A  Notation

\( \mathbb{N} \)  The set of nonnegative integers.
\( \mathbb{R} \)  The set of real numbers.
\( \mathbb{E} \)  The expectation operator.
\( \Pr \)  \( \Pr(x \mid y) \) denotes the probability of the stochastic variable \( x \) given \( y \).
\( \arg \max \)  \( x \in \Pi \) denotes an element \( x \in \Pi \) that maximizes the function \( f \).
\( \arg \max \)  \( \arg \max_{x \in \Pi} f(x) \) denotes an element \( x \in \Pi \) that maximizes the function \( f \).
\( \geq \)  For \( \lambda = (\lambda_1, ..., \lambda_J) \), \( \lambda \geq 0 \) denotes that \( \lambda_i \geq 0 \) for \( i = 1, ..., J \).
\( 1_X(x) \)  \( 1_X(x) = 1 \) if \( x \in \{X\} \) and 0 otherwise.
\( \mathbf{1}_n \)  \( \mathbf{1}_n = (1, 1, ..., 1) \in \mathbb{R}^n \).
\( N(t, s, a, b) \)  \( N(t, s, a, b) = \sum_{k=1}^{\ell} 1_{(s,a,b)}(s_k, a_k, b_k) \).
\( e \)  \( e : (s, a, o) \mapsto 1 \).
\( |S| \)  Denotes the number of elements in \( S \).

B  Examples

The following example from [Altman (1999)] describes in more detail a model where we have a Markov decision process with constraints and where the agent doesn't have model knowledge.

**Example 1** (Altman, 1999). Consider a discrete time single-server queue with a buffer of finite size \( L \). For a given time slot, we assume that at most one customer may join the system. The state of the system at a given time slot is the number of customers in the queue. There is a delay cost \( c(s) \) given a state \( s \in \{S_1, ..., S_n\} \) which one would like to keep as low as possible. The cost of a service to be completed is \( a^1 \), where \( 1/a_1 = \text{Quality of Service (QoS)} \). The probability of a queue arrival at time \( t \) is \( a^2 \). The actions are given by \( a^1 \) and \( a^2 \).

Let \( c^1(a^1) \) be the cost to complete the service (\( c^1 \) is increasing in \( a^1 \)). \( c^1 \) should be bounded by some value \( v^1 \). There is a cost corresponding to the throughput, \( c^2(a^2) \) (\( c^2 \) is decreasing in \( a^2 \)). \( c^2 \) should be bounded by some value \( v^2 \). We assume that the number of actions is finite and actions sets are given by \( a^1 \in \{A^1_1, ..., A^1_j\} \) and \( a^2 \in \{A^2_1, ..., A^2_j\} \) where \( 0 < A^1_1 \leq \cdots \leq A^1_j \leq 1 \) and \( 0 \leq A^2_1 \leq \cdots \leq A^2_j \leq 1 \). The transition probability \( P(s_{k+1}, s_k, a^1_k, a^2_k) \) from state \( s_k \) to \( s_{k+1} \) given actions \( a^1_k \) and \( a^2_k \) is given by

\[
P(s_{k+1}, s_k, a^1_k, a^2_k) = \begin{cases} 
(1 - a^2)a^1 & \text{if } L \geq s \geq 1, \\
0 & \text{if } L = s = 1, \\
a^2a^1 + (1 - a^2)(1 - a^1) & \text{if } L \geq s \geq 1, \\
0 & \text{if } L = s = 0, \\
a^2(1 - a^1) & \text{if } L \geq s \geq 0, \\
0 & \text{if } L = s = 0, \\
1 - a^2(1 - a^1) & \text{if } L \geq s \geq 0, \\
0 & \text{if } L = s = 0.
\end{cases}
\]

For \( \gamma \in (0, 1) \), the constrained Markov decision process problem is given by

\[
\min_{\pi^1, \pi^2} \mathbb{E} \left( \sum_{k=0}^{\infty} \gamma^k c^1(s_k) \right) \\
s. t. \mathbb{E} \left( \sum_{k=0}^{\infty} \gamma^k c^1(\pi^1(s_k)) \right) \leq v^1 \\
\mathbb{E} \left( \sum_{k=0}^{\infty} \gamma^k c^2(\pi^2(s_k)) \right) \leq v^2, \tag{28}
\]

For \( \gamma \in (0, 1) \), the constrained Markov decision process problem is given by

\[\min_{\pi^1, \pi^2} \mathbb{E} \left( \sum_{k=0}^{\infty} \gamma^k c^1(s_k) \right)\]

s. t. \[\mathbb{E} \left( \sum_{k=0}^{\infty} \gamma^k c^1(\pi^1(s_k)) \right) \leq v^1\]

\[\mathbb{E} \left( \sum_{k=0}^{\infty} \gamma^k c^2(\pi^2(s_k)) \right) \leq v^2, \tag{28}\]
Proof.  where converges to zero with probability 1 under the following assumptions:

**Proposition 2.** The proof will rely on the following result.

**Proof of Theorem 1**

A certain lower bound

One solution is to define a measure for the quality of service for the user, where it wants to maximize the rewards for the user and itself.

The search engine has multiple objectives here where it wants to maximize the rewards for the user and itself. One solution is to define a measure for the quality of service for the user, $R_u$, and at the same time satisfy a certain lower bound $R^c$ of its own reward, that is

\[
\begin{aligned}
\text{find } & \pi \\
\text{s. t. } & \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} E(u_i A_{\pi(i)}) \geq R^c \\
& \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} E(v_i A_{\pi(i)}) \geq R^u
\end{aligned}
\]

**Example 2 (Search Engine).** In a search engine, there is a number of documents that are related to a certain query. There are two values that are related to every document, the first being a (advertisement) value $u_i$ of document $i$ for the search engine and the second being a value $v_i$ for the user (could be a measure of how strongly related the document is to the user query). The task of the search engine is to display the documents in a row some order, where each row has an attention value, $A_j$ for row $j$. We assume that $u_i$ and $v_i$ are known to the search engine for all $i$, whereas the attention values $\{A_j\}$ are not known. The strategy $\pi$ of the search engine is to display document $i$ in position $j$, $\pi(i) = j$, with probability $p_{ij}$. Thus, the expected average reward for the search engine is

\[
R^c = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} E(u_i A_{\pi(i)})
\]

and for the user

\[
R^u = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} E(v_i A_{\pi(i)}).
\]

The search engine has multiple objectives here where it wants to maximize the rewards for the user and itself. One solution is to define a measure for the quality of service for the user, $R_u \geq R^u$ and at the same time satisfy a certain lower bound $R^c$ of its own reward, that is

**C Proof of Theorem 1**

The proof will rely on the following result.

**Proposition 2.** The random process $\{\Delta_k\}$ taking values in $\mathbb{R}$ and defined as

\[
\Delta_{k+1}(x) = (1 - \alpha_k(x)) \Delta_k(x) + \alpha_k(x) F_k(x)
\]

converges to zero with probability 1 under the following assumptions:

1. For all $x$, $0 \leq \alpha_k(x) \leq 1$, $\sum_k \alpha_k(x) = \infty$, and $\sum_k \alpha_k^2(x) < \infty$
2. $\|E(F_k(x) | \mathcal{F}_k)\|_{\infty} \leq \gamma \|\Delta_k\|_{\infty}$, with $\gamma < 1$
3. $E(F_k - E(F_k(x) | \mathcal{F}_k))^2 \leq C(1 + \|\Delta_k\|_{\infty}^2)$, for some constant $C > 0$

where $\mathcal{F}_k$ is the sigma algebra $\sigma(\Delta_t, F_{t-1}, \alpha_{t-1}, t \leq k)$.

**Proof.** Consult [Jaakkola et al. 1994].
Now let \( \Delta_k(s, a, o) = Q_k(s, a, o) - Q^*(s, a, o) \)

Subtracting \( Q^* \) from the right and left hand sides of the second equality in (12) implies that

\[
\Delta_{k+1}(s, a, o) = (1 - \alpha(s, a, o))\Delta_k(s, a, o) + \alpha(s, a, o)R(s, a, o) + \gamma E(Q_k(s_+, \pi_k(s_+), o)) - Q^*(s, a, o).
\]

We will show that \( \Delta_k \) satisfies the conditions of Proposition 2. Introduce the sigma algebra \( \mathcal{F}_k = \sigma(\Delta_t, F_{t-1}, a_{t-1}, t \leq k) \).

Define

\[
F_k(s, a, o) = 1_{(s, a, o)}(s_k, a_k, o_k) \times (R(s, a, o) + \gamma E(Q_k(s_+, \pi_k(s_+), o)) - Q^*(s, a, o))
\]

If \( (s, a, o) \neq (s_k, a_k, o_k) \), then \( F_k(s, a, o) = 0 \). Else,

\[
E(F_k(s, a, o) \mid \mathcal{F}_k) = \sum_{s_+} P(s, a, s_+) 1_{(s, a, o)}(s_k, a_k, o_k) \times (R(s, a, o) + \gamma E(Q_k(s_+, \pi_k(s_+), o)) - Q^*(s, a, o))
\]

\[
= \sum_{s_+} P(s, a, s_+) (R(s, a, o) + \gamma E(Q_k(s_+, \pi_k(s_+), o))) - R(s, a, o) - \gamma E(Q^*(s_+, \pi^*(s_+), o))
\]

\[
= \gamma \sum_{s_+} P(s, a, s_+) (E(Q_k(s_+, \pi_k(s_+), o)) - E(Q^*(s_+, \pi^*(s_+), o)))
\]

If \( E(Q_k(s_+, \pi_k(s_+), o_k)) \geq E(Q^*(s_+, \pi^*(s_+), o_k)) \), then

\[
\begin{align*}
&|E[Q_k(s_+, \pi_k(s_+), o_k)] - E[Q^*(s_+, \pi^*(s_+), o_k)]| \\
&= E(Q_k(s_+, \pi_k(s_+), o_k)) - E(Q^*(s_+, \pi^*(s_+), o_k)) \\
&= R(s, a, o_k) + E(Q_k(s_+, \pi_k(s_+), o_k)) - R(s, a, o_k) - E(Q^*(s_+, \pi^*(s_+), o_k)) \\
&\leq R(s, a, o_k) + E(Q_k(s_+, \pi_k(s_+), o_k)) - R(s, a, o^*) - E(Q^*(s_+, \pi^*(s_+), o^*)) \\
&\leq R(s, a, o^*) + E(Q_k(s_+, \pi_k(s_+), o^*)) - R(s, a, o^*) - E(Q^*(s_+, \pi^*(s_+), o^*)) \\
&= |E(Q_k(s_+, \pi_k(s_+), o^*)) - Q^*(s_+, \pi_k(s_+), o^*)| \\
&\leq \max_{s, a, o} |Q_k(s, a, o) - Q^*(s, a, o)| \\
&= \|Q_k - Q^*\|_\infty.
\end{align*}
\]

Else, if \( E(Q_k(s_+, \pi_k(s_+), o_k)) \leq E(Q^*(s_+, \pi^*(s_+), o_k)) \), then

\[
\begin{align*}
&|E[Q_k(s_+, \pi_k(s_+), o_k)] - E[Q^*(s_+, \pi^*(s_+), o_k)]| \\
&= E(Q^*(s_+, \pi^*(s_+), o_k)) - E(Q_k(s_+, \pi_k(s_+), o_k)) \\
&\leq E(Q^*(s_+, \pi^*(s_+), o_k)) - E(Q_k(s_+, \pi_k(s_+), o_k)) \\
&= |E(Q_k(s_+, \pi_k(s_+), o_k) - Q^*(s_+, \pi^*(s_+), o_k)| \\
&\leq \max_{s, a, o} |Q_k(s, a, o) - Q^*(s, a, o)| \\
&= \|Q_k - Q^*\|_\infty.
\end{align*}
\]
Thus,
\[
\|E(F_k(s, a, o))\|_\infty = \\
= \gamma \max_{s, a, o} \left| \sum_{s_+} P(s, a, s_+) (E(Q_k(s_+, \pi_k(s_+), \phi_k(s_+))) - E(Q^*(s_+, \pi^*(s_+), \phi_k(s_+))) \right| \\
\leq \gamma \max_{s, a, o} \left| \sum_{s_+} P(s, a, s_+) |(E(Q_k(s_+, \pi_k(s_+), \phi_k(s_+))) - E(Q^*(s_+, \pi^*(s_+), \phi_k(s_+)))| \right| \\
\leq \gamma \max_{s, a, o} \sum_{s_+} P(s, a, s_+ ||Q_k - Q^*||_\infty \\
= \gamma \|Q_k - Q^*\|_\infty \\
= \gamma \|\Delta_k\|_\infty
\]
where the first inequality follows from the triangle inequality and the fact that \(P(s, a, s_+) \geq 0\). Also, we have that
\[
E(F_k - E(F_k | F_k))^2 = \\
= \gamma^2 E \left( Q_k(s_+, \pi_k(s_+), \phi_k(s_+)) - Q^*(s_+, \pi^*(s_+), \phi_k(s_+)) - \\
- \sum_{s_+} P(s, a, s_+) (Q_k(s_+, \pi_k(s_+), \phi_k(s_+)) - Q^*(s_+, \pi^*(s_+), \phi_k(s_+))) \right)^2 \\
= \gamma^2 E \left( \Delta_k(s_+, \pi_k(s_+), \phi_k(s_+)) - \\
- \sum_{s_+} P(s, a, s_+) (\Delta_k(s_+, \pi_k(s_+), \phi_k(s_+))) \right)^2 \\
\leq C(1 + \|\Delta_k\|^2_\infty).
\]
Thus, \(\Delta_k = Q_k - Q^*\) satisfies the conditions of Proposition \ref{prop:convergence} and hence converges to zero with probability 1, i.e. \(Q_k\) converges to \(Q^*\) with probability 1.

### D Proof of Theorem \ref{thm:operator}

**Lemma 1.** Let the operator \(T\) be given by
\[
(TQ)(s, a, o) = \sum_{s_+} P(s, a, s_+) \max_{\pi \in \Pi, o \in \mathcal{O}} \min \{ R(s, a, o) + E(Q(s_+, \pi(s_+), o)) \}.
\]
(31)

Then,
\[
\|TQ_1 - TQ_2\|_\infty \leq \|Q_1 - Q_2\|_\infty.
\]

**Proof.**
\[
\|TQ_1 - TQ_2\|_\infty = \\
= \max_{s, a, o} \left| \sum_{s_+} P(s, a, s_+) \left( \max_{\pi \in \Pi, o \in \mathcal{O}} \min \{ R(s, a, o) + E(Q_1(s_+, \pi(s_+), o)) \} - \\
- \max_{\pi \in \Pi, o \in \mathcal{O}} \min \{ R(s, a, o) + E(Q_2(s_+, \pi(s_+), o)) \} \right) \right| \\
\leq \max_{s, a, o} \sum_{s_+} P(s, a, s_+) \left| \max_{\pi \in \Pi, o \in \mathcal{O}} \min \{ R(s, a, o) + E(Q_1(s_+, \pi(s_+), o)) \} - \\
- \max_{\pi \in \Pi, o \in \mathcal{O}} \min \{ R(s, a, o) + E(Q_2(s_+, \pi(s_+), o)) \} \right| \\
\]
(32)
where the last inequality follows from the triangle inequality and the fact that \( P(s, a, s_+) \geq 0 \). Without loss of generality, assume that

\[
\max_{\pi \in \Pi} \min_{o \in O} \left( R(s, a, o) + \mathbb{E}(Q_1(s_+, \pi(s_+), o)) \right) \\
\geq \max_{\pi \in \Pi} \min_{o \in O} \left( R(s, a, o) + \mathbb{E}(Q_2(s_+, \pi(s_+), o)) \right).
\]

Introduce

\[ (\pi_i, o_i) = \arg \max_{\pi \in \Pi} \min_{o \in O} R(s, a, o) + \mathbb{E}(Q_i(s_+, \pi(s_+), o)). \]

Then,

\[
\max_{\pi \in \Pi} \min_{o \in O} \left( R(s, a, o) + \mathbb{E}(Q_1(s_+, \pi(s_+), o)) \right) - \\
\min_{\pi \in \Pi} \max_{o \in O} \left( R(s, a, o) + \mathbb{E}(Q_2(s_+, \pi(s_+), o)) \right) \\
= \max_{\pi \in \Pi} \min_{o \in O} \left( R(s, a, o) + \mathbb{E}(Q_1(s_+, \pi(s_+), o)) \right) - \\
\min_{\pi \in \Pi} \max_{o \in O} \left( R(s, a, o) + \mathbb{E}(Q_2(s_+, \pi(s_+), o)) \right)
\]

Combining (32)–(36) implies that

\[
\| TQ_1 - TQ_2 \|_\infty \leq \max_{s_+, a} \sum_{s, a, o} P(s, a, s_+) \| Q_1 - Q_2 \|_\infty \\
= \| Q_1 - Q_2 \|_\infty
\]

and the proof is complete. \( \square \)

**Lemma 2.** The operator \( T \) given by (31) is a span semi-norm, that is

\[
\| TQ_1 - TQ_2 \|_s \leq \| Q_1 - Q_2 \|,
\]

where

\[
\| Q \|_s \triangleq \max_{s, a, o} Q(s, a, o) - \min_{s, a, o} Q(s, a, o).
\]

**Proof.** We start off by noting the trivial inequalities

\[
\max_{s', a', o'} \left( Q_1(s', a', o') - Q_2(s', a', o') \right) \\
\geq Q_1(s_+, a_+, o) - Q_2(s_+, a_+, o) \\
\geq \min_{s', a', o'} \left( Q_1(s', a', o') - Q_2(s', a', o') \right).
\]

Also, let

\[ o_i = \arg \min_{o \in O} R(s, a, o) + Q_i(s_+, \pi(s_+), o) \]

and

\[ a_i = \arg \max_{a \in A} Q_i(s, a, o), \quad i \neq j. \]
The definition of the span semi-norm implies that
\[
\|TQ_1 - TQ_2\|_s = \\
= \sum_{s,a} P(s, a, s_+) \left( \max_{\pi \in \Pi \ o \in O} (R(s, a, o) + \mathbb{E}(Q_1(s, \pi(s), o))) - \\
- \max_{\pi \in \Pi \ o \in O} (R(s, a, o) + \mathbb{E}(Q_2(s, \pi(s), o))) \right) \\
= \max_{s,a,o} \sum_{s,a} P(s, a, s_+) \left( \max_{\pi \in \Pi \ o \in O} (R(s, a, o) + \mathbb{E}(Q_1(s, \pi(s), o))) - \\
- \max_{\pi \in \Pi \ o \in O} (R(s, a, o) + \mathbb{E}(Q_2(s, \pi(s), o))) \right)
\]
\[
\leq \max_{s,a,o} \sum_{s,a} P(s, a, s_+) \left( \max_{\pi \in \Pi \ o \in O} (R(s, a, o2) + \mathbb{E}(Q_1(s, \pi(s), o2))) - \\
- \max_{\pi \in \Pi \ o \in O} (R(s, a, o2) + \mathbb{E}(Q_2(s, \pi(s), o2))) \right)
\]
\[
\leq \max_{s,a,o} \sum_{s,a} P(s, a, s_+) (Q_1(s, a1, o2) - Q_2(s, a1, o2)) \\
- \min_{s,a,o} \sum_{s,a} P(s, a, s_+) (Q_1(s, a2, o1) - Q_2(s, a2, o1))
\]
\[
\leq \max_{s,a,o} \sum_{s,a} P(s, a, s_+) \times \max_{s',a',o'} (Q_1(s', a', o') - Q_2(s', a', o')) \\
- \min_{s,a,o} \sum_{s,a} P(s, a, s_+) \times \min_{s',a',o'} (Q_1(s', a', o') - Q_2(s', a', o'))
\]
\[
= \max_{s',a',o'} (Q_1(s', a', o') - Q_2(s', a', o')) - \min_{s',a',o'} (Q_1(s', a', o') - Q_2(s', a', o'))
\]
\[
= \|Q_1 - Q_2\|_s.
\]

For convenience, let \(e: (s, a, o) \rightarrow 1\) be a constant tensor with all elements equal to 1.

**Lemma 3.** Let \(f \in \Phi\) be given, where the set \(\Phi\) is defined as in Definition 2 and let
\[
T'(Q) = T(Q) - f(Q) \cdot e
\]

The ordinary differential equation (ODE)
\[
\dot{Q}(t) = T'(Q(t)) - Q(t)
\]
has a unique globally asymptotically stable equilibrium \(Q^*\), with \(f(Q^*) = v^*\), where \(Q^*\) and \(v^*\) satisfy (17).

**Proof.** Introduce the operator
\[
\tilde{T}(Q) = T(Q) - v \cdot e.
\]
According to lemma 1 we have that
\[ \| TQ_1 - TQ_2 \|_\infty \leq \| Q_1 - Q_2 \|_\infty \]
and hence, T is Lipschitz. It’s easy to verify that
\[ \hat{T}(Q_1) - \hat{T}(Q_2) = T(Q_1) - T(Q_2) \]
and therefore
\[ \| \hat{T}(Q_1) - \hat{T}(Q_2) \|_\infty \leq \| Q_1 - Q_2 \|_\infty, \]
\[ \| \hat{T}(Q_1) - \hat{T}(Q_2) \|_S \leq \| Q_1 - Q_2 \|_S. \]

Now consider the ODE:s
\[ \dot{Q}(t) = \hat{T}(Q(t)) - Q(t) \quad (42) \]
and
\[ \dot{Q}(t) = T'(Q(t)) - Q(t) = \hat{T}(Q(t)) + (v - f(Q)) \cdot e. \quad (43) \]
Note that since T and f are Lipschitz, the ODE:s (42) and (43) are well posed.

Since T is Lipschitz and span semi-norm, the rest of the proof becomes identical to Theorem 3.4 along with Lemma 3.1, 3.2, and 3.3 in (Abounadi et al., 2001b) and hence omitted here.

**Proposition 3** (Borkar & Meyn, 2000: Theorem 2.5). Consider the asynchronous algorithm given by
\[ Q_{k+1} = Q_k + \alpha_k h(Q_k) + M_{k+1} \]
where \( \alpha_k(s, a, o) = 1_{(s, a, o)}(s_k, a_k, o_k) \times \beta_N(k, s, a, o) \cdot v \). Suppose that
1. \( M_k \) is a martingale sequence with respect to the sigma algebra \( F_k = \sigma(Q_t, M_t, t \leq k) \), that is
\[ E(M_{k+1} \mid F_k) = 0 \]
and that there exists a constant \( C_1 > 0 \) such that
\[ E(\| M_{k+1} \|^2 \mid F_k) \leq C_1 (1 + \| Q_k \|^2). \]
2. Assumptions 4 and 5 hold.
3. The limit
\[ h_\infty(X) = \lim_{z \to \infty} \frac{h(zX)}{z} \]
exists.
4. \( \dot{Q}(t) = h(Q(t)) \) has a unique globally asymptotically stable equilibrium \( Q^* \).

Then, \( Q_k \to Q^* \) with probability 1 as \( k \to \infty \) for any initial value \( Q(0) \).

**Proof of Theorem 3**. Introduce the operator
\[ (TQ)(s, a, o) = \sum_{s_+} P(s, a, s_+) \max_{\pi \in \Pi} \min_{o \in O} (R(s, a, o) + E(Q(s_+, \pi(s_+), o))). \]

For convenience, let
\[ \alpha_k(s, a, o) = 1_{(s, a, o)}(s_k, a_k, o_k) \cdot \beta_N(k, s, a, o), \]
\[ M_{k+1}(s, a, o) = \max_{\pi \in \Pi} \min_{o \in O} (R(s, a, o) + E(Q_k(s_{k+1}, \pi(s_{k+1}), o))) - (TQ_k)(s, a, o), \]
and
\[ h(Q) = TQ - f(Q) \cdot e - Q. \]
Then,
\[ Q_{k+1} = Q_k + \alpha_k(h(Q_k) + M_{k+1}). \]

We will now show that conditions 1 - 4 in Proposition 3 hold, and therefore \( Q_k \to Q^* \) with probability 1, where \( Q^* \) is the solution to (17).
1. Let $\mathcal{F}_k$ be the sigma algebra $\sigma(Q_t, M_t, t \leq k)$. Clearly,

$$E(M_{k+1} \mid \mathcal{F}_k) = 0$$

and

$$E(\|M_{k+1}\|^2 \mid \mathcal{F}_k) \leq C_1(1 + \|Q_k\|^2)$$

for some constant $C_1 > 0$.

2. We have supposed that assumptions 4 and 5 hold.

3. Let $h(X) = T(X) - X - f(X) \cdot e$ and introduce

$$(TQ)(s, a, o) = \max_{a_+ \in A} \sum_{s_+} P(s, a, s_+)Q(s_+, a_+, o). \quad (44)$$

Then, the limit

$$h_\infty(X) = \lim_{z \to \infty} h(zX)/z$$
$$= T(X) - X - f(X) \cdot e$$

exists.

4. By noting that

$$h(x) = T(X) - X - f(X) \cdot e = T'(X) - X$$

we can apply Lemma 3 and conclude that $\dot{Q}(t) = h(Q(t))$ has a unique globally asymptotically stable equilibrium $Q^*$.

Thus, according to Proposition 3, the iterators $Q_k$ in (18) converge to $Q^*$, where $h(Q^*) = 0$ and hence the unique solution to (17). Thus, the policy $\pi^* \in \Pi$ given by

$$\pi^*(s) = \arg \max_{\pi} \min_{o \in O} Q^*(s, \pi(s), o)$$

maximizes (13), and the proof is complete.

\[\square\]

**E Proof of Theorem 3**

Let

$$L(\pi, j) = E \left( \sum_{k=0}^{\infty} \gamma^k r^j(s_k, \pi(s_k)) \right).$$

Consider the zero-sum game

$$\max_{\pi \in \Pi} \min_{j \in [J]} L(\pi, j).$$

Suppose that $\pi$ is a policy such that

$$E \left( \sum_{k=0}^{\infty} \gamma^k r^j(s_k, \pi(s_k)) \right) < 0$$

for some $j$. Then,

$$L(\pi, j) < 0$$

which implies

$$\min_{j \in [J]} L(\pi, j) < 0.$$ 

Thus, if

$$\max_{\pi \in \Pi} \min_{j \in [J]} L(\pi, j) \geq 0$$
then, there must exist a policy $\pi$ that satisfies
\[
E \left( \sum_{k=0}^{\infty} \gamma^k r^j(s_k, \pi(s_k)) \right) \geq 0
\] (45)
for all $j$, and we get
\[
\min_{j \in [J]} \mathcal{L}(\pi, j) \geq 0.
\]
On the other hand, suppose that
\[
\max_{\pi \in \Pi} \min_{j \in [J]} \mathcal{L}(\pi, j) < 0.
\]
Then, there doesn’t exist a policy $\pi$ such that
\[
E \left( \sum_{k=0}^{\infty} \gamma^k r^j(s_k, \pi(s_k)) \right) \geq 0
\]
for all $j$, because it would imply that
\[
\max_{\pi \in \Pi} \min_{j \in [J]} \mathcal{L}(\pi, j) \geq 0
\]
which is a contradiction, and the proof is complete.

**F  Proof of Theorem 5**

Let
\[
\mathcal{L}(\pi, j) = \lim_{T \to \infty} E \left( \frac{1}{T} \sum_{k=0}^{T-1} r^j(s_k, \pi(s_k)) \right)
\]
where the expectation is taken over $s_k$ and $\pi$. The rest of the proof is similar to the proof of Theorem 3.

**G  Proof of Theorem 6**

According to Theorem 5, (23) is equivalent to the zero-sum Markov-Bandit game (24), which is equivalent to the zero-sum Markov-Bandit game given by the tuple $(S, A, O, P, R)$ with the objective
\[
\max_{\pi \in \Pi} \min_{o \in O} \lim_{T \to \infty} E \left( \frac{1}{T} \sum_{k=0}^{T-1} R(s_k, \pi(s_k), o) \right).
\] (46)

Assumption 3 implies that $|R(s, a, o)| \leq 2c$ for all $(s, a, o) \in S \times A \times O$. Now let $Q^*$ be the solution to the maximin optimality equation (17). According to Theorem 2, $Q_k$ in the recursion given by (18) converges to $Q^*$ with probability 1 under Assumptions 2, 3, 4, and 5. By definition, the optimal policy $\pi^*$ maximizes the expected average reward of the zero-sum Markov-Bandit game (46). Hence,
\[
\pi^*(s) = \arg \max_{\pi \in \Pi} \min_{o \in O} E(Q^*(s, \pi(s), o))
\]
and the proof is complete.

**H  Simulations**

In this section we will consider two additional examples for discounted rewards.
H.1 Static Process Example 1

In this subsection, we consider an example with 1 state (denoted as 1), 2 actions (denoted as 1, 2), and two constraints. Let the reward function for the two constraints, \( r^j(s,a) \) be given as

\[
\begin{array}{c}
r^1(1,1) = 1 \\
r^1(1,2) = -1 \\
r^2(1,1) = -1 \\
r^2(1,2) = 1
\end{array}
\] (47)

The aim of this example is to find a feasible policy that satisfies the discounted constraints. We let \( \gamma = \frac{1}{2} \) in this example. Since there is only a single state, we will ignore the first variable of state in the following. We note that the only stationary policy that satisfies the constraints in this example is \( \pi(1) = \pi(2) = 0.5 \) due to the symmetry of the two constraints. We will now illustrate that the proposed algorithm will achieve a feasible policy that satisfies the constraints.

First, we define the reward function \( R(a,o) \) for Markov zero-sum Bandit Game, \( a,o \in \{1,2\} \) as

\[
R(1,1) = 1 \quad R(2,1) = -1 \quad R(1,2) = -1 \quad R(2,2) = 1
\] (48)

We let the initial value for the Q-function be 0 and assume that the action for \( k = 0 \) is 1. For the learning rate, we adopt \( \alpha_k = \frac{1}{k+1} \). We also label the policy in time-step \( i \) as \( \pi_i \). According to Theorem 4, we can use the update rule in Eq. (12) to obtain the feasible policy. For \( k = 0 \), we have

\[
(\pi_0, o_0) = \arg \max_{\pi \in \Pi} \min_{o \in \mathcal{O}} Q_0(\pi_0(s), o)
\] (49)

Since \( Q_0 = 0 \) for all \( (a,o) \in \mathcal{A} \times \mathcal{O} \) and then the objective is not dependent on \( \pi \), any arbitrarily policy can be used. Let us choose \( \pi \) as a half-ball policy such that \( \pi_1(1) = \pi_1(2) = 0.5 \) and assume \( a_1 = 2 \). Similarly, \( o_0 \) can be arbitrary and we assume \( o_0 = 1 \). We also let \( a_0 = 1 \). Using \( a_0 = 1, o_0 = 1, \pi_1(1) = \pi_1(2) = 0.5 \), the Q-table update is given as

\[
Q_1(1,1) = (1 - \alpha_0(1,1))Q_0(1,1) + \alpha_0(1,1)(R(1,1) + \gamma E(Q_0(\pi_0, 1)))
= R(1,1) = 1
\] (50)

At the end of \( k = 0 \), we get \( Q_1(1,1) = 1 \) and \( Q_1(1,2) = Q_1(2,1) = Q_1(2,2) = 0 \).

For \( k = 1 \), we have

\[
(\pi_1, o_1) = \arg \max_{\pi \in \Pi} \min_{o \in \mathcal{O}} Q_1(s, \pi(s), o)
\] (51)

Since \( Q_1(2,1) = Q_1(2,2) = 0 \), the maxmin problem will again have result 0 whatever the policy \( \pi_2 \) is. Thus, we still assume that \( \pi_2(1) = \pi_2(2) = 0.5 \) and next action \( a_2 = 1 \). However, it follows that \( o_1 = 2 \) because \( Q_1(1,1) = 1 \). Since \( a_1 = 2, o_1 = 2, \pi_2(1) = \pi_2(2) = 0.5 \), the Q-table update is

\[
Q_2(2,2) = (1 - \alpha_1(2,2))Q_1(2,2) + \alpha_1(2,2)(R(2,2) + \gamma E(Q_1(\pi_1, 2)))
= 0.5 * 0 + 0.5 * (1 + 0.5 * 0) = 0.5
\] (52)

At the end of \( k = 1 \), we get \( Q_2(1,1) = 1, Q_2(2,2) = 0.5 \) and \( Q_2(1,2) = Q_2(2,1) = 0 \).

For \( k = 2 \), we have

\[
(\pi_2, o_2) = \arg \max_{\pi \in \Pi} \min_{o \in \mathcal{O}} Q_2(s, \pi(s), o)
\] (53)

To solve this problem, it is equivalent to solve the following problem

\[
\arg \max_z \quad z \\
\text{s.t.} \quad z \leq Q_2(s, \pi(s), o) \quad \text{for} \quad o = 1, 2
\] (54)

Assume \( \pi_3(1) = p, \pi_3(2) = 1 - p \), this is equivalent to solve the equation that \( p * Q_2(1,1) + (1 - p) * Q_2(1,2) = p * Q_2(1,2) + (1 - p) * Q_2(2,2) \), which gives the result \( \pi_3(1) = \frac{1}{3} \) and \( \pi_3(2) = \frac{2}{3} \) and we assume the next action
\(a_3 = 2\). Due to the equality in the above equation, \(o_2\) can again can be arbitrary and we assume \(a_2 = 2\). Since \(a_2 = 1\), the Q-table update is

\[
Q_3(1, 2) = (1 - \alpha_2(1, 2))Q_2(1, 2) + \alpha_2(1, 2)(R(1, 2) + \gamma \mathbb{E}(Q_2(\pi_2, 2)))
\]

\[
= \frac{2}{3} \cdot 0 + \frac{1}{3} \cdot [-1 + 0.5 \cdot \left(\frac{1}{3} \cdot Q_2(1, 2) + \frac{2}{3} \cdot Q_2(2, 2)\right)] = -\frac{5}{18}
\]

(55)

At the end of \(k = 2\), we get \(Q_3(1, 1) = 1\), \(Q_3(2, 2) = 0.5\) and \(Q_3(1, 2) = -\frac{5}{18}\) and \(Q_3(2, 2) = 0\).

For \(k = 3\), we have

\[
(\pi_4, o_3) = \arg \max_{\pi \in \mathcal{H}} \min_{o \in \mathcal{O}} Q_3(s, \pi(s), o)
\]

(56)

We need to solve the problem in the Equation (54) to get the result of \(\pi_4\) and the result is \(\pi_4(1) = \frac{7}{16}\) and \(\pi_4(2) = \frac{9}{16}\) and \(o_3\) can be arbitrary, thus we assume that \(a_3 = 2\). Since \(a_3 = 2\), the Q-table update is given as

\[
Q_4(2, 1) = (1 - \alpha_3(2, 1))Q_3(2, 1) + \alpha_3(2, 1)(R(2, 1) + \gamma \mathbb{E}(Q_3(1, 1)))
\]

\[
= \frac{3}{4} \cdot 0 + \frac{1}{4} \cdot (-1 + 0.5 \cdot \left(\frac{7}{16} \cdot Q_3(1, 1) + \frac{9}{16} \cdot Q_3(2, 1)\right)) = -\frac{3}{8} - \frac{1}{4} = -\frac{25}{128}
\]

(57)

At the end of \(k = 3\), we get \(Q_4(1, 1) = 1\), \(Q_4(2, 2) = 0.5\) and \(Q_4(1, 2) = -\frac{5}{18}\) and \(Q_4(2, 1) = -\frac{25}{128}\).

Based on these steps, we can keep on computing the update for Q-table. However, the computation is hard to do manually, and involves random choice of actions based on policy \(\pi\). Thus, we simulate the performance of the algorithm and the Q-values \(Q_k(i, j)\) for iterations \(k\) are depicted in Fig. 3.

![Figure 3: Convergence result for Example 1](image-url)

We note that \(Q_k(1, 1)\) and \(Q_k(2, 2)\) converges to 1, while \(Q_k(1, 2)\) and \(Q_k(2, 1)\) converges to \(-1\). According to the optimal Bellman equation,

\[
Q^*(s, a, o) = R(s, a, o) + \gamma \cdot \mathbb{E}(Q^*(s_+, \pi^*(s_+), o))
\]

(58)

we know \(Q^*(1, 1) = 1 + 0.5 \cdot [0.5 \cdot Q^*(1, 1) + 0.5 \cdot Q^*(2, 1)]\), which means

\[
3Q^*(1, 1) = 4 + Q^*(2, 1)
\]

(59)

Similarly, we have

\[
3Q^*(2, 1) = -4 + Q^*(1, 1)
\]

(60)

Combining these two equations, we have \(Q^*(1, 1) = -Q^*(2, 1) = -1\). Similarly, \(Q^*(2, 2) = -Q^*(1, 2) = -1\).

Thus, we see that the algorithm successfully have the whole Q table converges to \(Q^*\), which shows the correctness of the theorem. Moreover,

\[
\pi^* = \arg \max_{\pi \in \mathcal{H}} \min_{o \in \mathcal{O}} Q^*(s, \pi(s), o)
\]

(61)
which gives $\pi^*(1|s) = \pi^*(2|s) = 0.5$ and we know this is the only feasible policy. Thus, we see that the Q-values of the proposed algorithm converges to that of the optimal policy and the policy converges to the only feasible policy in this example.

H.2 Static Process Example 2

We consider a static process (that is, the state is constant) and an agent that takes action from the action set $A = \{1, 2, 3\}$. There are three objectives given by the reward functions $r_1, r_2, \text{and } r_3$ defined as

$$r^j(a) = \begin{cases} \frac{1}{2} & \text{if } a = j \\ 0 & \text{otherwise} \end{cases}$$

Note that we have dropped the dependence of the reward functions $r_j$ on the state $s$ as the state $s$ is assumed to be constant. Let the discount factor be $\gamma = \frac{1}{2}$ and let

$$\alpha_0 = \alpha_1 = \alpha_2 = \alpha = \frac{1}{3}.$$ 

The agent would then be looking for a probability distribution over the set $A$, $\Pr(a)$ for $a \in A$, that simultaneously satisfies the objectives

$$E \left( \sum_{k=0}^{\infty} \gamma^k r^j(a_k) \right) \geq \frac{1}{3}, \quad j = 1, 2, 3.$$ 

![Figure 4: A plot of the maximum of $|p_1 - \hat{p}_1| + |p_2 - \hat{p}_2| + |p_3 - \hat{p}_3|$ over 1000 iterations, as a function of the number of time steps.](image)

Now suppose that the agent takes action $a_k = 1$ with probability $p_1$. Then we have that

$$E \left( \sum_{k=0}^{\infty} \gamma^k r^1(a_k) \right) = p_1.$$ 

Similarly, we find that if the agent takes the action $a_k = j$ with probability $p_j$, $j = 2, 3$, then

$$E \left( \sum_{k=0}^{\infty} \gamma^k r^j(a_k) \right) = p_j.$$ 

Without loss of generality, suppose that $p_1 \leq p_2 \leq p_3$. Now the equality $p_1 + p_2 + p_3 = 1$ together with the Arithmetic-Geometric Mean Inequality imply that

$$\frac{1}{3} = \frac{p_1 + p_2 + p_3}{3} \geq \sqrt[3]{p_1 p_2 p_3} \geq p_1.$$
with equality if and only if \( p_1 = p_2 = p_3 = \frac{1}{3} \). Thus, in order to satisfy all of the three objectives, the agent’s mixed strategy is unique and given by \( p_1 = p_2 = p_3 = \frac{1}{3} \).

We have run 1000 iterations of a simulation of the learning algorithm as given by Theorem 4 over 5000 time steps (with respect to the time index \( k \)). As the above calculations showed, the probability distribution of the optimal policy is given by \( p_1 = p_2 = p_3 = \frac{1}{3} \). Let \( \hat{p}_1, \hat{p}_2, \hat{p}_3 \) be the estimated probabilities based on the \( Q \)-learning algorithm given by Theorem 4. In Figure 4, we see a plot of the maximum of the total error

\[
|p_1 - \hat{p}_1| + |p_2 - \hat{p}_2| + |p_3 - \hat{p}_3|
\]

over all iterations, as a function of the number of time steps. We see that it converges after 1000 time steps and stays stable for the rest of the simulation.