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 \(W(\mu^{(1)}, \mu^{(2)})\);

(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 \(W(\mu^{(1)} K, \mu^{(2)} K) \leq \mathbb{E} \left[\|f_1^{(1)} - f_2^{(2)}\|\right]\) and bound \(\mathbb{E} \left[\|f_1^{(1)} - f^{(2)}_0\|\right]\) 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 \(\pi\) using Monte Carlo returns. The update rule is given by:

\[
V_{n+1}(s) = (1 - \alpha) V_n(s) + \alpha G_n(s) \tag{MCE}
\]

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

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

**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:

\[
V_1^{(1)}(s) = (1 - \alpha) V_0^{(1)}(s) + \alpha G_n^{(1)}(s) \\
V_1^{(2)}(s) = (1 - \alpha) V_0^{(2)}(s) + \alpha G_n^{(2)}(s).
\]

for the same \(G_n^{(1)}(s)\) \tag{11}

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

\[
W(\mu^{(1)} K, \mu^{(2)} K) \leq \mathbb{E} \left[\|V_1^{(1)} - V_1^{(2)}\|\right]
\]

\[
= \mathbb{E} \left[\|(1 - \alpha) V_0^{(1)} + \alpha G_n^{(1)} - (1 - \alpha) V_0^{(2)} + \alpha G_n^{(2)}\|\right]
\]

\[
= \mathbb{E} \left[\|(1 - \alpha)(V_0^{(1)} - V_0^{(2)})\|\right]
\]

\[
= (1 - \alpha) \mathbb{E} \left[\|V_0^{(1)} - V_0^{(2)}\|\right] = (1 - \alpha) W(\mu^{(1)}, \mu^{(2)}).
\]

Since \(1 - \alpha < 1\), \(K_\alpha\) is a contraction mapping and we are done.

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 \(Q^*\). The updates are given by:

\[
\forall (s, a) \in S \times A : \quad Q_{n+1}(s, a) = (1 - \alpha) Q_n(s, a) + \alpha \left( r + \gamma \max_{a'} Q_n(s', a') \right), \tag{QL}
\]
where \( r \sim \mathcal{R}(s, a), s' \sim \mathcal{P}(s, a) \), and \( \alpha > 0 \).

**Theorem A.2.** For any constant step size \( 0 < \alpha \leq 1 \) and initialization \( Q_0 \sim \mu_0 \in \mathcal{M}(\mathbb{R}^{|S| \times |A|}) \), the sequence of random variables \( (Q_n)_{n \geq 0} \) defined by the recursion (QL) converges in distribution to a unique stationary distribution \( \xi_0 \in \mathcal{M}(\mathbb{R}^{|S|}) \).

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

\[
Q_1^{(1)}(s, a) = (1 - \alpha)Q_0^{(1)}(s, a) + \alpha \left( r + \gamma \max_{a'} Q_0^{(1)}(s', a') \right)
\]

\[
Q_1^{(2)}(s, a) = (1 - \alpha)Q_0^{(2)}(s, a) + \alpha \left( r + \gamma \max_{a'} Q_0^{(2)}(s', a') \right)
\]

for the same \( r \sim \mathcal{R}(s, a) \), \( s' \sim \mathcal{P}(s, a) \)

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

\[
\mathbb{E} \left[ \| \hat{T}(Q_1^{(1)}) - \hat{T}(Q_2^{(2)}) \| \right] = \mathbb{E} \left[ \max_{s, a} \left| r - r + \gamma \left( \max_{a'} Q_1^{(1)}(s', a') - \max_{a'} Q_2^{(2)}(s', a') \right) \right| \right]
\]

\[
\leq \gamma \mathbb{E} \left[ \max_{s, a} \left| \max_{a'} Q_1^{(1)}(s', a') - \max_{a'} Q_2^{(2)}(s', a') \right| \right]
\]

\[
\leq \gamma \mathbb{E} \left[ \max_{s, a} \left| Q_1^{(1)}(s', a') - Q_2^{(2)}(s', a') \right| \right] = \gamma \mathbb{E} \left[ \| Q_1^{(1)} - Q_2^{(2)} \| \right]
\]

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 \( Q_1^{(1)} \) and \( Q_2^{(2)} \) 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(\( \lambda \))

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

\[
\forall s : V_{n+1}(s) = (1 - \alpha)V_n(s) + \alpha(1 - \lambda) \sum_{k=1}^{\infty} \lambda^{k-1} \left( \sum_{i=0}^{k-1} \gamma^i r_i(s_i, a_i) + \gamma^k V_n(s_k) \right)
\]

(TD(\( \lambda \)))

where each \( n \)-step trajectory is sampled starting from \( s \) and following policy \( \pi \).

**Theorem A.3.** For any constant step size \( 0 < \alpha \leq 1 \) and initialization \( V_0 \sim \mu_0 \in \mathcal{M}(\mathbb{R}^{|S|}) \), the sequence of random variables \( (V_n)_{n \geq 0} \) defined by the recursion (TD(\( \lambda \))) converges in distribution to a unique stationary distribution \( \xi_0 \in \mathcal{M}(\mathbb{R}^{|S|}) \).

**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.

\[
V_{k+1}^{(1)}(s) = (1 - \alpha)V_k^{(1)}(s) + \alpha(1 - \lambda) \sum_{n=1}^{\infty} \lambda^{n-1} \left( \sum_{i=0}^{n-1} \gamma^i r_i(s_i, a_i) + \gamma^n V_k^{(1)}(s_n) \right)
\]

\[
V_{k+1}^{(2)}(s) = (1 - \alpha)V_k^{(2)}(s) + \alpha(1 - \lambda) \sum_{n=1}^{\infty} \lambda^{n-1} \left( \sum_{i=0}^{n-1} \gamma^i r_i(s_i, a_i) + \gamma^n V_k^{(2)}(s_n) \right)
\]

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 \( \hat{T}(V)(s) = \sum_{k=1}^{\infty} \lambda^{k-1} \left( \sum_{i=0}^{k-1} \gamma^i r_i(s_i, a_i) + \gamma^k V_n(s_k) \right) \) for the empirical Bell-
|\[ E \left[ \left\| \hat{T}(V(1)) - \hat{T}(V(2)) \right\| \right] = E \left[ \max_s \sum_{n=1}^{\infty} \lambda^{n-1} R_n^{(1)} - \sum_{n=1}^{\infty} \lambda^{n-1} R_n^{(2)} \right] \\
= E \left[ \max_s \sum_{n=1}^{\infty} \lambda^{n-1} \left( R_n^{(1)} - R_n^{(2)} \right) \right] \\
= E \left[ \max_s \sum_{n=1}^{\infty} \lambda^{n-1} \gamma^n \left( V^{(1)}(s_n) - V^{(2)}(s_n) \right) \right] \quad \text{(reward terms cancel)} \\
\leq E \sum_{n=1}^{\infty} \lambda^{n-1} \gamma^n \left| \max_s \left( V^{(1)}(s_n) - V^{(2)}(s_n) \right) \right| \quad \text{(triangle inequality)} \\
\leq \sum_{n=1}^{\infty} \lambda^{n-1} \gamma^n E \max_s \left| V^{(1)}(s) - V^{(2)}(s) \right| \quad \text{(by the coupling)} \\
= \sum_{n=1}^{\infty} \lambda^{n-1} \gamma^n E \left[ \left| V^{(1)}(s) - V^{(2)}(s) \right| \right] = \gamma \frac{1}{1 - \lambda} \gamma E \left[ \left| V^{(1)}(s) - V^{(2)}(s) \right| \right]. 
\]

Concluding the proof as before, we have \( W(\mu^{(1)} K, \mu^{(2)} K) \leq (1 - \alpha + \alpha \gamma 1 - \lambda) W(\mu^{(1)}, \mu^{(2)}). \) Since \( 1 - \alpha + \alpha \gamma 1 - \lambda \)

### A.4 SARSA with \( \varepsilon \)-greedy policies

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

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

where \( r \sim R(\cdot | s, a) \) and \( s' \sim P(\cdot | s, a) \) 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{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.

\[
E \left[ \left\| \hat{T}(Q^{(1)}) - \hat{T}(Q^{(2)}) \right\| \right] = P \left\{ \text{greedy action chosen} \right\} E \left[ \max_{s,a} \gamma \left| (\max_{a'} Q^{(1)}(s', a') - \max_{a'} Q^{(2)}(s', a')) \right| \right] \\
+ P \left\{ \text{non-greedy action chosen} \right\} E \left[ \max_{s,a} \gamma \left| (Q^{(1)}(s', a') - Q^{(2)}(s', a')) \right| \right] \\
\leq \varepsilon \gamma E \left[ \left\| Q^{(1)} - Q^{(2)} \right\| \right] + (1 - \varepsilon) \gamma E \left[ \left\| Q^{(1)} - Q^{(2)} \right\| \right] \\
= \gamma E \left[ \left\| Q^{(1)} - Q^{(2)} \right\| \right].
\]

The bound \( E \left[ \max_{s,a} \gamma \left| (\max_{a'} Q^{(1)}(s', a') - \max_{a'} Q^{(2)}(s', a')) \right| \right] \leq \gamma E \left[ \left\| Q^{(1)} - Q^{(2)} \right\| \right] \) follows from \( \max_{a'} Q_1(s, a') - \max_{a'} Q_2(s, a') \leq \max_{a'} \left| Q_1(s, a') - Q_2(s, a') \right| \), and since \( Q^{(1)} \) and \( Q^{(2)} \) sampled the
functions. Thus we set the state space to be

\[ \text{In both cases, we have} \]

\[ \text{Analogously, the update for} \]

\[ \text{In this example we will have to modify our state-space and introduce a new metric on pairs of} \]

\[ \text{A.6 Double Q-Learning} \]

\[ \text{Concluding the proof as before, we have that} \]

\[ \text{Theorem A.5. For any constant step size} \]

\[ \text{Proof. We jump straight to step (P2) of the proof template. We use the same-sampling coupling.} \]

\[ \text{We write} \]

\[ \text{We omit the subscripts on the} \]

\[ \text{This is the original algorithm, not the deep reinforcement learning version given in} \]

\[ \text{A.5 Expected SARSA with} \]

\[ \text{In this example we examine the Expected SARSA updates with} \]

\[ \text{Define} \]

\[ \text{The updates are as follow:} \]

\[ \text{where} \]

\[ \text{The bound follows similarly to the examples of} \]

\[ \text{We write} \]

\[ \text{The sequence of} \]

\[ \text{Concluding the proof as before, we have that} \]

\[ \text{A.6 Double Q-Learning} \]

\[ \text{In this example we will have to modify our state-space and introduce a new metric on pairs of} \]

\[ \text{The Double Q-Learning algorithm (Hasselt, 2010)\textsuperscript{1} maintains two random estimates} \]

\[ \text{Should} \]

\[ \text{Analogously, the update for} \]

\[ \text{In both cases, we have} \]
Theorem A.6. For any constant step size $0 < \alpha \leq 1$ and initialization $(Q_0^A, Q_0^B) \sim \mu_0 \in \mathcal{M}(\mathbb{R}^{[S] \times [A]} \times \mathbb{R}^{[S] \times [A]})$, 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 $\chi_\alpha \in \mathcal{M}(\mathbb{R}^{[S] \times [A]} \times \mathbb{R}^{[S] \times [A]})$.

Proof. As before, let $\mu^{(1)}, \mu^{(2)} \mathcal{M}(\mathbb{R}^{[S] \times [A]} \times \mathbb{R}^{[S] \times [A]})$ 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

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

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

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

Putting everything together, the full expectation is:

$$\mathbb{E} \left[ d((Q_1^A, Q_1^B), (R_1^A, R_1^B)) \right] = \mathbb{E} \left[ \|Q_1^A - R_1^A\| + \|Q_1^B - R_1^B\| \right]$$

$$= \mathbb{P} \{ A \text{ is updated} \} \mathbb{E} \left[ \|Q_1^A - R_1^A\| + \|Q_1^B - R_1^B\| \right]$$

$$+ \mathbb{P} \{ B \text{ is updated} \} \mathbb{E} \left[ \|Q_1^A - R_1^A\| + \|Q_1^B - R_1^B\| \right]$$

$$= p \mathbb{E} \left[ \|Q_1^A - R_1^A\| + \|Q_1^B - R_1^B\| \right]$$

$$+ (1 - p) \mathbb{E} \left[ \|Q_1^A - R_1^A\| + \|Q_1^B - R_1^B\| \right]$$

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

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

$$\leq \frac{1}{2} \left\{ 2 + \alpha \gamma - \alpha \right\} \mathbb{E} \left[ \|Q_0^A - R_0^A\| \right] + \mathbb{E} \left[ \|Q_0^B - R_0^B\| \right] \right\} \quad (p = \frac{1}{2})$$

$$= \frac{1}{2} \left\{ 2 + \alpha \gamma - \alpha \right\} \mathbb{E} \left[ d((Q_0^A, Q_0^B), (R_0^A, R_0^B)) \right]$$

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

Proof. Let $f_0$ be distributed according to $\psi_\alpha$. Rewriting equation (5):

$$f_1 = (1 - \alpha)f_0 + \alpha T^\pi f_0 + \alpha \xi(f_0),$$

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

$$\bar{f}_\alpha = (1 - \alpha)f_\alpha + \alpha \mathbb{E}[T^\pi f_0]$$

$$\alpha \bar{f}_\alpha = \alpha \mathbb{E} [R^\pi + \gamma P^\pi f_0]$$

$$\bar{f}_\alpha = R^\pi + \gamma P^\pi \mathbb{E}[f_0]$$

$$f_\alpha = T^\pi f_\alpha$$

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

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

$$C(f) := \mathbb{E}_\omega [(\hat{T}^\pi(f, \omega) - T^\pi f)(\hat{T}^\pi(f, \omega) - T^\pi f)^T]$$
to be the covariance of the zero-mean noise term $\tilde{T}_\pi(f, \omega) - T_\pi f$ for a given function $f$. Then, the covariance of $f_\alpha \sim \psi_\alpha$ is given by

$$
(1 - (1 - \alpha)^2) \mathbb{E} \left[ (f_\alpha - f^\pi)(f_\alpha - f^\pi)^T \right] = \alpha^2 (\gamma P^\pi) \mathbb{E} \left[ (f_\alpha - f^\pi)(f_\alpha - f^\pi)^T \right] (\gamma P^\pi)^T \\
+ \alpha (1 - \alpha) (\gamma P^\pi) \mathbb{E} \left[ (f_0 - f^\pi)(f_0 - f^\pi)^T \right] \\
+ \alpha (1 - \alpha) \mathbb{E} \left[ (f_\alpha - f^\pi)(f_\alpha - f^\pi)^T (\gamma P^\pi)^T \right] \\
+ \alpha^2 \int C(f) \psi_\alpha(df)
$$

Furthermore, we have that $\| \mathbb{E} \left[ (f_\alpha - f^\pi)(f_\alpha - f^\pi)^T \right] \|_o$ is monotonically decreasing with respect to $\alpha$, where $\| \cdot \|_o$ denotes the operator norm of a matrix. In particular, $\lim_{\alpha \to 0} \| \mathbb{E}((f_\alpha - f^\pi)(f_\alpha - f^\pi)^T) \|_o = 0$, 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 $\xi(f) := (\tilde{T}_\pi(f, \omega) - T_\pi f)$. In the same setup as Theorem 5.2:

$$
\mathbb{E} \left[ (f_\alpha - f^\pi)(T_\pi f_\alpha - f^\pi + \xi(f_0))^T \right] = \mathbb{E} \left[ (f_\alpha - f^\pi)(f_\alpha - f^\pi)^T \right] (\gamma P^\pi)^T \\
+ \mathbb{E} \left[ ((T_\pi f_\alpha - f^\pi) + \xi(f_\alpha)) ((T_\pi f_\alpha - f^\pi) + \xi(f_\alpha))^T \right] = (\gamma P^\pi) \mathbb{E} \left[ (f_\alpha - f^\pi)(f_\alpha - f^\pi)^T \right] (\gamma P^\pi)^T \\
+ \int C(v) \psi_\alpha(dv)
$$

**Proof.** Let $f_0 \sim \psi_\alpha$, by (5) we have $f_1 = (1 - \alpha) f_0 + \alpha (T_\pi f_0 + \xi(f_0))$ and $f_1 \sim \psi_\alpha$. Furthermore, the distribution of $f_0$ is independent of the distribution of $\omega$. By independence,

$$
\mathbb{E} \left[ (f_0 - f^\pi \xi(f_0)^T \right] = \mathbb{E}_{f_0} \mathbb{E}_\omega \left[ (f_0 - f^\pi \xi(f_0)^T \right] = \mathbb{E}_{f_0} \left[ (f_0 - f^\pi) (\mathbb{E}_\omega \xi(f_0)^T \right] = 0 \\
(\mathbb{E}_\omega [\xi(f)] = 0 \text{ for every } f)
$$

For the first identity, note that

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

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

For the second identity, expanding the outer product gives:

$$
\mathbb{E} \left[ ((T_\pi f_0 - f^\pi) + \xi(f_0)) ((T_\pi f_0 - f^\pi) + \xi(f_0))^T \right] = \mathbb{E} \left[ (T_\pi f_0 - f^\pi)(T_\pi f_0 - f^\pi)^T \right] \\
+ \mathbb{E} \left[ (\xi(f_0))(\xi(f_0))^T \right] \\
+ \mathbb{E} \left[ (T_\pi f_0 - f^\pi)(\xi(f_0))^T \right] \\
+ \mathbb{E} \left[ (\xi(f_0))(T_\pi f_0 - f^\pi)^T \right] \\
= \mathbb{E} \left[ (\gamma P^\pi(f_0 - f^\pi))(\gamma P^\pi(f_0 - f^\pi))^T \right] \\
+ \int C(v) \psi_\alpha(dv)
$$

$$
= (\gamma P^\pi) \mathbb{E} \left[ (f_0 - f^\pi)(f_0 - f^\pi)^T \right] (\gamma P^\pi)^T \\
+ \int C(v) \psi_\alpha(dv)
$$
where we used $E [(T^\pi f_0 - f^\pi)(\xi(f_0)) T] = 0$.

**Proof (of Theorem 5.2).** Again let $f_0$ be distributed according to $\psi_\alpha$. Subtracting $f^\pi$ from equation (12),

$$f_1 - f^\pi = (1 - \alpha)(f_0 - f^\pi) + \alpha (T^\pi f_0 - f^\pi + \xi(f_0)).$$

and taking outer products:

$$(f_1 - f^\pi)(f_1 - f^\pi)^T = (1 - \alpha)^2 (f_0 - f^\pi)(f_0 - f^\pi)^T + \alpha^2 (T^\pi f_0 - f^\pi + \xi(f_0))(T^\pi f_0 - f^\pi + \xi(f_0))^T + \alpha(1 - \alpha)(f_0 - f^\pi)(T^\pi f_0 - f^\pi + \xi(f_0))^T + \alpha(1 - \alpha)(T^\pi f_0 - f^\pi + \xi(f_0))(f_0 - f^\pi)^T.$$  

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

$$E [(f_1 - f^\pi)(f_1 - f^\pi)^T] = (1 - \alpha)^2 E [(f_0 - f^\pi)(f_0 - f^\pi)^T] + \alpha^2 E [(\gamma P^\pi)(f_0 - f^\pi)](\gamma P^\pi)^T + \alpha(1 - \alpha)E [(f_0 - f^\pi)(f_0 - f^\pi)^T] \gamma P^\pi)^T + \alpha(1 - \alpha)E [(f_0 - f^\pi)(f_0 - f^\pi)^T] \gamma P^\pi)^T + \alpha(1 - \alpha)E [(f_0 - f^\pi)(f_0 - f^\pi)^T] \gamma P^\pi)^T + \alpha^2 \int C(f) \psi_\alpha(df).$$

Since $E [(f_1 - f^\pi)(f_1 - f^\pi)^T] = E [(f_0 - f^\pi)(f_0 - f^\pi)^T]$ by stationarity, re-arranging to the LHS and factoring gives:

$$(1 - (1 - \alpha)^2)E [(f_\alpha - f^\pi)(f_\alpha - f^\pi)^T] = \alpha^2 (\gamma P^\pi) \otimes (\gamma P^\pi)^T \gamma P^\pi)^T + \alpha(1 - \alpha)E [(f_0 - f^\pi)(f_0 - f^\pi)^T] \gamma P^\pi)^T + \alpha(1 - \alpha)E [(f_0 - f^\pi)(f_0 - f^\pi)^T] \gamma P^\pi)^T + \alpha^2 \int C(f) \psi_\alpha(df).$$

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 $x \otimes y = xy^T$. By extension, the tensor product of two matrices $A, B$ is the operator defined by $(A \otimes B)X = AXB^T$. Then, the above expression can be re-written as:

$$(1 - (1 - \alpha)^2)E [(f_\alpha - f^\pi)(f_\alpha - f^\pi)^T] = \alpha^2 (\gamma P^\pi) \otimes (\gamma P^\pi)^T \gamma P^\pi)^T + \alpha(1 - \alpha)E [(f_0 - f^\pi)(f_0 - f^\pi)^T] \gamma P^\pi)^T + \alpha(1 - \alpha)E [(f_\alpha - f^\pi)(f_\alpha - f^\pi)^T] \gamma P^\pi)^T + \alpha^2 \int C(f) \psi_\alpha(df).$$

Factoring the tensor products further gives:

$$\left[ I - ((1 - \alpha)I + \alpha \gamma P^\pi) \otimes 2 \right] E [(f_\alpha - f^\pi) \otimes 2] = \alpha^2 \int C(f) \psi_\alpha(df).$$

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

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

(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 \( \| A^{-1} \| \leq \frac{1}{1 - \| I - A \|} \). We let \( A = I - ((1 - \alpha)I + \alpha\gamma P^\pi)^{\otimes 2} \), by the calculation in (13) we have \( \| I - A \| < 1 \). So we calculate the operator norm of the covariance matrix:

\[
\| \mathbb{E} [(f_0 - f^\pi)(f_0 - f^\pi)^T] \| = \alpha^2 \left\| \left[ I - ((1 - \alpha)I + \alpha\gamma P^\pi)^{\otimes 2} \right]^{-1} \right\| \int \mathcal{C}(v)\psi_\alpha(dv) \\
\leq \alpha^2 \left\| \left[ I - ((1 - \alpha)I + \alpha\gamma P^\pi)^{\otimes 2} \right]^{-1} \right\| \int \mathcal{C}(v)\psi_\alpha(dv) \\
\leq \alpha^2 \frac{1}{1 - \| I - ((1 - \alpha)I + \alpha\gamma P^\pi)^{\otimes 2} \|} \int \mathcal{C}(v)\psi_\alpha(dv) \\
= \alpha^2 \frac{1}{1 - \| (1 - \alpha)I + \alpha\gamma P^\pi \|} \int \mathcal{C}(v)\psi_\alpha(dv) \\
\leq \alpha^2 \frac{1}{1 - (1 - \alpha + \alpha\gamma)^2} \int \mathcal{C}(v)\psi_\alpha(dv)
\]

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

\[
\| \mathbb{E} [(f_0 - f^\pi)(f_0 - f^\pi)^T] \| \leq M \frac{\alpha^2}{1 - (1 - \alpha + \alpha\gamma)^2} \xrightarrow{\alpha \to 0} 0
\]

For the concentration inequality, we will use a multivariate Chebyshev inequality (Marshall and Olkin, 1960, Theorem 3.1), whose 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 \( A = \{ a \in \mathbb{R}^n : \langle a, x \rangle \geq 1 \ \text{for} \ x \in T_+ \} \), then

\[
\mathbb{P} \{ X \in T \} \leq \inf_{a \in A} \mathbb{a}^T\Sigma a
\]

Let \( \varepsilon > 0 \). We first bound \( a^T\Sigma a \) with the operator norm of \( \Sigma \). Note that

\[
a^T\Sigma a = \sum_i a_i (\Sigma a)_i \\
\leq \sum_i a_i \| \Sigma a \| \leq n \| \Sigma \|_{\text{op}} \| a \|^2
\]

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

\[
\mathbb{P} \{ f_\alpha - f^\pi \in T \} = \mathbb{P} \left\{ \min_i |f_\alpha(i) - f^\pi(i)| \geq \varepsilon \right\} \leq C \cdot \| \mathbb{E} [(f_\alpha - f^\pi)(f_\alpha - f^\pi)^T] \|_{\text{op}} \to 0
\]
Appendix C  Proofs of Section 6

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

We will prove an intermediate probability lemma. Let $X_1, \ldots, X_n$ 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 $i = 2, \ldots, n$. Note that

$$
\mathbb{P}\{X_1 \geq X_2, X_1 \geq X_3, \ldots, X_1 \geq X_n\} = \int_a^b \int_a^{x_1} \cdots \int_a^{x_n} d\mathbb{P}(x_1, \ldots, x_n)
= \int_a^b \int_a^{x_1} \cdots \int_a^{x_{n-1}} d\mathbb{P}_1(x_1) d\mathbb{P}_2(x_2) d\mathbb{P}_n(x_n)
\quad \text{by mutual independence}
= \int_a^b F_2(x_1) \cdots F_n(x_1) d\mathbb{P}_1(x_1)
= \mathbb{E}[F_2(X_1) F_3(X_1) \cdots F_n(X_1)].
$$

(14)

Then, we have:

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

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

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

Proof (Lemma C.1). Note that

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

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 $\mathbb{E}[F_1(X_1)] > 0$, 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

$$
Q^\pi(s, a_1) = \mathbb{E}[G^\pi(s, a_1)]
= \mathbb{P}\{G^\pi(s, a_1) \geq G^\pi(s, a)\} \mathbb{E}[G^\pi(s, a_1) | \{G^\pi(s, a_1) \geq G^\pi(s, a)\}]
+ \mathbb{P}\{G^\pi(s, a_1) < G^\pi(s, a)\} \mathbb{E}[G^\pi(s, a_1) | \{G^\pi(s, a_1) < G^\pi(s, a)\}]
= 0 + \mathbb{E}[G^\pi(s, a_1)|\{G^\pi(s, a_1) < G^\pi(s, a)\}]
< \mathbb{E}[G^\pi(s, a)] = Q^\pi(s, a),
$$

which contradicts the fact that $\pi'$ is greedy wrt $Q^\pi$. Hence $\mathbb{E}[F_1(X_1)] > 0$, 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

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

completing the proof. \qed
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:

\[
d_{TV}(\mu, \nu) = \sup_{B \in \mathcal{B}(\mathbb{R}^d)} |\mu(A) - \nu(A)|,
\]

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

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 \(\delta_0\)).

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,

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

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

\[
W(\delta_0, \delta_{f_n}) \to 0,
\]

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