Appendices

Appendix A Laundry List of Convergent Algorithms

We outline the general proof recipe, which will be re-using for the following examples.

Proof strategy

(P1) Let $\mu^{(1)}, \mu^{(2)}$ be initial distributions and $(f_0^{(1)}, f_0^{(2)})$ be the optimal coupling which minimizes $\mathcal{W}(\mu^{(1)}, \mu^{(2)})$;

(P2) Define an appropriate coupling $f_1^{(1)} \sim \mu^{(1)}K, f_1^{(2)} \sim \mu^{(2)}K$ – e.g. by defining them to follow the same trajectories if the updates sample from the same distributions;

(P3) Use the upper bound $\mathcal{W}(\mu^{(1)}K, \mu^{(2)}K) \leq \mathbb{E}\left[\|f_1^{(1)} - f_2^{(2)}\|\right]$ and bound $\mathbb{E}\left[\|f_1^{(1)} - f_1^{(2)}\|\right] \leq \rho\mathbb{E}\left[\|f_0^{(1)} - f_0^{(2)}\|\right]$ for some $\rho < 1$ (usually follows from the recursive nature of the updates) to show that $\mu \mapsto \mu K$ is a contraction.

A.1 Convergence of synchronous Monte Carlo Evaluation with constant step-sizes

We prove that Monte Carlo Evaluation with synchronous updates & constant step-size converges to a stationary distribution. The algorithm aims to evaluate the value function of a given policy π using Monte Carlo returns. The update rule is given by:

$$\forall\, s \in \mathcal{S}: \quad V_{n+1}(s) = (1-\alpha)V_n(s) + \alpha\mathcal{G}_n^\pi(s) \tag{MCE}$$

where $\mathcal{G}_n^\pi(s) = \sum_{n\geq 0}\gamma^n r_n(s_n, a_n)$ is the return of a random trajectory $(s_n, a_n, r_n)_{n\geq 0}$ starting from s, following $a_n \sim \pi(\cdot|s_n), r_n \sim \mathcal{R}(\cdot|s_n, a_n)$, and $s_{n+1} \sim \mathcal{P}(\cdot|s_n, a_n)$.

Theorem A.1. *For any constant step size $0 < \alpha \leq 1$ and initialization $V_0 \sim \mu_0 \in \mathcal{M}(\mathbb{R}^{|\mathcal{S}|})$, the sequence of random variables $(V_n)_{n\geq 0}$ defined by the recursion (MCE) converges in distribution to a unique stationary distribution $\varphi_\alpha \in \mathcal{M}(\mathbb{R}^{|\mathcal{S}|})$.*

Proof. Following the proof strategy outlined above, we skip to step **(P2)** of the proof. We define the coupling of the updates $(V_1^{(1)}, V_1^{(2)})$ to sample the same trajectories:

$$\left.\begin{aligned} V_1^{(1)}(s) &= (1-\alpha)V_0^{(1)}(s) + \alpha\mathcal{G}_k^\pi(s) \\ V_1^{(2)}(s) &= (1-\alpha)V_0^{(2)}(s) + \alpha\mathcal{G}_k^\pi(s). \end{aligned}\right\} \text{for the } \underline{\text{same}}\ \mathcal{G}_k^\pi(s) \tag{11}$$

Note that this is a valid coupling of $(\mu^{(1)}K_\alpha, \mu^{(2)}K_\alpha)$, since $V_1^{(1)}(s)$ and $V_1^{(2)}(s)$ have access to the same sampling distributions. We upper bound $\mathcal{W}(\mu^{(1)}K_\alpha, \mu^{(2)}K_\alpha)$ by the coupling defined in Equation (11). This gives:

$$\begin{aligned} \mathcal{W}(\mu^{(1)}K_\alpha, \mu^{(2)}K_\alpha) &\leq \mathbb{E}\left[\left\|V_1^{(1)} - V_1^{(2)}\right\|\right] \\ &= \mathbb{E}\left[\left\|(1-\alpha)V_0^{(1)} + \alpha\mathcal{G}_1^\pi - \left((1-\alpha)V_0^{(2)} + \alpha\mathcal{G}_1^\pi\right)\right\|\right] \\ &= \mathbb{E}\left[\left\|(1-\alpha)(V_0^{(1)} - V_0^{(2)})\right\|\right] \\ &= (1-\alpha)\mathbb{E}\left[\left\|V_0^{(1)} - V_0^{(2)}\right\|\right] = (1-\alpha)\mathcal{W}(\mu^{(1)}, \mu^{(2)}) \end{aligned}$$

Since $1-\alpha < 1$, K_α is a contraction mapping and we are done. $\square$

A.2 Convergence of synchronous Q-Learning with constant step-sizes

We prove that Q-Learning with synchronous updates & constant step-sizes converges to a stationary distribution. The algorithm aims to learn the optimal action-value function $Q^\star$. The updates are given by:

$$\forall\, (s,a) \in \mathcal{S}\times\mathcal{A}: \quad Q_{n+1}(s,a) = (1-\alpha)Q_n(s,a) + \alpha\left(r + \gamma\max_{a'} Q_n(s', a')\right), \tag{QL}$$

where $r \sim \mathcal{R}(\cdot|s,a)$, $s' \sim \mathcal{P}(\cdot|s,a)$, and $\alpha > 0$.

Theorem A.2. *For any constant step size $0 < \alpha \leq 1$ and initialization $Q_0 \sim \mu_0 \in \mathcal{M}(\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|})$, the sequence of random variables $(Q_n)_{n\geq 0}$ defined by the recursion (QL) converges in distribution to a unique stationary distribution $\xi_\alpha \in \mathcal{M}(\mathbb{R}^{|\mathcal{S}|})$.*

Proof. We use the proof outline given above, and jump straight to step (**P2**). We witness the same-sampling coupling again:

$$\left.\begin{aligned} Q_1^{(1)}(s,a) &= (1-\alpha)Q_0^{(1)}(s,a) + \alpha\left(r + \gamma \max_{a'} Q_0^{(1)}(s',a')\right) \\ Q_1^{(2)}(s.a) &= (1-\alpha)Q_0^{(2)}(s,a) + \alpha\left(r + \gamma \max_{a'} Q_0^{(2)}(s',a')\right) \end{aligned}\right\} \text{for the } \underline{\text{same}} \begin{array}{l} r \sim \mathcal{R}(s,a), \\ s' \sim \mathcal{P}(\cdot|s,a) \end{array}$$

The bound follows similarly, but with one additional step. Again we write $\widehat{\mathcal{T}}(Q)(s,a) = r + \gamma \max_{a'} Q(s'_{(s,a)}, a')$ for the empirical Bellman (optimality) operator.

$$\begin{aligned} \mathbb{E}\left[\left\|\widehat{\mathcal{T}}(Q^{(1)}) - \widehat{\mathcal{T}}(Q^{(2)})\right\|\right] &= \mathbb{E}\left[\max_{s,a}\left|r - r + \gamma\left(\max_{a'} Q^{(1)}(s'_{(s,a)},a') - \max_{a'} Q^{(2)}(s'_{(s,a)},a')\right)\right|\right] \\ &= \gamma\mathbb{E}\left[\max_{s,a}\left|\max_{a'} Q^{(1)}(s'_{(s,a)},a') - \max_{a'} Q^{(2)}(s'_{(s,a)},a')\right|\right] \\ &\leq \gamma\mathbb{E}\left[\max_{s,a}\max_{a'}\left|Q^{(1)}(s'_{(s,a)},a') - Q^{(2)}(s'_{(s,a)},a')\right|\right] \\ &\leq \gamma\mathbb{E}\left[\max_{s,a}\left|Q^{(1)}(s,a) - Q^{(2)}(s,a)\right|\right] = \gamma\mathbb{E}\left[\left\|Q^{(1)} - Q^{(2)}\right\|\right] \end{aligned}$$

□

The first inequality follows from $|\max_{a'} Q_1(s,a') - \max_{a'} Q_2(s,a')| \leq \max_{a'}|Q_1(s,a') - Q_2(s,a')|$, and the second inequality follows since $Q^{(1)}$ and $Q^{(2)}$ sampled the same s'. Concluding the proof as before we see that the kernel is contractive with Lipschitz constant $1 + \alpha - \alpha\gamma < 1$, and we are done.

A.3 TD(λ)

We prove that TD(λ) with synchronous updates & constant step-size converges to a stationary distribution. The algorithm aims to evaluate the value function of a given policy π using a convex combination of n-step returns. The update rule is given by:

$$\forall s: \; V_{n+1}(s) = (1-\alpha)V_n(s,a) + \alpha(1-\lambda)\sum_{k=1}^{\infty}\lambda^{k-1}\left(\sum_{i=0}^{k}\gamma^i r(s_i,a_i) + \gamma^k V_n(s_k)\right) \qquad \text{(TD}(\lambda))$$

where each n-step trajectory is sampled starting from s and following policy π.

Theorem A.3. *For any constant step size $0 < \alpha \leq 1$ and initialization $V_0 \sim \mu_0 \in \mathcal{M}(\mathbb{R}^{|\mathcal{S}|})$, the sequence of random variables $(V_n)_{n\geq 0}$ defined by the recursion (TD(λ)) converges in distribution to a unique stationary distribution $\zeta_\alpha \in \mathcal{M}(\mathbb{R}^{|\mathcal{S}|})$.*

Proof. Again, we jump straight to step (**P2**) of the template given above. We couple every n-step trajectory to sample the same n rewards, actions, and successors states.

$$\left.\begin{aligned} V_{k+1}^{(1)}(s) &= (1-\alpha)V_k^{(1)}(s) + \alpha(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\left(\sum_{i=0}^{n-1}\gamma^i r_i(s_i,a_i) + \gamma^n V_k^{(1)}(s_n)\right) \\ V_{k+1}^{(2)}(s) &= (1-\alpha)V_k^{(2)}(s) + \alpha(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\left(\sum_{i=0}^{n-1}\gamma^i r_i(s_i,a_i) + \gamma^n V_k^{(2)}(s_n)\right) \end{aligned}\right\} \begin{array}{l} \underline{\text{same}} \\ (s_i,a_i,r_i)_{i=0}^n \\ \forall n \end{array}$$

By the coupling, the reward terms will cancel in every n-step trajectory. We write $R_n^{(i)} = \sum_{i=0}^{n-1}\gamma^i r_i(s_i,a_i) + \gamma^n V_k^{(i)}(s_n)$ for the n-step return and $\hat{\mathcal{T}}(V)(s) = \sum_{k=1}^{\infty}\lambda^{k-1}\left(\sum_{i=0}^{k}\gamma^i r(s_i,a_i) + \gamma^k V_n(s_k)\right)$ for the empirical Bellman operator of TD(λ).

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\mathcal{T}}(V^{(1)}) - \hat{\mathcal{T}}(V^{(2)})\right\|\right] &= \mathbb{E}\left[\max_s \left|\sum_{n=1}^{\infty} \lambda^{n-1} R_n^{(1)} - \sum_{n=1}^{\infty} \lambda^{n-1} R_n^{(2)}\right|\right] \\
&= \mathbb{E}\left[\max_s \left|\sum_{n=1}^{\infty} \lambda^{n-1} \left(R_n^{(1)} - R_n^{(2)}\right)\right|\right] \\
&= \mathbb{E}\left[\max_s \left|\sum_{n=1}^{\infty} \lambda^{n-1} \gamma^n \left(V^{(1)}(s_n) - V^{(2)}(s_n)\right)\right|\right] && \text{(reward terms cancel)} \\
&\leq \mathbb{E}\left[\sum_{n=1}^{\infty} \lambda^{n-1} \gamma^n \max_s \left|\left(V^{(1)}(s_n) - V^{(2)}(s_n)\right)\right|\right] && \text{(triangle inequality)} \\
&\leq \sum_{n=1}^{\infty} \lambda^{n-1} \gamma^n \mathbb{E}\left[\max_s \left|V^{(1)}(s) - V^{(2)}(s)\right|\right] && \text{(by the coupling)} \\
&= \sum_{n=1}^{\infty} \lambda^{n-1} \gamma^n \mathbb{E}\left[\left\|V^{(1)} - V^{(2)}\right\|\right] = \gamma \frac{1}{1-\lambda\gamma} \mathbb{E}\left[\left\|V^{(1)} - V^{(2)}\right\|\right]
\end{aligned}$$

Concluding the proof as before, we have $\mathcal{W}(\mu^{(1)}K, \mu^{(2)}K) \leq (1-\alpha+\alpha\gamma\frac{1-\lambda}{1-\lambda\gamma})\mathcal{W}(\mu^{(1)}, \mu^{(2)})$. Since $1-\alpha+\alpha\gamma\frac{1-\lambda}{1-\lambda\gamma}$ ¡ 1 we are done. $\square$

A.4 SARSA with ε-greedy policies

In this example we will example the use of ε-greedy policies for control. In particular, we examine SARSA updates with ε-greedy policies. Let $\pi(\cdot|s)$ be some base policy. The updates are as follow:

$$Q_{k+1}(s,a) = \begin{cases} (1-\alpha)Q_k(s,a) + \alpha\left(r(s,a) + \gamma Q_k(s',a')\right) & \text{w.p. } \varepsilon \\ (1-\alpha)Q_k(s,a) + \alpha\left(r(s,a) + \gamma \max_{a'} Q_k(s',a')\right) & \text{w.p. } 1-\varepsilon \end{cases} \tag{SARSA}$$

where $r \sim \mathcal{R}(\cdot|s,a)$ and $s' \sim \mathcal{P}(\cdot|s,a)$ in both cases and $a' \sim \pi(\cdot|s')$ in the first case.

Theorem A.4. *For any constant step size $0 < \alpha \leq 1$ and initialization $Q_0 \sim \mu_0 \in \mathcal{M}(\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|})$, the sequence of random variables $(Q_n)_{n\geq 0}$ defined by the recursion (SARSA) converges in distribution to a unique stationary distribution $\theta_\alpha \in \mathcal{M}(\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|})$.*

Proof. We jump straight to step **(P2)** of the proof template. We use the same-sampling coupling, where $Q_1^{(1)}$ takes the greedy action if and only if $Q_1^{(2)}$ does. In the non-greedy case, they sample the same $a' \sim \pi(\cdot|s')$. In all cases, both functions sample the same $r(s,a)$ and s'.

We write $\hat{\mathcal{T}}(Q)(s,a) = \begin{cases} r + \gamma Q(s',a') \text{ w.p. } \varepsilon \\ r + \gamma \max_{a'} Q(s',a') \text{ w.p. } 1-\varepsilon \end{cases}$

The bound follows similarly to the examples of Q-learning and TD(0). We omit the subscripts on the Q-functions.

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\mathcal{T}}(Q^{(1)}) - \hat{\mathcal{T}}(Q^{(2)})\right\|\right] &= \mathbb{P}\left\{\text{greedy action chosen}\right\} \mathbb{E}\left[\max_{s,a} \gamma|(\max_{a'} Q^{(1)}(s',a') - \max_{a'} Q^{(2)}(s',a')|\right] \\
&\quad + \mathbb{P}\left\{\text{non-greedy action chosen}\right\} \mathbb{E}\left[\max_{s,a}|\gamma(Q^{(1)}(s',a') - Q^{(2)}(s',a'))|\right] \\
&\leq \varepsilon\gamma\mathbb{E}\left[\left\|Q^{(1)} - Q^{(2)}\right\|\right] + (1-\varepsilon)\gamma\mathbb{E}\left[\left\|Q^{(1)} - Q^{(2)}\right\|\right] \\
&= \gamma\mathbb{E}\left[\|Q^{(1)} - Q^{(2)}\|\right]
\end{aligned}$$

The bound $\mathbb{E}\left[\max_{s,a} \gamma|(\max_{a'} Q^{(1)}(s',a') - \max_{a'} Q^{(2)}(s',a')|\right] \leq \gamma\mathbb{E}\left[\|Q^{(1)} - Q^{(2)}\|\right]$ follows from $|\max_{a'} Q_1(s,a') - \max_{a'} Q_2(s,a')| \leq \max_{a'}|Q_1(s,a') - Q_2(s,a')|$, and since $Q^{(1)}$ and $Q^{(2)}$ sampled the same s' in the greedy case. The bound $\mathbb{E}\left[\max_{s,a}|\gamma(Q^{(1)}(s',a') - Q^{(2)}(s',a'))|\right] \leq \mathbb{E}\left[\left\|Q^{(1)} - Q^{(2)}\right\|\right]$ follows since $Q^{(1)}$ and $Q^{(2)}$ sampled the same state-action pair in the non-greedy case. Concluding the proof as before, we have that $\mathbb{E}\left[\|Q_1^{(1)} - Q_1^{(2)}\|\right] \leq (1-\alpha+\alpha\gamma)\mathbb{E}\left[\|Q_0^{(1)} - Q_0^{(2)}\|\right]$, and thus the kernel is a contraction. □

A.5 Expected SARSA with ε-greedy policies

In this example we examine the Expected SARSA updates with ε-greedy policies. Let $\pi(\cdot|s)$ be some base policy. Define $\pi_\varepsilon(\cdot|s)$ as the ε-greedy policy which takes the greedy action with probability 1-ε and π otherwise. The updates are as follow:

$$Q_{k+1}(s,a) = (1-\alpha)Q_k(s,a) + \alpha\left(r(s,a) + \gamma\sum_{a'}\pi_\varepsilon(a'|s)Q_k(s',a')\right) \tag{Expected-SARSA}$$

where $r \sim \mathcal{R}(\cdot|s,a)$ and $s' \sim \mathcal{P}(\cdot|s,a)$ in both cases and $a' \sim \pi(\cdot|s')$ in the first case.

Theorem A.5. *For any constant step size $0 < \alpha \leq 1$ and initialization $Q_0 \sim \mu_0 \in \mathcal{M}(\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|})$, the sequence of random variables $(Q_n)_{n\geq 0}$ defined by the recursion (Expected-SARSA) converges in distribution to a unique stationary distribution $\beta_\alpha \in \mathcal{M}(\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|})$.*

Proof. We jump straight to step **(P2)** of the proof template. We use the same-sampling coupling.

We write $\hat{\mathcal{T}}(Q)(s,a) = r + \gamma\sum_{a'}\pi(a'|s)Q(s',a')$. The bound follows similarly to the examples of Q-learning and TD(0). We omit the subscripts on the Q-functions.

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\mathcal{T}}(Q^{(1)}) - \hat{\mathcal{T}}(Q^{(2)})\right\|\right] &= \mathbb{E}\left[\max_{s,a}\gamma|\sum_{a'}\pi_\varepsilon(a')Q^{(1)}(s',a') - \sum_{a'}\pi_\varepsilon(a')Q^{(2)}(s',a')|\right] \\
&\leq \mathbb{E}\left[\max_{s,a}\gamma\sum_{a'}\pi_\varepsilon(a')|Q^{(1)}(s',a') - Q^{(2)}(s',a')|\right] \\
&\leq \mathbb{E}\left[\max_{s,a}\gamma\sum_{a'}\pi_\varepsilon(a')\left\|Q^{(1)}(s',a') - Q^{(2)}(s',a')\right\|\right] \\
&\leq \gamma\mathbb{E}\left[\|Q^{(1)} - Q^{(2)}\|\right]
\end{aligned}$$

Concluding the proof as before, we have that $\mathbb{E}\left[\|Q_1^{(1)} - Q_1^{(2)}\|\right] \leq (1-\alpha+\alpha\gamma)\mathbb{E}\left[\|Q_0^{(1)} - Q_0^{(2)}\|\right]$, and thus the kernel is a contraction. □

A.6 Double Q-Learning

In this example we will have to modify our state-space and introduce a new metric on pairs of Q-functions. The Double Q-Learning algorithm (Hasselt, 2010)[1] maintains two random estimates (Q^A, Q^B) and updates Q^A with probability p and Q^B with probability $1-p$. Should Q^A be chosen to be updated, the update is:

$$Q^A_{n+1}(s,a) = (1-\alpha)Q^A_n(s,a) + \alpha\left(r(s,a) + \gamma Q^B_n(s, \operatorname{argmax}_{a'} Q^A_n(s',a'))\right).$$

Analogously, the update for Q^B is:

$$Q^B_{n+1}(s,a) = (1-\alpha)Q^B_n(s,a) + \alpha\left(r(s,a) + \gamma Q^A_n(s, \operatorname{argmax}_{a'} Q^B_n(s',a'))\right).$$

In both cases, we have $s' \sim \mathcal{P}(\cdot|s,a)$. For this algorithm, the updates are Markovian on *pairs* of action-value functions. Thus we set the state space to be $\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} \times \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$. We choose the product metric defined by $d_1((Q^A,Q^B),(R^A,R^B)) = \left\|Q^A - R^A\right\| + \left\|Q^B - R^B\right\|$.

[1] This is the original algorithm, not the deep reinforcement learning version given in (Van Hasselt, Guez, and Silver, 2016).

Theorem A.6. *For any constant step size $0 < \alpha \leq 1$ and initialization $(Q_0^A, Q_0^B) \sim \mu_0 \in \mathcal{M}(\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} \times \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|})$, the sequence of random variables $(Q_n^A, Q_n^B)_{n\geq 0}$ defined by the Double Q-Learning recursion converges in distribution to a unique stationary distribution $\chi_\alpha \in \mathcal{M}(\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} \times \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|})$.*

Proof. As before, let $\mu^{(1)}, \mu^{(2)}\mathcal{M}(\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} \times \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|})$ be arbitrary initializations and (Q_0^A, Q_0^B) and (R_0^A, R_0^B) be the optimal coupling of $\mathcal{W}(\mu^{(1)}, \mu^{(2)})$. We couple (Q_1^A, Q_1^B) and (R_1^A, R_1^B) to sample the same function to be updated and the same s'. Assume for a moment that Q^A and R^A are chosen to be updated. Proceeding as in the proof of Q-Learning (cf. Theorem A.2), we find that

$$\mathbb{E}\left[\left\|Q_1^A - R_1^A\right\|\right] \leq (1-\alpha)\mathbb{E}\left[\left\|Q_0^A - R_0^A\right\|\right] + \alpha\gamma\mathbb{E}\left[\left\|Q_0^B - R_0^B\right\|\right].$$

Analogously, if Q^B and R^B are chosen to updated, we have:

$$\mathbb{E}\left[\left\|Q_1^B - R_1^B\right\|\right] \leq (1-\alpha)\mathbb{E}\left[\left\|Q_0^B - R_0^B\right\|\right] + \alpha\gamma\mathbb{E}\left[\left\|Q_0^A - R_0^A\right\|\right].$$

Putting everything together, the full expectation is:

$$\begin{aligned}
\mathbb{E}\left[d((Q_1^A, Q_1^B), (R_1^A, R_1^B))\right] &= \mathbb{E}\left[\left\|Q_1^A - R_1^A\right\| + \left\|Q_1^B - R_1^B\right\|\right] \\
&= \mathbb{P}\left\{\text{A is updated}\right\}\mathbb{E}\left[\left\|Q_1^A - R_1^A\right\| + \left\|Q_1^B - R_1^B\right\|\right] \\
&\quad + \mathbb{P}\left\{\text{B is updated}\right\}\mathbb{E}\left[\left\|Q_1^A - R_1^A\right\| + \left\|Q_1^B - R_1^B\right\|\right] \\
&= p\mathbb{E}\left[\left\|Q_1^A - R_1^A\right\| + \left\|Q_0^B - R_0^B\right\|\right] \\
&\quad + (1-p)\mathbb{E}\left[\left\|Q_0^A - R_0^A\right\| + \left\|Q_1^B - R_1^B\right\|\right] \\
&\leq p\left((1-\alpha)\mathbb{E}\left[\left\|Q_0^A - R_0^A\right\|\right] + (1+\alpha\gamma)\mathbb{E}\left[\left\|Q_0^B - R_0^B\right\|\right]\right) \\
&\quad + (1-p)\left((1+\alpha\gamma)\mathbb{E}\left[\left\|Q_0^A - R_0^A\right\|\right] + (1-\alpha)\mathbb{E}\left[\left\|Q_0^B - R_0^B\right\|\right]\right) \\
&\leq \frac{1}{2}(2+\alpha\gamma-\alpha)\left(\mathbb{E}\left[\left\|Q_0^A - R_0^A\right\|\right] + \mathbb{E}\left[\left\|Q_0^B - R_0^B\right\|\right]\right) \qquad (p = \tfrac{1}{2}) \\
&= \frac{1}{2}(2+\alpha\gamma-\alpha)\mathbb{E}\left[d((Q_0^A, Q_0^B), (R_0^A, R_0^B))\right]
\end{aligned}$$

Since $0 \leq 1/2(2+\alpha\gamma-\alpha) < 1$, so we are done. We note that the first equality only follows since, under the coupling, either A or B is updated for both functions. □

Appendix B Proofs of Section 5

Theorem B.1. *Suppose $\widehat{\mathcal{T}}^\pi$ is such that the updates (5) with step-size α converge to a stationary distribution ψ_α. If $\widehat{\mathcal{T}}$ is an empirical Bellman operator for some policy π, then $\mathbb{E}[f_\alpha] = f^\pi$ where $f_\alpha \sim \psi_\alpha$ and f^π is the fixed point of $\mathcal{T}^\pi$.*

Proof. Let f_0 be distributed according to ψ_α. Rewriting equation (5):

$$f_1 = (1-\alpha)f_0 + \alpha\mathcal{T}^\pi f_0 + \alpha\xi(f_0), \tag{12}$$

where $\xi(f_0) = \hat{\mathcal{T}}^\pi(f_0, \omega) - \mathcal{T}^\pi f_0$ is a zero-mean noise term. Taking expectations on both sides, and using that f_1 is also distributed according to ψ_α by stationarity and that $\mathbb{E}[\xi(f)] = 0$ for any f:

$$\begin{aligned}
\overline{f_\alpha} &= (1-\alpha)\overline{f_\alpha} + \alpha\mathbb{E}[\mathcal{T}^\pi f_0] \\
\alpha\overline{f_\alpha} &= \alpha\mathbb{E}[\mathcal{R}^\pi + \gamma\mathcal{P}^\pi f_0] \\
\overline{f_\alpha} &= \mathcal{R}^\pi + \gamma\mathcal{P}^\pi\mathbb{E}[f_0] \\
\overline{f_\alpha} &= \mathcal{T}^\pi\overline{f_\alpha}
\end{aligned}$$

And therefore $\overline{f_\alpha} = f^\pi$ since it is the unique fixed point of $\mathcal{T}^\pi$. □

Theorem B.2. *Suppose $\widehat{\mathcal{T}}^\pi$ is such that the updates (5) with step-size α converge to a stationary distribution ψ_α, and that $\widehat{\mathcal{T}}^\pi$ is an empirical Bellman operator for some policy π. Define*

$$\mathcal{C}(f) := \mathbb{E}_\omega[(\widehat{\mathcal{T}}^\pi(f,\omega) - \mathcal{T}^\pi f)(\widehat{\mathcal{T}}^\pi(f,\omega) - \mathcal{T}^\pi f)^\mathsf{T}]$$

to be the covariance of the zero-mean noise term $\widehat{\mathcal{T}}^\pi(f,\omega) - \mathcal{T}^\pi f$ for a given function f. Then, the covariance of $f_\alpha \sim \psi_\alpha$ is given by

$$\begin{aligned}(1-(1-\alpha)^2)\mathbb{E}\left[(f_\alpha - f^\pi)(f_\alpha - f^\pi)^\intercal\right] &= \alpha^2(\gamma\mathcal{P}^\pi)\mathbb{E}\left[(f_\alpha - f^\pi)(f_\alpha - f^\pi)^\intercal\right](\gamma\mathcal{P}^\pi)^\intercal \\ &+ \alpha(1-\alpha)(\gamma\mathcal{P}^\pi)\mathbb{E}\left[(f_0 - f^\pi)(f_0 - f^\pi)^\intercal\right] \\ &+ \alpha(1-\alpha)\mathbb{E}\left[(f_\alpha - f^\pi)(f_\alpha - f^\pi)^\intercal\right](\gamma\mathcal{P}^\pi)^\intercal \\ &+ \alpha^2\int \mathcal{C}(f)\psi_\alpha(\mathrm{d}f)\end{aligned}$$

Furthermore, we have that $\left\|\mathbb{E}\left[(f_\alpha - f^\pi)(f_\alpha - f^\pi)^\intercal\right]\right\|_{op}$ is monotonically decreasing with respect to α, where $\|\cdot\|_{op}$ denotes the operator norm of a matrix. In particular, $\lim_{\alpha\to 0}\left\|\mathbb{E}[(f_\alpha - f^\pi)(f_\alpha - f^\pi)^\intercal]\right\|_{op} = 0$, and we have that:

$$\mathbb{P}\left\{\min_i |f_\alpha(i) - f^\pi(i)| \geq \varepsilon\right\} \xrightarrow{\alpha\to 0} 0 \quad \forall\, \varepsilon > 0$$

We preface the proof with some useful identities. We will write the covariance in terms of the tensor product for ease of manipulations

Lemma B.1. *Write $\xi(f) := (\widehat{\mathcal{T}}^\pi(f,\omega) - \mathcal{T}^\pi f)$. In the same setup as Theorem 5.2:*

$$\mathbb{E}\left[(f_\alpha - f^\pi)(\mathcal{T}^\pi f_\alpha - f^\pi + \xi(f_0))^\intercal\right] = \mathbb{E}\left[(f_\alpha - f^\pi)(f_\alpha - f^\pi)^\intercal\right](\gamma\mathcal{P}^\pi)^\intercal$$

and

$$\begin{aligned}\mathbb{E}\left[((\mathcal{T}^\pi f_\alpha - f^\pi) + \xi(f_\alpha))\,((\mathcal{T}^\pi f_\alpha - f^\pi) + \xi(f_\alpha))^\intercal\right] &= (\gamma\mathcal{P}^\pi)\mathbb{E}\left[(f_\alpha - f^\pi)(f_\alpha - f^\pi)^\intercal\right](\gamma\mathcal{P}^\pi)^\intercal \\ &+ \int C(v)\psi_\alpha(\mathrm{d}v)\end{aligned}$$

Proof. Let $f_0 \sim \psi_\alpha$, by (5) we have $f_1 = (1-\alpha)f_0 + \alpha(\mathcal{T}^\pi f_0 + \xi(f_0))$ and $f_1 \sim \psi_\alpha$. Furthermore, the distribution of f_0 is independent of the distribution of ω. By independence,

$$\begin{aligned}\mathbb{E}\left[(f_0 - f^\pi)\xi(f_0)^\intercal\right] &= \mathbb{E}_{f_0}\mathbb{E}_\omega\left[(f_0 - f^\pi)\xi(f_0)^\intercal\right] && \text{(by independence of } f_0 \text{ and } \xi(\cdot)) \\ &= \mathbb{E}_{f_0}\left[(f_0 - f^\pi)(\mathbb{E}_\omega\xi(f_0))^\intercal\right] = 0 && (\mathbb{E}_\omega[\xi(f)] = 0 \text{ for every } f)\end{aligned}$$

For the first identity, note that

$$\begin{aligned}\mathbb{E}\left[(f_0 - f^\pi)(\mathcal{T}^\pi f_0 - f^\pi))^\intercal\right] &= \mathbb{E}\left[(f_0 - f^\pi)(\mathcal{R}^\pi + \gamma\mathcal{P}^\pi(f_0) - \mathcal{R}^\pi - \gamma\mathcal{P}^\pi(f^\pi))^\intercal\right] \\ &= \mathbb{E}\left[(f_0 - f^\pi)(\gamma\mathcal{P}^\pi(f_0 - f^\pi))^\intercal\right] \\ &= \mathbb{E}\left[(f_0 - f^\pi)(f_0 - f^\pi)^\intercal(\gamma\mathcal{P}^\pi)^\intercal\right] \\ &= \mathbb{E}\left[(f_0 - f^\pi)(f_0 - f^\pi)^\intercal\right](\gamma\mathcal{P}^\pi)^\intercal\end{aligned}$$

The first identity then follows by using $\mathbb{E}\left[(f_0 - f^\pi)\xi(f_0)^\intercal\right] = 0$ and linearity of expectations.

For the second identity, expanding the outer product gives:

$$\begin{aligned}\mathbb{E}\left[((\mathcal{T}^\pi f_0 - f^\pi) + \xi(f_0))\,((\mathcal{T}^\pi f_0 - f^\pi) + \xi(f_0))^\intercal\right] &= \mathbb{E}\left[(\mathcal{T}^\pi f_0 - f^\pi)(\mathcal{T}^\pi f_0 - f^\pi)^\intercal\right] \\ &+ \mathbb{E}\left[(\xi(f_0))(\xi(f_0)))^\intercal\right] \\ &+ \cancel{\mathbb{E}\left[(\mathcal{T}^\pi f_0 - f^\pi)(\xi(f_0))^\intercal\right]} \\ &+ \cancel{\mathbb{E}\left[\xi(f_0)(\mathcal{T}^\pi f_0 - f^\pi)^\intercal\right]} \\ &= \mathbb{E}\left[(\gamma\mathcal{P}^\pi(f_0 - f^\pi))(\gamma\mathcal{P}^\pi(f_0 - f^\pi))^\intercal\right] \\ &+ \int \mathcal{C}(v)\psi_\alpha(\mathrm{d}v) \\ &= (\gamma P^\pi)\mathbb{E}\left[(f_0 - f^\pi)(f_0 - f^\pi)^\intercal\right](\gamma P^\pi)^\intercal \\ &+ \int \mathcal{C}(v)\psi_\alpha(\mathrm{d}v)\end{aligned}$$

where we used $\mathbb{E}\left[(\mathcal{T}^\pi f_0 - f^\pi)(\xi(f_0))^\top\right] = 0$. □

Proof (*of Theorem 5.2*). Again let f_0 be distributed according to ψ_α. Subtracting f^π from equation (12),

$$f_1 - f^\pi = (1-\alpha)(f_0 - f^\pi) + \alpha\left(\mathcal{T}^\pi f_0 - f^\pi + \xi(f_0)\right).$$

and taking outer products:

$$\begin{aligned}(f_1 - f^\pi)(f_1 - f^\pi)^\top = &(1-\alpha)^2 (f_0 - f^\pi)(f_0 - f^\pi)^\top \\ &+ \alpha^2 (\mathcal{T}^\pi f_0 - f^\pi + \xi(f_0))(\mathcal{T}^\pi f_0 - f^\pi + \xi(f_0))^\top \\ &+ \alpha(1-\alpha)(f_0 - f^\pi)(\mathcal{T}^\pi f_0 - f^\pi + \xi(f_0))^\top \\ &+ \alpha(1-\alpha)(\mathcal{T}^\pi f_0 - f^\pi + \xi(f_0))(f_0 - f^\pi)^\top.\end{aligned}$$

Taking expectations on both sides, and using Lemma B.1:

$$\begin{aligned}\mathbb{E}\left[(f_1 - f^\pi)(f_1 - f^\pi)^\top\right] = &(1-\alpha)^2\mathbb{E}\left[(f_0 - f^\pi)(f_0 - f^\pi)^\top\right] + \alpha^2(\gamma\mathcal{P}^\pi)\mathbb{E}[(f_0 - f^\pi)](\gamma\mathcal{P}^\pi)^\top \\ &+ \alpha^2\int \mathcal{C}(v)\psi_a(\mathrm{d}v) \\ &+ \alpha(1-\alpha)(\gamma\mathcal{P}^\pi)\mathbb{E}\left[(f_0 - f^\pi)(f_0 - f^\pi)^\top\right] \\ &+ \alpha(1-\alpha)\mathbb{E}\left[(f_0 - f^\pi)(f_0 - f^\pi)^\top\right](\gamma\mathcal{P}^\pi)^\top\end{aligned}$$

Since $\mathbb{E}\left[(f_1 - f^\pi)(f_1 - f^\pi)^\top\right] = \mathbb{E}\left[(f_0 - f^\pi)(f_0 - f^\pi)^\top\right]$ by stationarity, re-arranging to the LHS and factoring gives:

$$\begin{aligned}(1-(1-\alpha)^2)\mathbb{E}\left[(f_\alpha - f^\pi)(f_\alpha - f^\pi)^\top\right] = &\alpha^2(\gamma\mathcal{P}^\pi)\mathbb{E}\left[(f_\alpha - f^\pi)(f_\alpha - f^\pi)^\top\right](\gamma\mathcal{P}^\pi)^\top \\ &+ \alpha(1-\alpha)(\gamma\mathcal{P}^\pi)\mathbb{E}\left[(f_0 - f^\pi)(f_0 - f^\pi)^\top\right] \\ &+ \alpha(1-\alpha)\mathbb{E}\left[(f_\alpha - f^\pi)(f_\alpha - f^\pi)^\top\right](\gamma\mathcal{P}^\pi)^\top \\ &+ \alpha^2\int \mathcal{C}(f)\psi_\alpha(\mathrm{d}f)\end{aligned}$$

For the remainder of the proof we re-write the above expression in terms of tensor products. The tensor product of two vectors x, y is the matrix defined by $x \otimes y = xy^\top$. By extension, the tensor product of two matrices A, B is the operator defined by $(A \otimes B)X = AXB^\top$. Then, the above expression can be re-written as:

$$\begin{aligned}(1-(1-\alpha)^2)\mathbb{E}\left[(f_\alpha - f^\pi)(f_\alpha - f^\pi)^\top\right] = &\alpha^2(\gamma\mathcal{P}^\pi)^{\otimes 2}\mathbb{E}\left[(f_\alpha - f^\pi)(f_\alpha - f^\pi)^\top\right] \\ &+ \alpha(1-\alpha)(\gamma\mathcal{P}^\pi \otimes \mathrm{I})\mathbb{E}\left[(f_0 - f^\pi)(f_0 - f^\pi)^\top\right] \\ &+ \alpha(1-\alpha)(\mathrm{I} \otimes \gamma\mathcal{P}^\pi)\mathbb{E}\left[(f_\alpha - f^\pi)(f_\alpha - f^\pi)^\top\right] \\ &+ \alpha^2\int \mathcal{C}(f)\psi_\alpha(\mathrm{d}f).\end{aligned}$$

Factoring the tensor products further gives:

$$\left[I - ((1-\alpha)I + \alpha\gamma P^\pi)^{\otimes 2}\right]\mathbb{E}\left[(f_\alpha - f^\pi)^{\otimes 2}\right] = \alpha^2\int \mathcal{C}(f)\psi_\alpha(\mathrm{d}f)$$

We show that the matrix on the LHS is invertible. By (Puterman, 2014, Corollary C.4) it will follow from showing that $\rho\left(((1-\alpha)I + \alpha\gamma P^\pi)^{\otimes 2}\right) < 1$, where $\rho(A)$ is the spectral radius of matrix A. Writing $\|A\|_{\text{op}} = \max_i \sum_j |A(i,j)|$ for the operator norm of a matrix A, and using that $\rho(A) \leq \|A\|_{\text{op}}$, $\|A \otimes B\|_{\text{op}} = \|A\|_{\text{op}}\|B\|_{\text{op}}$, and $\|P^\pi\|_{\text{op}} = \|I\|_{\text{op}} = 1$:

$$\left\|((1-\alpha)I + \alpha\gamma P^\pi)^{\otimes 2}\right\|_{\text{op}} = \|(1-\alpha)I + \alpha\gamma P^\pi\|_{\text{op}}^2 \leq ((1-\alpha) + \alpha\gamma)^2 < 1, \tag{13}$$

where the last inequality followed since $\gamma < 1$. Finally, for the limit $\alpha \to 0$, we use the following identity: if A is such that $\|I - A\| \leq 1$ then $\left\|A^{-1}\right\| \leq \frac{1}{1-\|I-A\|}$. We let $A = I - ((1-\alpha)I + \alpha\gamma\mathcal{P}^\pi)^{\otimes 2}$, by the calculation in (13) we have $\|I - A\| < 1$. So we calculate the operator norm of the covariance matrix:

$$\begin{aligned}
\left\|\mathbb{E}\left[(f_0 - f^\pi)(f_0 - f^\pi)^\mathsf{T}\right]\right\| &= \alpha^2 \left\|\left[I - ((1-\alpha)I + \alpha\gamma P^\pi)^{\otimes 2}\right]^{-1} \int \mathcal{C}(v)\psi_\alpha(\mathrm{d}v)\right\| \\
&\leq \alpha^2 \left\|\left[I - ((1-\alpha)I + \alpha\gamma P^\pi)^{\otimes 2}\right]^{-1}\right\| \left\|\int \mathcal{C}(v)\psi_\alpha(\mathrm{d}v)\right\| \\
&\leq \alpha^2 \frac{1}{1 - \left\|I - I + ((1-\alpha)I + \alpha\gamma P^\pi)^{\otimes 2}\right\|} \left\|\int \mathcal{C}(v)\psi_\alpha(\mathrm{d}v)\right\| \\
&= \alpha^2 \frac{1}{1 - \left\|((1-\alpha)I + \alpha\gamma P^\pi)^{\otimes 2}\right\|} \left\|\int \mathcal{C}(v)\psi_\alpha(\mathrm{d}v)\right\| \\
&= \alpha^2 \frac{1}{1 - \|((1-\alpha)I + \alpha\gamma P^\pi)\|^2} \left\|\int \mathcal{C}(v)\psi_\alpha(\mathrm{d}v)\right\| \\
&\leq \alpha^2 \frac{1}{1 - (1-\alpha+\alpha\gamma)^2} \left\|\int \mathcal{C}(v)\psi_\alpha(\mathrm{d}v)\right\|
\end{aligned}$$

Finally, since the state space is bounded in $[0, \text{RMAX}/(1-\gamma)]^n$, we have $(\widehat{\mathcal{T}}f)_i \leq \text{RMAX}/(1-\gamma)$ and $(\mathcal{T}f)_i \leq \text{RMAX}/(1-\gamma)$ for each i. Then, we have $|\xi_\omega(f)_i\xi_\omega(f)_j| = |(\widehat{\mathcal{T}}f)_i(\mathcal{T}f)_j - (\mathcal{T}f)_i(\widehat{\mathcal{T}}f)_j - (\mathcal{T}f)_j(\widehat{\mathcal{T}}f)_j + (\mathcal{T}f)_j(\mathcal{T}f)_i| \leq 4\frac{\text{RMAX}^2}{(1-\gamma)^2}$ Thus we have $\|\mathcal{C}(f)\| \leq 4\frac{\text{RMAX}^2}{(1-\gamma)^2} := M$ and thus

$$\left\|\mathbb{E}\left[(f_0 - f^\pi)(f_0 - f^\pi)^\mathsf{T}\right]\right\| \leq M \frac{\alpha^2}{1 - (1-\alpha+\alpha\gamma)^2} \overset{\alpha\to 0}{\longrightarrow} 0$$

For the concentration inequality, we will use a multivariate Chebyshev inequality (Marshall and Olkin, 1960, Theorem 3.1), whos statement is as follows:

Theorem B.3. *Let $X = (X_1, ..., X_n)$ be a random vector with $\mathbb{E}X = 0$ and $\mathbb{E}[X^TX] = \Sigma$. Let $T = T_+ \cup \{x : -x \in T_+\}$, where $T_+ \subseteq \mathbb{R}^n$ is a closed, convex set. If $A = \{a \in \mathbb{R}^n : \langle a, x\rangle \geq 1\ \forall x \in T_+\}$, then*

$$\mathbb{P}\{X \in T\} \leq \inf_{a\in A} a^\mathsf{T}\Sigma a$$

Let $\varepsilon > 0$. We first bound $a^\mathsf{T}\Sigma a$ with the operator norm of Σ. Note that

$$\begin{aligned}
a^\mathsf{T}\Sigma a &= \sum_i a_i(\Sigma a)_i \\
&\leq \sum_i a_i \|\Sigma a\| \leq n\|\Sigma\|_{\text{op}} \|a\|^2
\end{aligned}$$

We define T_+ to be the intersection of half-planes the $\{x|x_i \geq \varepsilon\}$, so that $T_+ = \{x|x_i \geq \varepsilon\ \forall i\}$. Since the half-planes are closed and convex, T_+ is also closed and convex since it is an intersection of closed and convex sets.Then, $T = T_+ \cup \{x : -x \in T_+\} = \{x|x_i \geq \varepsilon\ \forall i \text{ or } x_i \leq -\varepsilon\ \forall i\}$. Note that $x \in T \iff \min_i|x_i| \geq \varepsilon$. We define $X = f_\alpha - f^\pi$ which has zero-mean. Finally, Theorem B.3 states that

$$\mathbb{P}\{X \in T\} = \mathbb{P}\{f_\alpha - f^\pi \in T\} \leq \inf_{a\in A} a^T\Sigma a \leq n\|\Sigma\|_{\text{op}} \inf_{a\in A}\|a\|^2.$$

Note that $\inf_a\|a\|^2$ is bounded since $a = (\frac{1}{n\varepsilon}, \frac{1}{n\varepsilon}, ..., \frac{1}{n\varepsilon})$ is in A and $\|a\|^2 = \frac{1}{(n\varepsilon)^2}$. So $n\inf_{a\in A}\|a\|^2 \leq C$ for some constant C independent of α. From the previous result, we can take the limit of $\alpha \to 0$ of $\|\Sigma\|_{\text{op}} = \left\|\mathbb{E}\left[(f_\alpha - f^\pi)(f_\alpha - f^\pi)^\mathsf{T}\right]\right\|_{\text{op}}$ and obtain:

$$\mathbb{P}\{f_\alpha - f^\pi \in T\} = \mathbb{P}\left\{\min_i|f_\alpha(i) - f^\pi(i)| \geq \varepsilon\right\} \leq C \cdot \left\|\mathbb{E}\left[(f_\alpha - f^\pi)(f_\alpha - f^\pi)^\mathsf{T}\right]\right\|_{\text{op}} \to 0$$

$\square$

Appendix C Proofs of Section 6

Lemma C.1. *Suppose $\pi'(s) = \operatorname{argmax}_a Q^\pi(s,a)$ for each s. Then $K(\pi,\pi') = \mathbb{P}\{\pi'$ is greedy with respect to $\mathcal{G}^\pi\} > 0$.*

We will prove an intermediate probability lemma. Let $X_1, ..., X_n$ be mutually independent random variables bounded in $[a,b]$, and $F_i(x) = \mathbb{P}\{X_i \leq x\}$ denote the cumulative density functions of X_i for $i = 2, .., n$. Note that

$$\begin{aligned}
\mathbb{P}\{X_1 \geq X_2, X_1 \geq X_3, ..., X_1 \geq X_n\} &= \int_a^b \int_a^{x_1} \cdots \int_a^{x_1} \mathrm{d}\mathbb{P}(x_1, ..., x_n) \\
&= \int_a^b \int_a^{x_1} \cdots \int_a^{x_1} \mathrm{d}\mathbb{P}_1(x_1)\mathrm{d}\mathbb{P}_2(x_2)\mathrm{d}\mathbb{P}_n(x_n) \quad \text{by mutual independence} \\
&= \int_a^b F_2(x_1)\cdots F_n(x_1)\mathrm{d}\mathbb{P}_1(x_1) \\
&= \mathbb{E}\left[F_2(X_1)F_3(X_1)\cdots F_n(X_1)\right]. \qquad (14)
\end{aligned}$$

Then, we have:

Lemma C.2. *Suppose that $\mathbb{E}[F_i(X_1)] > 0\ \forall i = 2, ..., n$. Then also*

$$\mathbb{E}\left[F_2(X_1)\cdots F_n(X_1)\right] > 0$$

Proof. It is easy to see that $H(x_1) = \Pi_{i=2}^n F_i(x_1)$ is also a CDF. In particular, H starts at 0, ends at 1, and it monotone and right-continuous. In fact, by Equation (14) it corresponds to the CDF of $\max(X_2, ..., X_n)$. Assume for a contradiction that $\mathbb{E}\left[F_2(X_1)\cdots F_n(X_1)\right] = 0$. By positivity, monotonicity, and right-continuity, we have that $H(x_1) = 0\ \forall x_1 \in [a,b)$. Then, for every x we have

$$H(x) = 0 \implies F_i(x) = 0 \text{ for some } i.$$

Since we have $H(b) = 1$ and $H(x) = 0$ otherwise, note that there must exist one i' such that $F_{i'}(b) = 1$ and $F_{i'}(x) = 0$ otherwise. If not, then for all i there exists a $\varepsilon_i > 0$ such that $F_i(b - \varepsilon_i) > 0$. By monotonicity, $F_i(b - \min_i \varepsilon_i) > 0\ \forall i$, and thus $H(b - \min_i \varepsilon_i) > 0$. Thus we have $\mathbb{E}[F_{i'}(x)] = 0$, a contradiction. □

Proof (*Lemma C.1*). Note that

$$K(\pi,\pi') = \mathbb{P}\{\pi' \text{ is greedy with respect to } \mathcal{G}^\pi\} = \mathbb{P}\{\text{for each } s, \mathcal{G}^\pi(s,\pi'(s)) \geq \mathcal{G}^\pi(s,a)\ \forall a\}.$$

Fix a state s, write $X_i(s) := G^\pi(s,a_i)$, and without loss of generality assume that $\pi'(s) = a_1$. We first show that $\mathbb{E}[F_i(X_1)] > 0$, i.e. $\mathbb{P}\{G^\pi(s,a_1) \geq G^\pi(s,a)\} > 0$ for all a. Suppose that it is not so, and pick a such that $\mathbb{P}\{G^\pi(s,a_1) \geq G^\pi(s,a)\} = 0$. Then

$$\begin{aligned}
Q^\pi(s,a_1) &= \mathbb{E}\left[\mathcal{G}^\pi(s,a_1)\right] \\
&= \mathbb{P}\{\mathcal{G}^\pi(s,a_1) \geq \mathcal{G}^\pi(s,a)\}\mathbb{E}\left[\mathcal{G}^\pi(s,a_1) \mid \{\mathcal{G}^\pi(s,a_1) \geq \mathcal{G}^\pi(s,a)\}\right] \\
&+ \mathbb{P}\{\mathcal{G}^\pi(s,a_1) < \mathcal{G}^\pi(s,a)\}\mathbb{E}\left[\mathcal{G}^\pi(s,a_1) \mid \{\mathcal{G}^\pi(s,a_1) < \mathcal{G}^\pi(s,a)\}\right] \\
&= 0 + \mathbb{E}\left[\mathcal{G}^\pi(s,a_1) | \{\mathcal{G}^\pi(s,a_1) < \mathcal{G}^\pi(s,a)\}\right] \\
&< \mathbb{E}\left[\mathcal{G}^\pi(s,a)\right] = Q^\pi(s,a),
\end{aligned}$$

which contradicts the fact that π' is greedy wrt Q^π. Hence $\mathbb{E}[F_i(X_1)] > 0$, and we apply Lemma C.2 to this set to conclude that for each s,

$$\mathbb{P}\{G^\pi(s,a_1) \geq G^\pi(s,a), \forall a\} > 0.$$

Because the returns are mutually independent, we further know that

$$\mathbb{P}\{G^\pi(s,a_1) \geq G^\pi(s,a), \forall s,a\} = \prod_{s\in\mathcal{S}} \mathbb{P}\{G^\pi(s,a_1) \geq G^\pi(s,a), \forall a\} > 0,$$

completing the proof. □

Appendix D On weak convergence and total variation convergence

Recall the definition of the Total Variation metric:

Definition D.1. The total variation metric between probability measures is defined by:

$$d_{\mathtt{TV}}(\mu,\nu) = \sup_{\mathcal{B}\in \mathrm{Borel}(\mathbb{R}^{\mathrm{d}})} |\mu(A) - \nu(A)|,$$

for $\mu, \nu \in \mathscr{P}(\mathbb{R}^{\mathrm{d}})$.

Consider a bandit with a single arm that has a deterministic reward of 0. Consider any of the classic algorithms covered in this paper, which will sample a target of 0 at every iteration. It is easy to see that the unique stationary distribution of the algorithm in this instance is a Dirac distribution at 0 (denoted δ_0).

Suppose a step-size of $\alpha < 1$. If we initialize with some $f_0 \neq 0$ then we can see that the algorithm will never converge to the true stationary distribution in Total Variation distance. This is because a Dirac distribution at any $x \neq 0$ is always a constant distance of 1 away from a Dirac at 0. In other words,

$$d_{\mathtt{TV}}(\delta_0, \delta_{f_n}) = 1 \quad \forall n$$

despite the fact that $f_n \to 0$. On the other hand, we have

$$\mathcal{W}(\delta_0, \delta_{f_n}) \to 0,$$

since the Wasserstein metric takes into consideration the underlying metric structure of the space.