \boldsymbol{\phi}_i \Big] + C_k^{t-1} \Big[\bar{\delta}_k^{k+1} + \epsilon_{k+1}^t + \gamma_t \boldsymbol{\theta}^\top \boldsymbol{\phi}_t \Big] && \text{(using (33))}\\
&\qquad + (1-\gamma_{k+1}\lambda_{k+1}) \left(R_{k+1} - \boldsymbol{\theta}^\top \boldsymbol{\phi}_k \right) + \gamma_{k+1}(1-\lambda_{k+1})\boldsymbol{\theta}^\top \boldsymbol{\phi}_{k+1} \Bigg) && \text{(using } C_k^k = 1\text{)}\\
&= \rho_k \Bigg(\sum_{i=k+2}^{t-1} C_k^{i-1} \Big[(1-\gamma_i\lambda_i)\epsilon_{k+1}^i + \gamma_i(1-\lambda_i)\boldsymbol{\theta}^\top \boldsymbol{\phi}_i \Big] + C_k^{t-1} \Big[\epsilon_{k+1}^t + \gamma_t \boldsymbol{\theta}^\top \boldsymbol{\phi}_t \Big] + \sum_{i=k+2}^{t-1} C_k^{i-1}(1-\gamma_i\lambda_i)\bar{\delta}_k^{k+1} + C_k^{t-1}\bar{\delta}_k^{k+1} \\
&\qquad + R_{k+1} - \boldsymbol{\theta}^\top \boldsymbol{\phi}_k + \gamma_{k+1}\boldsymbol{\theta}^\top \boldsymbol{\phi}_{k+1} - \gamma_{k+1}\lambda_{k+1} \left(R_{k+1} - \boldsymbol{\theta}^\top \boldsymbol{\phi}_k + \boldsymbol{\theta}^\top \boldsymbol{\phi}_{k+1} \right) \Bigg)\\
&= \rho_k \Bigg(\sum_{i=k+2}^{t-1} \gamma_{k+1}\lambda_{k+1}\rho_{k+1} C_{k+1}^{i-1} \Big[(1-\gamma_i\lambda_i)\epsilon_{k+1}^i + \gamma_i(1-\lambda_i)\boldsymbol{\theta}^\top \boldsymbol{\phi}_i \Big] + \gamma_{k+1}\lambda_{k+1}\rho_{k+1} C_{k+1}^{t-1} \Big[\epsilon_{k+1}^t + \gamma_t \boldsymbol{\theta}^\top \boldsymbol{\phi}_t \Big] \\
&\qquad + \sum_{i=k+2}^{t-1} C_k^{i-1}(1-\gamma_i\lambda_i)\bar{\delta}_k^{k+1} + C_k^{t-1}\bar{\delta}_k^{k+1} + \delta_k - \gamma_{k+1}\lambda_{k+1}\bar{\delta}_k^{k+1} \Bigg)\\
&= \rho_k \left(\gamma_{k+1}\lambda_{k+1}\delta_{k+1,t}^{\lambda\rho} + \left[\sum_{i=k+2}^{t-1} C_k^{i-1}(1-\gamma_i\lambda_i) + 1 - \gamma_{k+1}\lambda_{k+1} + C_k^{t-1} - 1 \right] \bar{\delta}_k^{k+1} + \delta_k \right) && \text{(using (34))}\\
&= \rho_k \left(\gamma_{k+1}\lambda_{k+1}\delta_{k+1,t}^{\lambda\rho} + \left[\sum_{i=k+2}^{t-1} C_k^{i-1}(1-\gamma_i\lambda_i) + C_k^k(1-\gamma_{k+1}\lambda_{k+1}) + C_k^{t-1} - 1 \right] \bar{\delta}_k^{k+1} + \delta_k \right)\\
&= \rho_k \left(\gamma_{k+1}\lambda_{k+1}\delta_{k+1,t}^{\lambda\rho} + \left[\sum_{i=k+1}^{t-1} C_k^{i-1}(1-\gamma_i\lambda_i) + C_k^{t-1} - 1 \right] \bar{\delta}_k^{k+1} + \delta_k \right)\\
&= \rho_k \left(\delta_k + \left[D_k^t - 1 \right] \bar{\delta}_k^{k+1} + \gamma_{k+1}\lambda_{k+1}\delta_{k+1,t}^{\lambda\rho} \right). && \square
\end{aligned}$$

S.4 Proof of Theorem 3 (On-policy and off-policy expectations for PTD)

All definitions here are from Sections 2-4 (the state-value or PTD case).

Theorem 3 (On-policy and off-policy expectations)**.** *For any state s,*

$$\mathbb{E}_b\left[\delta_{k,t}^{\lambda\rho}\middle|S_k=s\right] = \mathbb{E}_\pi\left[\delta_{k,t}^{\lambda 1}\middle|S_k=s\right], \tag{35}$$

where $\mathbb{E}_b$ and $\mathbb{E}_\pi$ denote expectations under the behavior and target policies, and $\delta_{k,t}^{\lambda 1}$ denotes $\delta_{k,t}^{\lambda\rho}$ with $\rho_t=1$ for all t.

Proof. First we note that

$$\begin{aligned}
\mathbb{E}_b\left[\rho_k C_k^t\middle|S_k=s\right] &= \mathbb{E}_b\left[\rho_k\prod_{i=k+1}^{t}\gamma_i\lambda_i\rho_i\middle|S_k=s\right]\\
&= \sum_a b(a|s)\sum_{s'}p(s'|s,a)\frac{\pi(a|s)}{b(a|s)}\gamma(s')\lambda(s')\mathbb{E}_b\left[\prod_{i=k+2}^{t}\gamma_i\lambda_i\rho_i\middle|S_k=s,A_k=a,S_{k+1}=s'\right]\\
&= \sum_a \pi(a|s)\sum_{s'}p(s'|s,a)\gamma(s')\lambda(s')\mathbb{E}_b\left[\rho_{k+1}\prod_{i=k+2}^{t}\gamma_i\lambda_i\rho_i\middle|S_{k+1}=s'\right]\\
&= \sum_a \pi(a|s)\sum_{s'}p(s'|s,a)\gamma(s')\lambda(s')\sum_{a'}\pi(a'|s')\sum_{s''}p(s''|s',a')\gamma(s'')\lambda(s'')\cdots\\
&= \mathbb{E}_\pi\left[\prod_{i=k+1}^{t}\gamma_i\lambda_i\middle|S_k=s\right],
\end{aligned}$$

from which one can show

$$\begin{aligned}
\mathbb{E}_b\left[\rho_k D_k^t\middle|S_k=s\right] &= \mathbb{E}_b\left[\rho_k\left(\sum_{i=k+1}^{t-1}C_k^{i-1}(1-\gamma_i\lambda_i)+C_k^{t-1}\right)\middle|S_k=s\right] && \text{(using (31))}\\
&= \mathbb{E}_\pi\left[\sum_{i=k+1}^{t-1}\prod_{i=k+1}^{i}\gamma_i\lambda_i(1-\gamma_i\lambda_i)+\prod_{i=k+1}^{t}\gamma_i\lambda_i\middle|S_k=s\right]\\
&= 1\,. && (36)
\end{aligned}$$

Now we can start directly from the left-hand side of the theorem statement:

$$\begin{aligned}
\mathbb{E}_b\left[\delta_{k,t}^{\lambda\rho}\middle|S_k=s\right] &= \mathbb{E}_b\left[\rho_k\left(\delta_k+\left(D_k^t-1\right)\bar{\delta}_k^{k+1}+\gamma_{k+1}\lambda_{k+1}\delta_{k+1,t}^{\lambda\rho}\right)\middle|S_k=s\right] && \text{(using Lemma 1)}\\
&= \mathbb{E}_b\left[\rho_k\left(\delta_k+\gamma_{k+1}\lambda_{k+1}\delta_{k+1,t}^{\lambda\rho}\right)\middle|S_k=s\right] && \text{(using (36))}\\
&= \sum_a b(a|s)\frac{\pi(a|s)}{b(a|s)}\left(\mathbb{E}_b[\delta_k|S_k=s,A_k=a]+\mathbb{E}_b\left[\gamma_{k+1}\lambda_{k+1}\delta_{k+1,t}^{\lambda\rho}\middle|S_k=s,A_k=a\right]\right)\\
&= \sum_a \pi(a|s)\left(\mathbb{E}_b[\delta_k|S_k=s,A_k=a]+\mathbb{E}_b\left[\gamma_{k+1}\lambda_{k+1}\delta_{k+1,t}^{\lambda\rho}\middle|S_k=s,A_k=a\right]\right)\\
&= \mathbb{E}_\pi[\delta_k|S_k=s]+\sum_a\pi(a|s)\sum_{s'}p(s'|s,a)\gamma(s')\lambda(s')\mathbb{E}_b\left[\delta_{k+1,t}^{\lambda\rho}\middle|S_{k+1}=s'\right]\\
&= \mathbb{E}_\pi\left[\delta_k+\gamma_{k+1}\lambda_{k+1}\mathbb{E}_b\left[\delta_{k+1,t}^{\lambda\rho}\middle|S_{k+1}\right]\middle|S_k=s\right]\\
&= \mathbb{E}_\pi\left[\delta_k+\gamma_{k+1}\lambda_{k+1}\delta_{k+1}+\gamma_{k+2}\lambda_{k+2}\mathbb{E}_b\left[\delta_{k+2,t}^{\lambda\rho}\middle|S_{k+2}\right]\middle|S_k=s\right]\\
&\ \ \vdots
\end{aligned}$$

$$= \mathbb{E}_\pi \left[\sum_{j=k}^{t-1} \left(\prod_{i=k+1}^{j} \gamma_i \lambda_i \right) \delta_j \middle| S_k = s \right].$$

It thus only remains to show that $\delta_{k,t}^{\lambda 1}$ is equal to this sum, which we can show directly from (11) and the definition of $\delta_{k,t}^{\lambda 1}$:

$$\begin{aligned}
\delta_{k,t}^{\lambda 1} &= \delta_{k,t-1}^{\lambda 1} + \left(\prod_{i=k+1}^{t-1} \gamma_i \lambda_i \right) \delta_{t-1} \\
&= \delta_{k,t-2}^{\lambda 1} + \left(\prod_{i=k+1}^{t-2} \gamma_i \lambda_i \right) \delta_{t-2} + \left(\prod_{i=k+1}^{t-1} \gamma_i \lambda_i \right) \delta_{t-1} \\
&\vdots \\
&= \sum_{j=k}^{t-1} \left(\prod_{i=k+1}^{j} \gamma_i \lambda_i \right) \delta_j .
\end{aligned}$$

□

S.5 Lemma 2: PQ recursion in k

This lemma is the analog of Lemma 1 for the action-value case, showing how $\delta_{k,t}^{\lambda\rho}$ depends on $\delta_{k+1,t}^{\lambda\rho}$ when these errors are defined by (18–24). This lemma assists in proving Theorem 5 below. All definitions here are as in Section 5 (the action-value or PQ case), plus D_k^t as in Lemma 1.

Lemma 2 (PQ error recursion in k)**.** *For all $k < t-1$,*

$$\delta_{k,t}^{\lambda\rho} = \delta_k + \gamma_{k+1}\lambda_{k+1}\rho_{k+1}\delta_{k+1,t}^{\lambda\rho} + \gamma_{k+1}\lambda_{k+1}\boldsymbol{\theta}^\top(\boldsymbol{\phi}_{k+1} - \bar{\boldsymbol{\phi}}_{k+1}^\pi) + (D_k^t - 1)\left(\epsilon_k^{k+1} + \boldsymbol{\theta}^\top\boldsymbol{\phi}_{k+1}\right) . \quad (37)$$

Proof. The proof is analogous to that of Lemma 1. Here we have the helper identities

$$\bar{\delta}_k^i = \epsilon_k^i + \boldsymbol{\theta}^\top \bar{\boldsymbol{\phi}}_i^\pi, \quad (38)$$

and

$$\epsilon_k^i = \epsilon_{k+1}^i + \epsilon_k^{k+1} + \boldsymbol{\theta}^\top \boldsymbol{\phi}_{k+1}. \quad (39)$$

Then we can proceed directly:

$$\begin{aligned}
\delta_{k,t}^{\lambda\rho} &= \sum_{i=k+1}^{t-1} C_k^{i-1} \Big[(1-\gamma_i)\epsilon_k^i + \gamma_i(1-\lambda_i)\bar{\delta}_k^i \Big] + C_k^{t-1} \left[(1-\gamma_t)\epsilon_k^t + \gamma_t \bar{\delta}_k^t \right] && \text{(as in (23))} \\
&= \sum_{i=k+1}^{t-1} C_k^{i-1} \Big[(1-\gamma_i)\epsilon_k^i + \gamma_i(1-\lambda_i)\left(\epsilon_k^i + \boldsymbol{\theta}^\top\bar{\boldsymbol{\phi}}_i^\pi\right) \Big] + C_k^{t-1} \Big[(1-\gamma_t)\epsilon_k^t + \gamma_t\left(\epsilon_k^t + \boldsymbol{\theta}^\top\bar{\boldsymbol{\phi}}_t^\pi\right) \Big] && \text{(using (38))} \\
&= \sum_{i=k+1}^{t-1} C_k^{i-1} \Big[(1-\gamma_i\lambda_i)\epsilon_k^i + \gamma_i(1-\lambda_i)\boldsymbol{\theta}^\top\bar{\boldsymbol{\phi}}_i^\pi \Big] + C_k^{t-1} \Big[\epsilon_k^t + \gamma_t\boldsymbol{\theta}^\top\bar{\boldsymbol{\phi}}_t^\pi \Big] && (40) \\
&= \sum_{i=k+2}^{t-1} C_k^{i-1} \Big[(1-\gamma_i\lambda_i)\epsilon_k^i + \gamma_i(1-\lambda_i)\boldsymbol{\theta}^\top\bar{\boldsymbol{\phi}}_i^\pi \Big] + C_k^{t-1} \Big[\epsilon_k^t + \gamma_t\boldsymbol{\theta}^\top\bar{\boldsymbol{\phi}}_t^\pi \Big] \\
&\qquad + C_k^k \Big[(1-\gamma_{k+1}\lambda_{k+1})\epsilon_k^{k+1} + \gamma_{k+1}(1-\lambda_{k+1})\boldsymbol{\theta}^\top\bar{\boldsymbol{\phi}}_{k+1}^\pi \Big] \\
&= \sum_{i=k+2}^{t-1} C_k^{i-1} \Big[(1-\gamma_i\lambda_i)\left(\epsilon_{k+1}^i + \epsilon_k^{k+1} + \boldsymbol{\theta}^\top\boldsymbol{\phi}_{k+1}\right) + \gamma_i(1-\lambda_i)\boldsymbol{\theta}^\top\bar{\boldsymbol{\phi}}_i^\pi \Big] + C_k^{t-1} \Big[\epsilon_{k+1}^t + \epsilon_k^{k+1} + \boldsymbol{\theta}^\top\boldsymbol{\phi}_{k+1} + \gamma_t\boldsymbol{\theta}^\top\bar{\boldsymbol{\phi}}_t^\pi \Big] \\
&\qquad + (1-\gamma_{k+1}\lambda_{k+1})\left(R_{k+1} - \boldsymbol{\theta}^\top\boldsymbol{\phi}_k\right) + \gamma_{k+1}(1-\lambda_{k+1})\boldsymbol{\theta}^\top\bar{\boldsymbol{\phi}}_{k+1}^\pi && \text{(using (39) and } C_k^k = 1)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=k+2}^{t-1} C_k^{i-1}\Big[(1-\gamma_i\lambda_i)\epsilon_{k+1}^i + \gamma_i(1-\lambda_i)\boldsymbol{\theta}^\top\bar{\boldsymbol{\phi}}_i^\pi\Big] + C_k^{t-1}\Big[\epsilon_{k+1}^t + \gamma_t\boldsymbol{\theta}^\top\bar{\boldsymbol{\phi}}_t^\pi\Big] \\
&\qquad + \sum_{i=k+2}^{t-1} C_k^{i-1}\Big[(1-\gamma_i\lambda_i)\left(\epsilon_k^{k+1} + \boldsymbol{\theta}^\top\boldsymbol{\phi}_{k+1}\right)\Big] + C_k^{t-1}\Big[\epsilon_k^{k+1} + \boldsymbol{\theta}^\top\boldsymbol{\phi}_{k+1}\Big] \\
&\qquad + R_{k+1} - \boldsymbol{\theta}^\top\boldsymbol{\phi}_k + \gamma_{k+1}\boldsymbol{\theta}^\top\bar{\boldsymbol{\phi}}_{k+1}^\pi - \gamma_{k+1}\lambda_{k+1}\left(R_{k+1} - \boldsymbol{\theta}^\top\boldsymbol{\phi}_k + \boldsymbol{\theta}^\top\bar{\boldsymbol{\phi}}_{k+1}^\pi\right) \\
&= \sum_{i=k+2}^{t-1} \gamma_{k+1}\lambda_{k+1}\rho_{k+1}C_{k+1}^{i-1}\Big[(1-\gamma_i\lambda_i)\epsilon_{k+1}^i + \gamma_i(1-\lambda_i)\boldsymbol{\theta}^\top\bar{\boldsymbol{\phi}}_i^\pi\Big] + \gamma_{k+1}\lambda_{k+1}\rho_{k+1}C_{k+1}^{t-1}\Big[\epsilon_{k+1}^t + \gamma_t\boldsymbol{\theta}^\top\bar{\boldsymbol{\phi}}_t^\pi\Big] \\
&\qquad + \sum_{i=k+2}^{t-1} C_k^{i-1}\Big[(1-\gamma_i\lambda_i)\left(\epsilon_k^{k+1} + \boldsymbol{\theta}^\top\boldsymbol{\phi}_{k+1}\right)\Big] + C_k^{t-1}\Big[\epsilon_k^{k+1} + \boldsymbol{\theta}^\top\boldsymbol{\phi}_{k+1}\Big] + \delta_k - \gamma_{k+1}\lambda_{k+1}\bar{\delta}_k^{k+1} \\
&= \gamma_{k+1}\lambda_{k+1}\rho_{k+1}\delta_{k+1,t}^{\lambda\rho} + \left(\sum_{i=k+2}^{t-1} C_k^{i-1}(1-\gamma_i\lambda_i) + C_k^{t-1}\right)\left(\epsilon_k^{k+1} + \boldsymbol{\theta}^\top\boldsymbol{\phi}_{k+1}\right) + \delta_k - \gamma_{k+1}\lambda_{k+1}\bar{\delta}_k^{k+1} \\
&\qquad \text{(using (40); now add and subtract } (1-\gamma_{k+1}\lambda_{k+1})(\epsilon_k^{k+1} + \boldsymbol{\theta}^\top\boldsymbol{\phi}_{k+1})\text{, the first element of the summation)} \\
&= \gamma_{k+1}\lambda_{k+1}\rho_{k+1}\delta_{k+1,t}^{\lambda\rho} + \left(\sum_{i=k+1}^{t-1} C_k^{i-1}(1-\gamma_i\lambda_i) + C_k^{t-1} - 1 + \gamma_{k+1}\lambda_{k+1}\right)\left(\epsilon_k^{k+1} + \boldsymbol{\theta}^\top\boldsymbol{\phi}_{k+1}\right) + \delta_k - \gamma_{k+1}\lambda_{k+1}\bar{\delta}_k^{k+1} \\
&= \gamma_{k+1}\lambda_{k+1}\rho_{k+1}\delta_{k+1,t}^{\lambda\rho} + \left(D_k^t - 1 + \gamma_{k+1}\lambda_{k+1}\right)\left(\epsilon_k^{k+1} + \boldsymbol{\theta}^\top\boldsymbol{\phi}_{k+1}\right) + \delta_k - \gamma_{k+1}\lambda_{k+1}\bar{\delta}_k^{k+1} \\
&= \delta_k + \gamma_{k+1}\lambda_{k+1}\rho_{k+1}\delta_{k+1,t}^{\lambda\rho} + \gamma_{k+1}\lambda_{k+1}(\epsilon_k^{k+1} + \boldsymbol{\theta}^\top\boldsymbol{\phi}_{k+1} - \bar{\delta}_k^{k+1}) + (D_k^t - 1)\left(\epsilon_k^{k+1} + \boldsymbol{\theta}^\top\boldsymbol{\phi}_{k+1}\right) \\
&= \delta_k + \gamma_{k+1}\lambda_{k+1}\rho_{k+1}\delta_{k+1,t}^{\lambda\rho} + \gamma_{k+1}\lambda_{k+1}\boldsymbol{\theta}^\top(\boldsymbol{\phi}_{k+1} - \bar{\boldsymbol{\phi}}_{k+1}^\pi) + (D_k^t - 1)\left(\epsilon_k^{k+1} + \boldsymbol{\theta}^\top\boldsymbol{\phi}_{k+1}\right)\,. \qquad \text{(using (38))}
\end{aligned}
$$

□

S.6 Theorem 5: On-policy and off-policy expectations for PQ

All definitions here are from Section 5 (the state-value or PQ case), plus D_k^t from Lemma 1.

Theorem 5 (On-policy and off-policy expectations)**.** *For any state s,*

$$
\mathbb{E}_b\Big[\delta_{k,t}^{\lambda\rho}\Big|S_k = s, A_k = a\Big] = \mathbb{E}_\pi\Big[\delta_{k,t}^{\lambda 1}\Big|S_k = s, A_k = a\Big]\,,
$$

where $\mathbb{E}_b$ and $\mathbb{E}_\pi$ denote expectations under the behavior and target policies, and $\delta_{k,t}^{\lambda 1}$ denotes $\delta_{k,t}^{\lambda\rho}$ with $\rho_t = 1$ for all t.

Proof. The proof is analogous to that for Theorem 3. Using Lemma 2, the left-hand side can be written

$$
\begin{aligned}
&\mathbb{E}_b\Big[\delta_{k,t}^{\lambda\rho}\Big|S_k = s, A_k = a\Big] \\
&= \mathbb{E}_b\Big[\delta_k + \gamma_{k+1}\lambda_{k+1}\rho_{k+1}\delta_{k+1,t}^{\lambda\rho} + \gamma_{k+1}\lambda_{k+1}\boldsymbol{\theta}^\top(\boldsymbol{\phi}_{k+1} - \bar{\boldsymbol{\phi}}_{k+1}^\pi) + (D_k^t - 1)\left(\epsilon_k^{k+1} + \boldsymbol{\theta}^\top\boldsymbol{\phi}_{k+1}\right)\Big|S_k = s, A_k = a\Big] \\
&= \mathbb{E}_b\Big[\delta_k + \gamma_{k+1}\lambda_{k+1}\rho_{k+1}\delta_{k+1,t}^{\lambda\rho} + \gamma_{k+1}\lambda_{k+1}\boldsymbol{\theta}^\top(\boldsymbol{\phi}_{k+1} - \bar{\boldsymbol{\phi}}_{k+1}^\pi)\Big|S_k = s, A_k = a\Big]
\end{aligned}
$$

(using $\mathbb{E}_b[X|S_k = s, A_k = a] = \mathbb{E}_\pi[\mathbb{E}_b[X|S_{k+1}]|S_k = s, A_k = a]$)

$$
= \mathbb{E}_\pi\Big[\mathbb{E}_b\Big[\delta_k + \gamma_{k+1}\lambda_{k+1}\rho_{k+1}\delta_{k+1,t}^{\lambda\rho} + \gamma_{k+1}\lambda_{k+1}\boldsymbol{\theta}^\top(\boldsymbol{\phi}_{k+1} - \bar{\boldsymbol{\phi}}_{k+1}^\pi)\Big|S_{k+1}\Big]\Big|S_k = s, A_k = a\Big]
$$

(using $\mathbb{E}_\pi[\mathbb{E}_b[X|S_{k+1}]|S_k = s, A_k = a] = \mathbb{E}_\pi[X|S_k = s, A_k = a]$ for all X not depending on A_{k+1})

$$
= \mathbb{E}_\pi\Big[\delta_k + \gamma_{k+1}\lambda_{k+1}\mathbb{E}_b\Big[\rho_{k+1}\delta_{k+1,t}^{\lambda\rho}\Big|S_{k+1}\Big] + \gamma_{k+1}\lambda_{k+1}\boldsymbol{\theta}^\top(\boldsymbol{\phi}_{k+1} - \bar{\boldsymbol{\phi}}_{k+1}^\pi)\Big|S_k = s, A_k = a\Big]
$$

(using, as in Theorem 3, $\mathbb{E}_b[\rho_k X|S_k = s] = \mathbb{E}_\pi[\mathbb{E}_b[X|A_k = a]|S_k = s]$)

$$
= \mathbb{E}_\pi\Big[\delta_k + \gamma_{k+1}\lambda_{k+1}\mathbb{E}_\pi\Big[\delta_{k+1,t}^{\lambda\rho}\Big|S_{k+1}, A_{k+1}\Big] + \gamma_{k+1}\lambda_{k+1}\boldsymbol{\theta}^\top(\boldsymbol{\phi}_{k+1} - \bar{\boldsymbol{\phi}}_{k+1}^\pi)\Big|S_k = s, A_k = a\Big]
$$

$\vdots$ (repeatedly expand the $\delta^{\lambda\rho}$ term until, finally, $\delta^{\lambda\rho}_{t-1,t} = \delta_{t-1}$)

$$= \mathbb{E}_\pi \left[\sum_{j=k}^{t-1} \left(\prod_{i=k+1}^{j} \gamma_i \lambda_i \right) \left(\delta_j + \boldsymbol{\theta}^\top (\boldsymbol{\phi}_j - \bar{\boldsymbol{\phi}}^\pi_j) \right) - \boldsymbol{\theta}^\top (\boldsymbol{\phi}_k - \bar{\boldsymbol{\phi}}^\pi_k) \middle| S_k = s, A_k = a \right].$$

It thus only remains to show that $\delta^{\lambda 1}_{k,t}$ is equal to the quantity whose expectation is being taken here:

$$\begin{aligned}
\delta^{\lambda 1}_{k,t} &= \delta_k + \gamma_{k+1}\lambda_{k+1}\boldsymbol{\theta}^\top(\boldsymbol{\phi}_{k+1} - \bar{\boldsymbol{\phi}}^\pi_{k+1}) + \gamma_{k+1}\lambda_{k+1}\delta^{\lambda\rho}_{k+1,t} \\
&= \delta_k + \gamma_{k+1}\lambda_{k+1}\delta_{k+1} + \gamma_{k+1}\lambda_{k+1}\boldsymbol{\theta}^\top(\boldsymbol{\phi}_{k+1} - \bar{\boldsymbol{\phi}}^\pi_{k+1}) + \gamma_{k+1}\lambda_{k+1}\gamma_{k+2}\lambda_{k+2}\boldsymbol{\theta}^\top(\boldsymbol{\phi}_{k+1} - \bar{\boldsymbol{\phi}}^\pi_{k+1}) + \gamma_{k+2}\lambda_{k+2}\delta^{\lambda\rho}_{k+2,t} \\
&\ \ \vdots \\
&= \sum_{j=k}^{t-1} \left(\prod_{i=k+1}^{j} \gamma_i\lambda_i \right) \left(\delta_j + \boldsymbol{\theta}^\top(\boldsymbol{\phi}_j - \bar{\boldsymbol{\phi}}^\pi_j) \right) \ - \ \boldsymbol{\theta}^\top(\boldsymbol{\phi}_k - \bar{\boldsymbol{\phi}}^\pi_k).
\end{aligned}$$

The last term is there because $\delta^{\lambda 1}_{t-1,t} = \delta_{t-1}$, so in the summation the indices of the δ range from k to $t-1$, but the indices on the other terms range from $k+1$ to t. □

S.7 Additional detail on the provisional-weight updates (15) and (27)

A key step in the derivation of (15) is the transition from the second to the third equation, involving a re-writing of $\bar{\delta}^t_k$ in terms of $\bar{\delta}^{t-1}_k$. Here we spell it out more fully:

$$\begin{aligned}
\bar{\delta}^t_k &= R_{k+1} + \cdots + R_{t-1} + R_t + \boldsymbol{\theta}^\top\boldsymbol{\phi}_t - \boldsymbol{\theta}^\top\boldsymbol{\phi}_k && \text{(from (4))} \\
&= R_{k+1} + \cdots + R_{t-1} + R_t + \boldsymbol{\theta}^\top\boldsymbol{\phi}_t - \boldsymbol{\theta}^\top\boldsymbol{\phi}_k + \boldsymbol{\theta}^\top\boldsymbol{\phi}_{t-1} - \boldsymbol{\theta}^\top\boldsymbol{\phi}_{t-1} \\
&= \underbrace{R_{k+1} + \cdots + R_{t-1} + \boldsymbol{\theta}^\top\boldsymbol{\phi}_{t-1} - \boldsymbol{\theta}^\top\boldsymbol{\phi}_k}_{\bar{\delta}^{t-1}_k} + \underbrace{R_t + \boldsymbol{\theta}^\top\boldsymbol{\phi}_t - \boldsymbol{\theta}^\top\boldsymbol{\phi}_{t-1}}_{\bar{\delta}^t_{t-1}} && \text{(regrouping)} \\
&= \bar{\delta}^{t-1}_k + \bar{\delta}^t_{t-1}.
\end{aligned}$$

The derivation for PQ's provisional weight update (27) is similar to that for PTD, but was not included in the main text to save space. We include it here:

$$\begin{aligned}
\boldsymbol{u}_t &= \alpha\gamma_t\lambda_t \sum_{k=0}^{t-1} C^{t-1}_k \bar{\delta}^t_k \boldsymbol{\phi}_k \\
&= \alpha\gamma_t\lambda_t \left[\sum_{k=0}^{t-2} C^{t-1}_k \bar{\delta}^t_k \boldsymbol{\phi}_k + C^{t-1}_{t-1}\bar{\delta}^t_{t-1}\boldsymbol{\phi}_{t-1} \right] \\
&= \alpha\gamma_t\lambda_t \left[\sum_{k=0}^{t-2} C^{t-1}_k \left[\bar{\delta}^{t-1}_k + \bar{\delta}^t_{t-1} + \boldsymbol{\theta}^\top(\boldsymbol{\phi}_{t-1} - \bar{\boldsymbol{\phi}}^\pi_{t-1}) \right] \boldsymbol{\phi}_k + \bar{\delta}^t_{t-1}\boldsymbol{\phi}_{t-1} \right] \\
&= \gamma_t\lambda_t \Big(\rho_{t-1}\boldsymbol{u}_{t-1} + \alpha\bar{\delta}^t_{t-1}\boldsymbol{e}_{t-1} + \ \alpha\boldsymbol{\theta}^\top \left(\boldsymbol{\phi}_{t-1} - \bar{\boldsymbol{\phi}}^\pi_{t-1} \right) \left(\boldsymbol{e}_{t-1} - \boldsymbol{\phi}_{t-1} \right) \Big). && (27)
\end{aligned}$$

As in the PTD derivation, the key step is moving from the second to the third equation by writing $\bar{\delta}^t_k$ in terms of $\bar{\delta}^{t-1}_k$, as follows:

$$\begin{aligned}
\bar{\delta}^t_k &= R_{k+1} + \cdots + R_{t-1} + R_t + \boldsymbol{\theta}^\top\bar{\boldsymbol{\phi}}^\pi_t - \boldsymbol{\theta}^\top\boldsymbol{\phi}_k && \text{(from (21))} \\
&= R_{k+1} + \cdots + R_{t-1} + R_t + \boldsymbol{\theta}^\top\bar{\boldsymbol{\phi}}^\pi_t - \boldsymbol{\theta}^\top\boldsymbol{\phi}_k + \boldsymbol{\theta}^\top\bar{\boldsymbol{\phi}}^\pi_{t-1} - \boldsymbol{\theta}^\top\bar{\boldsymbol{\phi}}^\pi_{t-1} + \boldsymbol{\theta}^\top\boldsymbol{\phi}_{t-1} - \boldsymbol{\theta}^\top\boldsymbol{\phi}_{t-1} \\
&= (R_{k+1} + \cdots + R_{t-1} + \boldsymbol{\theta}^\top\bar{\boldsymbol{\phi}}^\pi_{t-1} - \boldsymbol{\theta}^\top\boldsymbol{\phi}_k) + (R_t + \boldsymbol{\theta}^\top\bar{\boldsymbol{\phi}}^\pi_t - \boldsymbol{\theta}^\top\boldsymbol{\phi}_{t-1}) - \boldsymbol{\theta}^\top\bar{\boldsymbol{\phi}}^\pi_{t-1} + \boldsymbol{\theta}^\top\boldsymbol{\phi}_{t-1} && \text{(regrouping)} \\
&= \bar{\delta}^{t-1}_k + \bar{\delta}^t_{t-1} + \boldsymbol{\theta}^\top \left(\boldsymbol{\phi}_{t-1} - \bar{\boldsymbol{\phi}}^\pi_{t-1} \right).
\end{aligned}$$