Appendix

7 A supporting result

We first present an alternative characterisation of the projection operator $\Pi_{\mathcal{C}}$ which will be useful for the analysis that follows. Throughout, for a probability measure $\nu \in \mathscr{P}(\mathbb{R})$, we write F_ν for its CDF.

Proposition 6. *For each $i = 1, \ldots, K$, define $h_{z_i} : \mathbb{R} \to [0, 1]$ to be the (possibly asymmetric) hat function centered in z_i defined by*

$$h_{z_i}(x) = \begin{cases} \frac{z_{i+1}-x}{z_{i+1}-z_i} & \text{for } x \in [z_i, z_{i+1}] \quad \text{and } 1 \le i < K, \\ \frac{x-z_{i-1}}{z_i - z_{i-1}} & \text{for } x \in [z_{i-1}, z_i] \quad \text{and } 1 < i \le K, \\ 1 & \text{for } x \le z_1 \quad \text{and } i = 1, \\ 1 & \text{for } x \ge z_K \quad \text{and } i = K, \\ 0 & \text{otherwise.} \end{cases}$$

Then defining $\Pi_{\mathcal{C}}\nu = \sum_{i=1}^K \mathbb{E}_{w\sim\nu}[h_{z_i}(w)]\delta_{z_i}$ for all probability distributions $\nu \in \mathscr{P}(\mathbb{R})$, is consistent with the earlier definition in (7) for mixtures of Diracs. Further, $F_{\Pi_{\mathcal{C}}\nu}(z_i)$ is equal to the average value of F_ν in the interval $[z_i, z_{i+1}]$, for $i = 1, \ldots, K-1$, and $F_{\Pi_{\mathcal{C}}\nu}(z_K) = 1$.

Proof. The consistency of the definition $\Pi_{\mathcal{C}}\nu = \sum_{i=1}^K \mathbb{E}_{w\sim\nu}[h_{z_i}(w)]\delta_{z_i}$ with (7) follows immediately by observing directly that the definitions agree when ν is a Dirac measure, and then observing that the definition of $\Pi_{\mathcal{C}}$ in the statement of the proposition is also affine.

For the characterisation of $F_{\Pi_{\mathcal{C}}\nu}(z_i)$ for $i = 1, \ldots, K-1$, we note that

$$\begin{aligned} F_{\Pi_{\mathcal{C}}\nu}(z_i) &= \sum_{j=1}^{i} \mathbb{E}_{w\sim\nu}[h_{z_j}(w)] \\ &= \mathbb{E}_{w\sim\nu}\left[\sum_{j=1}^{i} h_{z_j}(w)\right] \\ &= \mathbb{E}_{w\sim\nu}\left[\mathbb{1}_{w \le z_i} + \mathbb{1}_{w\in(z_i, z_{i+1}]}\frac{z_{i+1}-w}{z_{i+1}-z_i}\right] \\ &= \frac{1}{z_{i+1}-z_i}\int_{z_i}^{z_{i+1}} F_\nu(w)\mathrm{d}w \,, \end{aligned}$$

as required. Finally, since $\Pi_{\mathcal{C}}\nu$ is supported on $\{z_1, \ldots, z_K\}$, it immediately follows that $F_{\Pi_{\mathcal{C}}\nu}(z_K) = 1$. □

8 Mixture update version of categorical policy evaluation and categorical Q-learning

Here we give a precise specification of the mixture update versions of categorical policy evaluation and categorical Q-learning, as described in the main paper in Section 4.3. The difference from Algorithm 1 is highlighted in red.

Algorithm 2 CDRL mixture update

Require: $\eta_t^{(x,a)} = \sum_{k=1}^K p_{t,k}^{(x,a)} \delta_{z_k}$ for each (x,a)
Sample transition (x_t, a_t, r_t, x_{t+1})
# Compute distributional Bellman target
if Categorical policy evaluation **then**
  $a^* \sim \pi(\cdot|x_{t+1})$
else if Categorical Q-learning **then**
  $a^* \leftarrow \arg\max_a \mathbb{E}_{R\sim\eta_t^{(x_{t+1},a)}}[R]$
end if
$\widehat{\eta}_*^{(x_t,a_t)} \leftarrow (f_{r_t,\gamma})_\# \eta_t^{(x_{t+1},a^*)}$
Project target onto support
$\widehat{\eta}_t^{(x_t,a_t)} \leftarrow \Pi_\mathcal{C} \widehat{\eta}_*^{(x_t,a_t)}$
Compute mixture update
Generate new estimates according to mixture rule: $\eta_{t+1}^{(x_t,a_t)} = (1-\alpha_t(x_t,a_t))\eta_t^{(x_t,a_t)} + \alpha_t(x_t,a_t)\widehat{\eta}_t^{(x_t,a_t)}$
return η_{t+1}

9 Proof of results in Section 4

Lemma 2. *The operator $\Pi_\mathcal{C}\mathcal{T}^\pi$ is in general not a contraction in $\overline{d}_p$, for $p > 1$.*

Proof. We exhibit a simple counterexample; it is enough to demonstrate that $\Pi_\mathcal{C}$ can act as an expansion. Take $z_1 = 0, z_2 = 1$, and consider two Dirac delta distributions, $\nu_1 = \delta_{1/4}$ and $\nu_2 = \delta_{3/4}$. We have $d_p(\nu_1, \nu_2) = ((1/2)^p)^{1/p} = 1/2$. Now $\Pi_\mathcal{C}\nu_1 = \frac{3}{4}\delta_0 + \frac{1}{4}\delta_1$, and $\Pi_\mathcal{C}\nu_2 = \frac{1}{4}\delta_0 + \frac{3}{4}\delta_1$, and hence $d_p(\Pi_\mathcal{C}\nu_1, \Pi_\mathcal{C}\nu_2) = ((1/2)\times 1^p)^{1/p} = 2^{-1/p} > 1/2$. □

Proposition 1. *The Cramér metric ℓ_2 endows a particular subset of $\mathscr{P}(\mathbb{R})$ with a notion of orthogonal projection, and the orthogonal projection onto the subset $\mathcal{P}$ is exactly the heuristic projection $\Pi_\mathcal{C}$. Consequently, $\Pi_\mathcal{C}$ is a non-expansion with respect to ℓ_2.*

Proof. We begin by setting out a Hilbert space structure of a subset of $\mathscr{P}(\mathbb{R})$. Let $\mathcal{M}(\mathbb{R})$ be the vector space of all finite signed measures on $\mathbb{R}$. First, observe that the following subspace of signed measures:

$$\mathcal{M}_0(\mathbb{R}) = \left\{ \nu \in \mathcal{M}(\mathbb{R}) \middle| \nu(\mathbb{R}) = 0\,, \int_\mathbb{R} F_\nu(x)^2 \mathrm{d}x < \infty \right\}\,,$$

where $F_\nu(x) = \nu((-\infty, x])$ for each $x \in \mathbb{R}$, is isometrically isomorphic to a subspace of the Hilbert space $L^2(\mathbb{R})$ with inner product given by

$$\langle \nu_1, \nu_2 \rangle_{\ell_2} = \int_\mathbb{R} F_{\nu_1}(x) F_{\nu_2}(x) \mathrm{d}x\,. \tag{10}$$

Now consider the affine space $\delta_0 + \mathcal{M}_0(\mathbb{R})$ (i.e. the translation of $\mathcal{M}_0(\mathbb{R})$ in $\mathcal{M}(\mathbb{R})$ by the measure δ_0). This affine space consists of signed measures of total mass 1, with sufficiently quickly decaying tails. In particular, it contains the set of probability measures $\nu \in \mathscr{P}(\mathbb{R})$ satisfying

$$\int_{-\infty}^0 F_\nu(x)^2 \mathrm{d}x < \infty \quad \text{and} \quad \int_0^\infty (1 - F_\nu(x))^2 \mathrm{d}x < \infty\,.$$

As $\delta_0 + \mathcal{M}_0(\mathbb{R})$ is an affine translation of a Hilbert space, it inherits the inner product defined in (10) from $\mathcal{M}_0(\mathbb{R})$, which is now defined for differences of elements. Now consider the affine subspace consisting of measures supported on $\{z_1, \ldots, z_K\}$. It is clear that this is a closed affine subspace (since it is finite-dimensional), and therefore there exists an orthogonal projection (with respect to the inner product defined above) onto this subspace, which we denote by Π. Given a probability measure $\nu \in \delta_0 + \mathcal{M}_0(\mathbb{R})$, $\Pi\nu = \sum_{i=1}^K p_i \delta_{z_i}$, where the p_i satisfy $\sum_{i=1}^K p_i = 1$, and subject to this constraint, minimise $\langle \Pi\nu - \nu, \Pi\nu - \nu \rangle_{\ell_2}$. But note that

$$\langle \Pi\nu - \nu, \Pi\nu - \nu \rangle_{\ell_2} = \int_\mathbb{R} (F_{\Pi\nu}(x) - F_\nu(x))^2 \mathrm{d}x\,. \tag{11}$$

By construction, $F_{\Pi\nu}$ is constant on the open intervals (z_i, z_{i+1}) for $i = 1, \ldots, K-1$, and also on the intervals $(-\infty, z_1)$ and $(z_K, +\infty)$. Therefore $F_{\Pi\nu}$, and hence $\Pi\nu$ itself, is determined by the values of $F_{\Pi\nu}(z_i)$ for $i = 1, \ldots, K$. The optimal values (i.e. those minimising (11)) are easily verified to be: $F_{\Pi\nu}(z_K) = 1$, and $F_{\Pi\nu}(z_i)$ is equal to the average of F_ν on the interval (z_i, z_{i+1}), for $i = 1, \ldots, K-1$. Note then that $\Pi\nu$ is a probability distribution (since $F_{\Pi\nu}$ is non-decreasing), and in fact matches the characterisation of $\Pi_{\mathcal{C}}\nu$ obtained in Proposition 6. Therefore we have established that $\Pi_{\mathcal{C}}$ is exactly orthogonal projection in the affine Hilbert space $\delta_0 + \mathcal{M}_0(\mathbb{R})$. Further, we have verified that the norm between elements in the space is exactly ℓ_2, and hence it follows that $\Pi_{\mathcal{C}}$ is a non-expansion with respect to ℓ_2. □

Lemma 3 (Pythagorean theorem)**.** *Let $\mu \in \mathscr{P}([z_1, z_K])$, and let $\nu \in \mathscr{P}(\{z_1, \ldots, z_K\})$. Then*

$$\ell_2^2(\mu, \nu) = \ell_2^2(\mu, \Pi_{\mathcal{C}}\mu) + \ell_2^2(\Pi_{\mathcal{C}}\mu, \nu)\,.$$

Proof. Denote by F_μ, $F_{\Pi_{\mathcal{C}}\mu}$ and F_ν the CDFs of the measures μ, $\Pi_{\mathcal{C}}\mu$ and ν respectively. Now note

$$\begin{aligned}
\ell_2^2(\mu, \nu) &= \int_{z_1}^{z_K} (F_\mu(x) - F_\nu(x))^2 \mathrm{d}x \\
&= \int_{z_1}^{z_K} (F_\mu(x) - F_{\Pi_{\mathcal{C}}\mu}(x) + F_{\Pi_{\mathcal{C}}\mu}(x) - F_\nu(x))^2 \mathrm{d}x \\
&= \int_{z_1}^{z_K} (F_\mu(x) - F_{\Pi_{\mathcal{C}}\mu}(x))^2 \mathrm{d}x + \int_{z_1}^{z_K} (F_\nu(x) - F_{\Pi_{\mathcal{C}}\mu}(x))^2 \mathrm{d}x \\
&\qquad - 2\int_{z_1}^{z_K} (F_\mu(x) - F_{\Pi_{\mathcal{C}}\mu}(x))(F_\nu(x) - F_{\Pi_{\mathcal{C}}\mu}(x)) \mathrm{d}x\,.
\end{aligned}$$

Finally, observe that

$$\begin{aligned}
&\int_{z_1}^{z_K} (F_\mu(x) - F_{\Pi_{\mathcal{C}}\mu}(x))(F_\nu(x) - F_{\Pi_{\mathcal{C}}\mu}(x)) \mathrm{d}x \\
=& \sum_{k=1}^{K-1} (F_\nu(z_k) - F_{\Pi_{\mathcal{C}}\mu}(z_k)) \int_{z_k}^{z_{k+1}} (F_\mu(x) - F_{\Pi_{\mathcal{C}}\mu}(x)) \mathrm{d}x \\
=& 0\,,
\end{aligned}$$

since by Proposition 6, $F_{\Pi_{\mathcal{C}}\mu}$ is constant on (z_k, z_{k+1}), and is equal to the average of F_μ on the same interval. □

Proposition 2. *The operator $\Pi_{\mathcal{C}}\mathcal{T}^\pi$ is a $\sqrt{\gamma}$-contraction in $\overline{\ell}_2$. Further, there is a unique distribution function $\eta_{\mathcal{C}} \in \mathcal{P}^{\mathcal{X}\times\mathcal{A}}$ such that given any initial distribution function $\eta_0 \in \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$, we have*

$$(\Pi_{\mathcal{C}}\mathcal{T}^\pi)^m \eta_0 \to \eta_{\mathcal{C}} \text{ in } \overline{\ell}_2 \text{ as } m \to \infty\,.$$

Proof. First, we show that the true distributional Bellman operator $\mathcal{T}^\pi$ is a $\sqrt{\gamma}$-contraction in $\overline{\ell}_2$. Note that through notions of scale sensitivity, as discussed by Bellemare et al. [2017b], the ideas here may be extended to other distances over probability measures. Let $\eta, \mu \in \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$. Then

$$\begin{aligned}
\ell_2^2((\mathcal{T}^\pi\eta)^{(x,a)}, (\mathcal{T}^\pi\mu)^{(x,a)}) =& \ell_2^2\Bigg(\int_{\mathbb{R}} \sum_{(x',a')\in\mathcal{X}\times\mathcal{A}} \pi(a'|x')p(\mathrm{d}r, x'|x,a)(f_{r,\gamma})_\# \eta^{(x',a')}, \\
& \qquad \int_{\mathbb{R}} \sum_{(x',a')\in\mathcal{X}\times\mathcal{A}} \pi(a'|x')p(\mathrm{d}r, x'|x,a)(f_{r,\gamma})_\# \mu^{(x',a')}\Bigg) \\
\leq& \int_{\mathbb{R}} \sum_{(x',a')\in\mathcal{X}\times\mathcal{A}} \pi(a'|x')p(\mathrm{d}r, x'|x,a)\ell_2^2((f_{r,\gamma})_\# \eta^{(x',a')}, (f_{r,\gamma})_\# \mu^{(x',a')}) \\
=& \int_{\mathbb{R}} \sum_{(x',a')\in\mathcal{X}\times\mathcal{A}} \pi(a'|x')p(\mathrm{d}r, x'|x,a)\gamma\ell_2^2(\eta^{(x',a')}, \mu^{(x',a')}) \\
\leq& \gamma\overline{\ell}_2^2(\eta, \mu)\,,
\end{aligned}$$

with the first inequality following from Jensen's inequality, and the equality coming from the follow general fact about the Cramér distance and probability measures $\nu_1, \nu_2 \in P(\mathbb{R})$:

$$\begin{aligned}
\ell_2^2((f_{r,\gamma})_\# \nu_1, (f_{r,\gamma})_\# \nu_2) &= \int_{\mathbb{R}} (F_{(f_{r,\gamma})_\# \nu_1}(t) - F_{(f_{r,\gamma})_\# \nu_2}(t))^2 dt \\
&= \int_{\mathbb{R}} (F_{\nu_1}(f_{r,\gamma}^{-1}(t)) - F_{\nu_2}(f_{r,\gamma}^{-1}(t)))^2 dt \\
&= \int_{\mathbb{R}} \left(F_{\nu_1}\left(\frac{t-r}{\gamma}\right) - F_{\nu_2}\left(\frac{t-r}{\gamma}\right)\right)^2 dt \\
&= \gamma \int_{\mathbb{R}} \left(F_{\nu_1}(t') - F_{\nu_2}(t')\right)^2 dt' \\
&= \gamma \ell_2^2(\nu_1, \nu_2) \, .
\end{aligned}$$

Now by Proposition 1, $\Pi_{\mathcal{C}}$ is a non-expansion in $\overline{\ell}_2$. Therefore $\Pi_{\mathcal{C}}\mathcal{T}^\pi$ is the composition of a non-expansion in $\overline{\ell}_2$ with a $\sqrt{\gamma}$-contraction in $\overline{\ell}_2$, and is therefore itself a $\sqrt{\gamma}$-contraction in $\overline{\ell}_2$. The second claim of the proposition then follows immediately from the Banach fixed point theorem. □

Proposition 3. *Let $\eta_{\mathcal{C}}$ be the limiting return distribution function of Proposition 2. If $\eta_\pi^{(x,a)}$ is supported on $[z_1, z_K]$ for all $(x,a) \in \mathcal{X} \times \mathcal{A}$, then we have:*

$$\overline{\ell}_2^2(\eta_{\mathcal{C}}, \eta_\pi) \le \frac{1}{1-\gamma} \max_{1 \le i < K} (z_{i+1} - z_i) \, .$$

Proof. By Lemma 3, we have:

$$\begin{aligned}
\overline{\ell}_2^2(\eta_{\mathcal{C}}, \eta_\pi) &= \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} \ell_2^2(\eta_{\mathcal{C}}^{(x,a)}, \eta_\pi^{(x,a)}) \\
&= \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} \left[\ell_2^2(\eta_{\mathcal{C}}^{(x,a)}, (\Pi_{\mathcal{C}}\eta_\pi)^{(x,a)}) + \ell_2^2((\Pi_{\mathcal{C}}\eta_\pi)^{(x,a)}, \eta_\pi^{(x,a)}) \right] \\
&\le \overline{\ell}_2^2(\eta_{\mathcal{C}}, \Pi_{\mathcal{C}}\eta_\pi) + \overline{\ell}_2^2(\Pi_{\mathcal{C}}\eta_\pi, \eta_\pi) \\
&= \overline{\ell}_2^2(\Pi_{\mathcal{C}}\mathcal{T}^\pi\eta_{\mathcal{C}}, \Pi_{\mathcal{C}}\mathcal{T}^\pi\eta_\pi) + \overline{\ell}_2^2(\Pi_{\mathcal{C}}\eta_\pi, \eta_\pi) \\
&\le \gamma\overline{\ell}_2^2(\eta_{\mathcal{C}}, \eta_\pi) + \overline{\ell}_2^2(\Pi_{\mathcal{C}}\eta_\pi, \eta_\pi) \, ,
\end{aligned} \tag{12}$$

where in the final line we have used the contractivity of $\Pi_{\mathcal{C}}\mathcal{T}^\pi$ under $\overline{\ell}_2$ from Proposition 2. Due to Proposition 6 (see Section 7) we have that $F_{\Pi_{\mathcal{C}}\eta_\pi^{(x,a)}}$ is constant on the intervals (z_i, z_{i+1}) for $i = 1, \ldots, K-1$, and moreover, due to the formula for the mass placed at the locations $z_{1:K}$, we also have

$$F_{\Pi_{\mathcal{C}}\eta_\pi^{(x,a)}}(z_i) \in [F_{\eta_\pi^{(x,a)}}(z_i), F_{\eta_\pi^{(x,a)}}(z_{i+1})] \quad \text{for } i = 1, \ldots, K-1 \quad , F_{\Pi_{\mathcal{C}}\eta_\pi^{(x,a)}}(z_K) = 1 \, .$$

Therefore,

$$\begin{aligned}
\ell_2^2(\Pi_{\mathcal{C}}\eta_\pi^{(x,a)}, \eta_\pi^{(x,a)}) &\le \sum_{i=1}^{K-1} (z_{i+1} - z_i)(F_{\eta_\pi^{(x,a)}}(z_{i+1}) - F_{\eta_\pi^{(x,a)}}(z_i))^2 \\
&\le \left[\sup_{1\le i<K} (z_{i+1} - z_i)\right] \sum_{i=1}^{K-1} (F_{\eta_\pi^{(x,a)}}(z_{i+1}) - F_{\eta_\pi^{(x,a)}}(z_i))^2 \\
&\le \left[\sup_{1\le i<K} (z_{i+1} - z_i)\right] \left[\sum_{i=1}^{K-1} (F_{\eta_\pi^{(x,a)}}(z_{i+1}) - F_{\eta_\pi^{(x,a)}}(z_i))\right]^2 \\
&\le \sup_{1\le i<K} (z_{i+1} - z_i) \, ,
\end{aligned}$$

for each $(x,a) \in \mathcal{X} \times \mathcal{A}$, yielding

$$\overline{\ell}_2^2(\Pi_{\mathcal{C}}\eta_\pi, \eta_\pi) \le \sup_{1\le i<K} (z_{i+1} - z_i) \, .$$

Thus, taking (12), applying the upper bound on $\overline{\ell}_2^2(\Pi_{\mathcal{C}}\eta_\pi, \eta_\pi)$ and rearranging, we obtain

$$\overline{\ell}_2^2(\eta_{\mathcal{C}}, \eta_\pi) \leq \frac{1}{1-\gamma} \sup_{1 \leq i < K} (z_{i+1} - z_i) \,.$$

$\square$

Proposition 4. *Let $\eta_{\mathcal{C}}$ be the limiting return distribution function of Proposition 2. Suppose $\eta_\pi^{(x,a)}$ is supported on an interval $[z_1 - \delta, z_K + \delta]$ containing $[z_1, z_K]$ for each $(x,a) \in \mathcal{X} \times \mathcal{A}$, and $\eta_\pi^{(x,a)}([z_1 - \delta, z_1] \cup [z_K, z_K + \delta]) \leq q$ for some $q \in \mathbb{R}$ and for all $(x,a) \in \mathcal{X} \times \mathcal{A}$ – q bounds the excess mass lying outside the region $[z_1, z_K]$. Then we have*

$$\overline{\ell}_2^2(\eta_{\mathcal{C}}, \eta_\pi) \leq \frac{1}{1-\gamma} \left(\max_{1 \leq i < K} (z_{i+1} - z_i) + 2q^2\delta \right) \,.$$

Proof. The proof proceeds as for that of Proposition 3, obtaining the inequality

$$\overline{\ell}_2^2(\eta_{\mathcal{C}}, \eta_\pi) \leq \frac{1}{1-\gamma} \overline{\ell}_2^2(\Pi_{\mathcal{C}}\eta_\pi, \eta_\pi) \,.$$

We now bound the right-hand side as follows:

$$\begin{aligned}
\ell_2^2(\Pi_{\mathcal{C}}\eta_\pi^{(x,a)}, \eta_\pi^{(x,a)}) &\leq q^2 \times (z_1 - (z_1 - \delta)) + q^2((z_K + \delta) - z_K) + \sum_{i=1}^{K-1} (z_{i+1} - z_i)(F_{\eta_\pi^{(x,a)}}(z_{i+1}) - F_{\eta_\pi^{(x,a)}}(z_i))^2 \\
&\leq 2q^2\delta + \sup_{1 \leq i < K} (z_{i+1} - z_i) \,,
\end{aligned}$$

which yields the result as required. $\square$

Proposition 5. *The distributional Bellman operator $\mathcal{T}^\pi : \mathscr{P}(\mathbb{R})^{\mathcal{X} \times \mathcal{A}} \rightarrow \mathscr{P}(\mathbb{R})^{\mathcal{X} \times \mathcal{A}}$ is a monotone map with respect to the partial ordering on $\mathscr{P}(\mathbb{R})^{\mathcal{X} \times \mathcal{A}}$ given by element-wise stochastic dominance. Further, the Cramér projection $\Pi_{\mathcal{C}} : \mathscr{P}(\mathbb{R})^{\mathcal{X} \times \mathcal{A}} \rightarrow \mathscr{P}(\mathbb{R})^{\mathcal{X} \times \mathcal{A}}$ is a monotone map, from which it follows that the Cramér-Bellman operator $\Pi_{\mathcal{C}}\mathcal{T}^\pi$ is also monotone.*

Proof. Let $\eta, \mu \in \mathscr{P}(\mathbb{R})^{\mathcal{X} \times \mathcal{A}}$, and suppose that $\eta \leq \mu$. This is equivalent to $F_{\eta^{(x,a)}} \geq F_{\mu^{(x,a)}}$ pointwise, for each $(x,a) \in \mathcal{X} \times \mathcal{A}$. We now compute the CDFs of $(\mathcal{T}^\pi\eta)^{(x,a)}$ and $(\mathcal{T}^\pi\mu)^{(x,a)}$, for each $(x,a) \in \mathscr{P}(\mathbb{R})^{\mathcal{X} \times \mathcal{A}}$, and show that stochastic dominance still holds. Indeed, by conditioning on the value of the tuple (r, x', a'), we obtain, for each

$$\begin{aligned}
(\mathcal{T}^\pi\eta)^{(x,a)}((-\infty, y]) &= \sum_{(x',a') \in \mathcal{X} \times \mathcal{A}} \int_{\mathbb{R}} p(\mathrm{d}r, x'|x,a)\pi(a'|x')(f_{r,\gamma})_\# \eta^{(x',a')}((-\infty, y]) \\
&= \sum_{(x',a') \in \mathcal{X} \times \mathcal{A}} \int_{\mathbb{R}} p(\mathrm{d}r, x'|x,a)\pi(a'|x') \eta^{(x',a')}((-\infty, (y-r)/\gamma]) \\
&\geq \sum_{(x',a') \in \mathcal{X} \times \mathcal{A}} \int_{\mathbb{R}} p(\mathrm{d}r, x'|x,a)\pi(a'|x') \mu^{(x',a')}((-\infty, (y-r)/\gamma]) \\
&= \sum_{(x',a') \in \mathcal{X} \times \mathcal{A}} \int_{\mathbb{R}} p(\mathrm{d}r, x'|x,a)\pi(a'|x')(f_{r,\gamma})_\# \mu^{(x',a')}((-\infty, y]) \\
&= (\mathcal{T}^\pi\mu)^{(x,a)}((-\infty, y]) \,,
\end{aligned}$$

as required, with the inequality coming from the fact that $\mu^{(x',a')}$ stochastically dominates $\eta^{(x',a')}$. This concludes the proof that the distributional Bellman operator $\mathcal{T}^\pi$ is monotone with respect to the partial order of element-wise stochastic dominance.

The monotonocity of the Cramér projection $\Pi_{\mathcal{C}}$ may be established from the expression given for the projection in Proposition 6. Suppose we have two distributions $\nu_1, \nu_2 \in \mathscr{P}(\mathbb{R})$, and suppose further that $\nu_1 \leq \nu_2$. Then recall from Proposition 6 that we have $F_{\Pi_\mathcal{C}\nu_1}(w)$ and $F_{\Pi_\mathcal{C}\nu_2}(w)$ equal to 0 for $w < z_1$ and equal to 1 for $w \geq z_K$. For $w \in [z_i, z_{i+1})$ for some $i \in \{1, \ldots, K-1\}$, recall again from Proposition 6 that we have

$$F_{\Pi_\mathcal{C}\nu_j}(w) = \frac{1}{z_{i+1} - z_i} \int_{z_i}^{z_{i+1}} F_{\nu_j}(t) \mathrm{d}t \,, \quad \text{for } j = 1, 2 \,. \tag{13}$$

Since by assumption we have $F_{\nu_1} \geq F_{\nu_2}$ pointwise, it follows from (13) that $F_{\Pi_\mathcal{C}\nu_1} \geq F_{\Pi_\mathcal{C}\nu_2}$ pointwise, and therefore $\Pi_\mathcal{C}\nu_1 \leq \Pi_\mathcal{C}\nu_2$, as required. □

9.1 Proof of Theorem 1

Theorem 1. *In the context of policy evaluation for some policy π, suppose that:*

(i) the stepsizes $(\alpha_t(x,a) | t \geq 0, (x,a) \in \mathcal{X} \times \mathcal{A})$ satisfy the Robbins-Monro conditions:
- $\sum_{t=0}^\infty \alpha_t(x,a) = \infty$
- $\sum_{t=0}^\infty \alpha_t^2(x,a) < C < \infty$

almost surely, for all $(x,a) \in \mathcal{X} \times \mathcal{A}$;

(ii) we have initial estimates $\eta_0^{(x,a)}$ of the distribution of returns for each state-action pair $(x,a) \in \mathcal{X} \times \mathcal{A}$, each with support contained in $[z_1, z_K]$.

Then, for the updates given by Algorithm 2, in the case of evaluation of the policy π, we have almost sure convergence of η_t to $\eta_\mathcal{C}$ in $\overline{\ell}_2$, where $\eta_\mathcal{C}$ is the limiting return distribution function of Proposition 2. That is,

$$\overline{\ell}_2(\eta_t, \eta_\mathcal{C}) \to 0 \text{ as } t \to \infty \quad \text{almost surely.}$$

The proof structure is based on that of Theorem 2 of Tsitsiklis [1994]; our Lemmas 5 and 6 are variants of Lemmas 5 and 6 of Tsitsiklis [1994]. The high-level argument of the proof proceeds as follows.

Define:

$$U_0^{(x,a)} = \delta_{z_K} \,, \qquad L_0^{(x,a)} = \delta_{z_1}$$
$$U_{k+1}^{(x,a)} = \frac{1}{2} U_k^{(x,a)} + \frac{1}{2} (\Pi_\mathcal{C} \mathcal{T}^\pi U_k)^{(x,a)} \,, \qquad L_{k+1}^{(x,a)} = \frac{1}{2} L_k^{(x,a)} + \frac{1}{2} (\Pi_\mathcal{C} \mathcal{T}^\pi L_k)^{(x,a)} \,,$$

iteratively for each $(x,a) \in \mathcal{X} \times \mathcal{A}$.

Lemma 5. *We have $U_{k+1} \leq U_k$, for each $k \in \mathbb{N}_0$, and $L_{k+1} \geq L_k$, for each $k \in \mathbb{N}_0$. Further, we have $U_k \to \eta_\mathcal{C}$ in $\overline{\ell}_2$, and also $L_k \to \eta_\mathcal{C}$ in $\overline{\ell}_2$.*

Finally, we argue that, for each $k \in \mathbb{N}_0$, the return distribution functions U_k and L_k sandwich all but finitely many of the return distribution estimators η_t, in a sense made precise by the following lemma.

Lemma 6. *Given $k \in \mathbb{N}_0$, there exists a random time T_k taking values in $\mathbb{N}_0$ such that*

$$L_k \leq \eta_t \leq U_k \quad \text{for all } t > T_k, \text{ almost surely.}$$

Now, from Lemma 6 the conclusion of Theorem 1 is reached as follows. Let $\varepsilon > 0$, and pick $k \in \mathbb{N}_0$ sufficiently large so that $\overline{\ell}_2(L_k, \eta_\mathcal{C}), \overline{\ell}_2(U_k, \eta_\mathcal{C}) < \varepsilon$, which can be done by Lemma 5. Note then by the triangle inequality that $\overline{\ell}_2(U_k, L_k) < 2\varepsilon$, and further, we have:

$$\overline{\ell_2}(\eta_t, \eta_\mathcal{C}) \leq \overline{\ell}_2(\eta_t, L_k) + \overline{\ell}_2(L_k, U_k) + \overline{\ell}_2(U_k, \eta_\mathcal{C}) \,.$$

Since, by Lemma 6, we have that $L_k \leq \eta_t \leq U_k$ for all $t > T_k$ almost surely, it follows that $\overline{\ell}_2(\eta_t, L_k) \leq \overline{\ell}_2(L_k, U_k)$ for all $t > T_k$ almost surely, and so we obtain

$$\overline{\ell_2}(\eta_t, \eta_\mathcal{C}) \leq 2\overline{\ell}_2(L_k, U_k) + \overline{\ell}_2(U_k, \eta_\mathcal{C}) < 5\varepsilon \text{ for all } t > T_k \text{ almost surely} \,,$$

which yields the statement of Theorem 1. It now remains to establish Lemmas 5 and 6.

9.2 Proof of Lemma 5

We firstly show that $U_{k+1} \leq U_k$ for each $k \in \mathbb{N}_0$. The proof that $L_{k+1} \geq L_k$ for each $k \in \mathbb{N}_0$ is entirely analogous.

First, observe that $U_1 \leq U_0$, since each distribution $U_1^{(x,a)}$ is supported on $[z_1, z_K]$, and $U_0^{(x,a)}$ was chosen to stochastically dominate all distributions supported on $[z_1, z_K]$. For the inductive step, suppose $U_{k+1} \leq U_k$ for some $k \in \mathbb{N}_0$. Then by monotonicity of $\Pi_\mathcal{C}\mathcal{T}^\pi$, we have $\Pi_\mathcal{C}\mathcal{T}^\pi U_{k+1} \leq \Pi_\mathcal{C}\mathcal{T}^\pi U_k$. Hence,

$$U_{k+2}^{(x,a)} = \frac{1}{2}U_{k+1}^{(x,a)} + \frac{1}{2}(\Pi_\mathcal{C}\mathcal{T}^\pi U_{k+1})^{(x,a)} \leq \frac{1}{2}U_k^{(x,a)} + \frac{1}{2}(\Pi_\mathcal{C}\mathcal{T}^\pi U_k)^{(x,a)} = U_{k+1}^{(x,a)} \, ,$$

which completes the inductive proof. To establish convergence of U_k to $\eta_\mathcal{C}$, we make use of the following general result.

Lemma 7. *Let $(\nu_k)_{k=0}^\infty$ be a sequence of probability measures over $\{z_1, \ldots, z_K\}$, with the property that $\nu_{k+1} \leq \nu_k$ for each $k \in \mathbb{N}_0$. Then there exists a probability measure ν^* over $\{z_1, \ldots, z_K\}$ such that $\nu_k \to \nu^*$ in ℓ_2.*

Proof. We work with CDFs. Denote the CDF of ν_k by F_k, for $k \in \mathbb{N}_0$. Recall that the stochastic dominance condition $\nu_{k+1} \leq \nu_k$ implies that $F_{k+1} \geq F_k$ pointwise. Therefore for each $x \in \mathbb{R}$, we have that $(F_k(x))_{k\in\mathbb{N}_0}$ is an increasing sequence, trivially upper-bounded by 1. Therefore the sequence converges, and so there exists a limit function $F : \mathbb{R} \to \mathbb{R}$, defined by $F^*(x) = \lim_{k\to\infty} F_k(x)$. It is straightforward to see that this limit function takes values in $[0, 1]$, is non-decreasing, right-continuous and is constant away from the set $\{z_1, \ldots, z_K\}$. It is therefore the CDF of a probability distribution ν^* supported on $\{z_1, \ldots, z_K\}$. Since $\widetilde{F}^*$ is constant away from $\{z_1, \ldots, z_K\}$, ν^* is supported on $\{z_1, \ldots, z_K\}$. To show that $\nu_k \to \nu^*$ in ℓ_2, we must establish that $\int_\mathbb{R}(F_k(x) - F^*(x))^2 \mathrm{d}x \to 0$. Since $\nu^* \leq \nu_{k+1} \leq \nu_k$ for each $k \in \mathbb{N}_0$, it follows that $\int_\mathbb{R}(F_k(x) - F^*(x))^2 \mathrm{d}x$ is a non-increasing sequence, and so it suffices to show that it is not lower-bounded by a positive number to establish the sequence's convergence to 0. To that end, let $\varepsilon > 0$. Pick $k \in \mathbb{N}_0$ such that $|F_k(z_i) - F^*(z_i)| < \varepsilon$, for each $i = 1, \ldots, K-1$. Then observe that

$$\int_\mathbb{R}(F_k(x) - F^*(x))^2 \mathrm{d}x \leq \sum_{i=1}^{K-1}(z_{i+1} - z_i)\varepsilon^2 \, ,$$

which demonstrates that no positive lower bounded exists, as required. □

Applying Lemma 7 to each of the sequences $(U_k^{(x,a)})_{k=0}^\infty$, for each state-action pair $(x, a) \in \mathcal{X} \times \mathcal{A}$, we obtain the convergence of $(U_k)_{k=0}^\infty$ to some set of return distributions η^* in $\overline{\ell}_2$. Finally, due to the continuity of $\Pi_\mathcal{C}\mathcal{T}^\pi$ with respect to $\overline{\ell}_2$, this limiting set of return distributions η^* must satisfy $\eta^* = \frac{1}{2}\eta^* + \frac{1}{2}\Pi_\mathcal{C}\mathcal{T}^\pi\eta^*$, implying that $\eta^* = \Pi_\mathcal{C}\mathcal{T}^\pi\eta^*$, so the limiting set of return distributions is indeed the fixed point $\eta_\mathcal{C}$ of $\Pi_\mathcal{C}\mathcal{T}^\pi$. Analogously, we may show that $L_k \to \eta_\mathcal{C}$ in $\overline{\ell}_2$.

9.3 Proof of Lemma 6

We prove this lemma by induction. The result is clear for $k = 0$, as in this case $U_0^{(x,a)}$ stochastically dominates all distributions supported on $[z_1, z_K]$, and $L_0^{(x,a)}$ is stochastically dominated by all distributions supported on $[z_1, z_K]$. Now assume the result holds for some $k \geq 0$; that is, there exists some random time T_k such that $L_k \leq \eta_t \leq U_k$ for all $t \geq T_k$ almost surely. Here, we follow the structure of the proof of Lemma 6 of [Tsitsiklis, 1994] closely. We will show there exists a random time T_{k+1} such that $\eta_t \leq U_{k+1}$ for all $t \geq T_{k+1}$ almost surely; the claim that $L_{k+1} \leq \eta_t$ for all $t \geq T_{k+1}$ may be proven analogously.

Now define

$$H_{T_k}^{(x,a)} = U_k^{(x,a)} \, , \; H_{t+1}^{(x,a)} = (1 - \alpha_t(x, a))H_t^{(x,a)} + \alpha_t(x, a)(\Pi_\mathcal{C}\mathcal{T}^\pi U_k)^{(x,a)} \, , \text{ for } t \geq T_k \tag{14}$$

$$W_{T_k}^{(x,a)} = 0 \in \mathcal{M}(\mathbb{R}) \, , \; W_{t+1}^{(x,a)} = (1 - \alpha_t(x, a))W_t^{(x,a)} + \alpha_t(x, a)\left[(\Pi_\mathcal{C}(f_{r,\gamma})_\# \eta_t)^{(x',a')} - (\Pi_\mathcal{C}\mathcal{T}^\pi\eta_t)^{(x,a)}\right] , \text{for } t \geq T_k \, ,$$

where $\mathcal{M}(\mathbb{R})$ is the space of signed measures on $\mathbb{R}$, and $0 \in \mathcal{M}(\mathbb{R})$ represents the zero measure; that is, the signed measure that assigns measure 0 to every Borel subset of $\mathbb{R}$. Note that the process $(W_t)_{t\geq T_k}$ takes values in the space of collections of finite signed measures indexed by state-action pairs, each with overall mass 0; that is, $W_t^{(x,a)}(\mathbb{R}) = 0$ for all $(x, a) \in \mathcal{X} \times \mathcal{A}$, for all $t \geq T_k$.

We now argue that $\eta_t^{(x,a)} \le H_t^{(x,a)} + W_t^{(x,a)}$ for all $t \ge T_k$ and for all $(x,a) \in \mathcal{X}\times\mathcal{A}$ almost surely. For $t = T_k$, this following from the definitions in (14) and the dominance relation $\eta_{T_k} \le U_k$. To complete the proof, we proceed inductively. Suppose that $\eta_t^{(x,a)} \le H_t^{(x,a)} + W_t^{(x,a)}$ for all $(x,a) \in \mathcal{X}\times\mathcal{A}$, for some $t \ge T_k$. Then note, assuming $\alpha_t(x,a) = 0$ if the distribution corresponding to the state-action pair (x,a) is not updated at time t, we have

$$\begin{aligned}
\eta_{t+1}^{(x,a)} =& (1-\alpha_t(x,a))\eta_t^{(x,a)} + \alpha_t(x,a)\Pi_{\mathcal{C}}(f_{r,\gamma})_\#\eta_t^{(x',a')} \\
=&(1-\alpha_t(x,a))\eta_t^{(x,a)} + \alpha_t(x,a)(\Pi_{\mathcal{C}}\mathcal{T}^\pi\eta_t)^{(x,a)} + \alpha_t(x,a)(\Pi_{\mathcal{C}}(f_{r,\gamma})_\#\eta_t^{(x',a')} - (\Pi_{\mathcal{C}}\mathcal{T}^\pi\eta_t)^{(x,a)}) \\
\overset{(i)}{\le}&(1-\alpha_t(x,a))(H_t^{(x,a)} + W_t^{(x,a)}) + \alpha_t(x,a)(\Pi_{\mathcal{C}}\mathcal{T}^\pi U_k)^{(x,a)} + \alpha_t(x,a)(\Pi_{\mathcal{C}}(f_{r,\gamma})_\#\eta_t^{(x',a')} - (\Pi_{\mathcal{C}}\mathcal{T}^\pi\eta_t)^{(x,a)}) \\
=&(1-\alpha_t(x,a))H_t^{(x,a)} + \alpha_t(x,a)(\Pi_{\mathcal{C}}\mathcal{T}^\pi U_k)^{(x,a)} + (1-\alpha_t(x,a))W_t^{(x,a)} \\
&\qquad + \alpha_t(x,a)(\Pi_{\mathcal{C}}(f_{r,\gamma})_\#\eta_t^{(x',a')} - (\Pi_{\mathcal{C}}\mathcal{T}^\pi\eta_t)^{(x,a)}) \\
=&H_{t+1}^{(x,a)} + W_{t+1}^{(x,a)} \, ,
\end{aligned}$$

as required. In the above derivation, (i) comes from the stochastic dominance relations $\eta_t \le H_t + W_t$ (by induction hypothesis) and $\eta_t \le U_k$ and the monotonicity of $\Pi_{\mathcal{C}}\mathcal{T}^\pi$. Note that we have the following expression for $H_t^{(x,a)}$:

$$H_t^{(x,a)} = \left(\prod_{\tau=T_k}^{t-1}(1-\alpha_\tau(x,a))\right)U_k + \left(1 - \prod_{\tau=T_k}^{t-1}(1-\alpha_\tau(x,a))\right)(\Pi_{\mathcal{C}}\mathcal{T}^\pi U_k)^{(x,a)}$$

Since by assumption we have $\sum_{k=0}^\infty \alpha_k(x,a) = \infty$ for all $(x,a) \in \mathcal{X}\times\mathcal{A}$ almost surely, we have that there exists a random time $\widetilde{T}_{k+1}$ such that $\prod_{\tau=T_k}^{t-1}(1-\alpha_\tau(x,a)) \le 1/4$ for all $(x,a) \in \mathcal{X}\times\mathcal{A}$, and for all $t \ge \widetilde{T}_{k+1}$ almost surely. Since $\Pi_{\mathcal{C}}\mathcal{T}^\pi U_k \le U_k$, for all $t \ge \widetilde{T}_k$, we have:

$$\begin{aligned}
\eta_t &\le H_t + W_t \\
&\le \frac{1}{4}U_k + \frac{3}{4}\Pi_{\mathcal{C}}\mathcal{T}^\pi U_k + W_t \\
&= \frac{1}{2}U_k + \frac{1}{2}\Pi_{\mathcal{C}}\mathcal{T}^\pi U_k + W_t - \frac{1}{4}(U_k - \Pi_{\mathcal{C}}\mathcal{T}^\pi U_k) \\
&= U_{k+1} + W_t - \frac{1}{4}(U_k - \Pi_{\mathcal{C}}\mathcal{T}^\pi U_k)\,. \qquad (15)
\end{aligned}$$

Now note that if $U_k^{(x,a)}((\infty, z_i]) = \Pi_{\mathcal{C}}\mathcal{T}^\pi U_k^{(x,a)}((\infty, z_i])$, then we have $U_{k+1}^{(x,a)}((-\infty, z_i]) = U_k^{(x,a)}((-\infty, z_i])$. Let δ, then, be the smallest non-zero value of $|(\Pi_{\mathcal{C}}\mathcal{T}^\pi U_k)^{(x,a)}((-\infty, z_i]) - U_k^{(x,a)}((-\infty, z_i])|$ across all state-action pairs $(x,a) \in \mathcal{X}\times\mathcal{A}$ and all support points $z_i \in \{z_1, \ldots, z_K\}$. Crucially, we observe that the additive "noise" term appearing in the definition of $W_{t+1}^{(x,a)}$ in Equation (14) is mean-zero, in the following sense: as a random measure, the expectation of the noise term is the 0 measure. More concretely for our purposes, we have, as stated in Lemma 4 in the main paper, for all $z_i \in \{z_1, \ldots, z_K\}$:

$$\mathbb{E}_{r,x',a'}\left[((\Pi_{\mathcal{C}}(f_{r,\gamma})_\#\eta_t)^{(x',a')} - (\Pi_{\mathcal{C}}\mathcal{T}^\pi\eta_t)^{(x,a)})\right]((-\infty, z_i]) = 0\,.$$

Standard stochastic approximation theory (e.g. [Tsitsiklis, 1994]), via Assumption (i), then yields that $W_t^{(x,a)}((-\infty, z_i]) \to 0$ almost surely, for all $(x,a) \in \mathcal{X}\times\mathcal{A}$, and for all $z_i \in \{z_1, \ldots, z_K\}$. We can now take $T_{k+1} > \widetilde{T}_{k+1}$ sufficiently large so that $|W_t^{(x,a)}((-\infty, z_i]))| < \delta/4$ for all $t \ge T_{k+1}$ and all $(x,a) \in \mathcal{X}\times\mathcal{A}$. Then (15) yields that $\eta_t \le U_{k+1}$ for all $t \ge T_{k+1}$, completing the inductive step, and therefore completing the proof of Lemma 6.

9.4 Proof of Theorem 2

Theorem 2. *Suppose that Assumptions (i)–(ii) of Theorem 1 hold, and that all unprojected target distributions $\widehat{\eta}_*^{(x_t,a_t)}$ arising in Algorithm 2 are supported within $[z_1, z_K]$ almost surely. Assume further that there is a unique optimal policy π^* for the MDP. Then, for the updates given in Algorithm 2, in the case of control, we have almost sure convergence of $(\eta_t^{(x,a)})_{(x,a)\in\mathcal{X}\times\mathcal{A}}$ in $\overline{\ell}_2$ to some limit $\eta_{\mathcal{C}}^*$, and furthermore the greedy policy with respect to $\eta_{\mathcal{C}}^*$ is the optimal policy π^*.*

Proof. We first note that the updates induced by the algorithm on the *expected returns* are exactly those of standard (non-distributional) Q-learning. More precisely, denoting the expected returns $\mathbb{E}_{R\sim\eta_t^{(x,a)}}[R]$ at state-action pair $(x,a) \in \mathcal{X} \times \mathcal{A}$ at time t by $Q_t(x,a)$, we have that these Q-values follow the standard dynamics of Q-learning. This holds because the maximum and minimum possible estimated rewards lie within the support of the parametrised distributions, by the assumptions of the theorem. We may therefore apply the non-distributional theory [Tsitsiklis, 1994] to argue that the expectations $(Q_t(x,a)|(x,a) \in \mathcal{X} \times \mathcal{A})$ converge almost-surely to the true optimal expected returns $(Q^{\pi^*}(x,a)|(x,a) \in \mathcal{X} \times \mathcal{A})$. Since the state space and action space are finite, this convergence is almost-surely uniform across all state-action pairs. Therefore, given $\varepsilon > 0$, there exists a random variable N such that for $t > N$, we have

$$\sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} |Q_t(x,a) - Q^{\pi^*}(x,a)| < \varepsilon \qquad \text{almost surely}\,.$$

Now take ε to be equal to half the minimum action gap across all states for the optimal action-value function Q^{π^*}; that is, take $\varepsilon = \frac{1}{2}\min_{x\in\mathcal{X}}[Q^{\pi^*}(x,\pi^*(x)) - \max_{a\neq\pi^*(x)} Q^{\pi^*}(x,a)]$ (which is greater than zero by the assumption of a unique optimal policy and finite state and action spaces). Then for $t > N$, the Q-learning updates are exactly the same as policy evaluation updates for the optimal policy π^*. Under these updates, we proved in Theorem 1 that the return distributions converge to the approximate return distribution function $\eta_\mathcal{C}$. Note however, that N is not a stopping time; we must be particularly careful with the analysis that follows.

We therefore proceed according to a coupling argument. We define the following set of independent stochastic distributional Bellman operators: $(\widehat{\mathcal{T}}_t^\pi)$ across all deterministic policies π, and timesteps $t \in \mathbb{N}$. The idea is to define a π^* categorical policy evaluation algorithm with these operators, and also a categorical Q-learning algorithm, and couple these processes together with probability tending to 1 as the number of steps of each algorithm increases. Since the return distribution ensemble computed by the policy evaluation algorithm will converge to the approximate return distribution function $\eta_\mathcal{C}$ associated with π^* almost surely, we will then be able to argue that the same is true of the distributions computed by the Q-learning algorithm.

More precisely, we first construct the π^* categorical policy evaluation algorithm by taking an initial return distribution function $(\eta_0^{(x,a)}|(x,a) \in \mathcal{X} \times \mathcal{A})$, and defining:

$$\eta_{k+1} = \Pi_\mathcal{C}\widehat{\mathcal{T}}_k^{\pi^*}\eta_k\,,$$

for each $k \geq 0$. We construct the Q-learning algorithm by taking the same initial return distribution function $(\eta_0^{(x,a)}|(x,a) \in \mathcal{X} \times \mathcal{A})$, and defining the following updates, letting $\widetilde{\eta}_0 = \eta_0$:

$$\begin{aligned}&\text{Let } \pi_k \text{ be greedy wrt } \widetilde{\eta}_k\,,\\ &\widetilde{\eta}_{k+1} = \Pi_\mathcal{C}\widehat{\mathcal{T}}_k^{\pi_k}\widetilde{\eta}_k\,,\end{aligned}$$

for each $k \geq 0$.

By the remarks above, we have $\pi_k = \pi^*$ for all k sufficiently large almost surely. Let $A_k = \{\pi_l = \pi^* \text{ for all } l \geq k\}$, for each $k \in \mathbb{N}$. Then $A_k \subseteq A_{k+1}$, and $\mathbb{P}(A_k) \uparrow 1$. Let B be the event of probability 1 for which the policy evaluation algorithm converges. Now, on the event $B \cap A_k$, we have

$$\overline{\ell}_2^2(\eta_l, \eta_\mathcal{C}) \to 0\,,$$

where $\eta_\mathcal{C}$ is the limiting distribution function for the policy π^*, as in Theorem 1. Note then that if $\overline{\ell}_2^2(\widetilde{\eta}_l, \eta_l) \to 0$ on this event too, then by the triangle inequality, we have $\overline{\ell}_2(\widetilde{\eta}_l, \eta_\mathcal{C}) \to 0$, and hence Q-learning converges on $A_k \cap B$, and since $\mathbb{P}(A_k \cap B) \uparrow 1$, the statement of the theorem immediately follows. We first observe that

$\overline{\ell}_2^2(\widetilde{\eta}_l, \eta_l)$, for $l \geq k$, is eventually a non-increasing positive sequence on the event A_k:

$$
\begin{aligned}
&\ell_2^2(\widetilde{\eta}_{l+1}^{(x,a)}, \eta_{l+1}^{(x,a)}) \\
=& \left\| \left((1-\alpha_l(x,a))\widetilde{\eta}_l^{(x,a)} + \alpha_l(x,a)(\Pi_{\mathcal{C}}\widehat{\mathcal{T}}_l^{\pi^*}\widetilde{\eta}_l)^{(x,a)} \right) - \left((1-\alpha_l(x,a))\eta_l^{(x,a)} + \alpha_l(x,a)(\Pi_{\mathcal{C}}\widehat{\mathcal{T}}_l^{\pi^*}\eta_l)^{(x,a)} \right) \right\|_{\ell_2}^2 \\
=& (1-\alpha_l(x,a))^2 \left\| \widetilde{\eta}_l^{(x,a)} - \eta_l^{(x,a)} \right\|_{\ell_2}^2 + \alpha_l(x,a)^2 \left\| (\Pi_{\mathcal{C}}\widehat{\mathcal{T}}_l^{\pi^*}\widetilde{\eta}_l)^{(x,a)} - (\Pi_{\mathcal{C}}\widehat{\mathcal{T}}_l^{\pi^*}\eta_l)^{(x,a)} \right\|_{\ell_2}^2 \\
& \qquad + 2\alpha_l(x,a)(1-\alpha_l(x,a))\langle \widetilde{\eta}_l^{(x,a)} - \eta_l^{(x,a)}, (\Pi_{\mathcal{C}}\widehat{\mathcal{T}}_l^{\pi^*}\widetilde{\eta}_l)^{(x,a)} - (\Pi_{\mathcal{C}}\widehat{\mathcal{T}}_l^{\pi^*}\eta_l)^{(x,a)} \rangle_{\ell_2} \\
\leq& (1-\alpha_l(x,a))^2 \overline{\ell}_2^2(\widetilde{\eta}_l, \eta_l) + \alpha_l(x,a)^2 \gamma \overline{\ell}_2^2(\widetilde{\eta}_l, \eta_l) + 2\alpha_l(x,a)(1-\alpha_l(x,a))\sqrt{\gamma}\overline{\ell}_2^2(\widetilde{\eta}_l, \eta_l) \\
=& (1-\alpha_l(x,a)(1-\sqrt{\gamma}))^2 \overline{\ell}_2^2(\widetilde{\eta}_l, \eta_l) \, .
\end{aligned}
\tag{16}
$$

Therefore, on this event, $\overline{\ell}_2(\widetilde{\eta}_l, \eta_l)$ has a limit almost surely. Denote Z as the limit of the sequence, and on the event that $Z > 0$, pick $\delta > 0$ such that $\sqrt{\gamma}(Z + \delta) < Z$. Letting $\tau > 0$ such that $\overline{\ell}_2(\widetilde{\eta}_l, \eta_l) < Z + \delta$ for all $l \geq \tau$, we observe that for $l \geq \tau$, following the calculations in Equation (16), we obtain the inequality

$$
\begin{aligned}
\ell_2^2(\widetilde{\eta}_{l+1}^{(x,a)}, \eta_{l+1}^{(x,a)}) &\leq (1-\alpha_l(x,a))^2 \ell_2^2(\widetilde{\eta}_l^{(x,a)}, \eta_l^{(x,a)}) + \alpha_l(x,a)^2\gamma(Z+\delta) + 2\alpha_l(x,a)(1-\alpha_l(x,a))\sqrt{\gamma}(Z+\delta) \\
&\leq (1 - 2\alpha_l(x,a) + \alpha_l(x,a)^2)\ell_2^2(\widetilde{\eta}_l^{(x,a)}, \eta_l^{(x,a)}) + (2\alpha_l(x,a) - \alpha_l(x,a)^2)\sqrt{\gamma}(Z+\delta) \, .
\end{aligned}
$$

By Assumption (i) of the theorem, we have $\limsup_l \ell_2(\widetilde{\eta}_l^{(x,a)}, \eta_l^{(x,a)}) \leq \sqrt{\gamma}(Z + \delta) < Z$ for all $(x, a) \in \mathcal{X} \times \mathcal{A}$, a contradiction. Therefore $\overline{\ell}_2^2(\widetilde{\eta}_l, \eta_l) \to 0$ holds on $A_k \cap B$ almost surely, as required. □