Reinforcement Learning
Proofs
E. Le Pennec
Fall 2018

1 History Dependent or Markov Policies

**Proposition 1** (Equivalence of History Dependent and Markov Policies). Let \( \pi \) be a stochastic history dependent policy. For each state \( s_0 \in S \), there exists a stochastic policy \( \pi' \) such that \( V^{\pi'}(s_0) = V^\pi(s_0) \).

**Proof.** Let \( \pi'(a_t|s_t) = \mathbb{E}[\pi(a_t|H_t)|S_t = s_t, S_0 = s_0] \), we can prove by recursion that

\[
\mathbb{P}_{\pi'}(S_t = s_t, A_t = a_t|S_0 = s_0) = \mathbb{P}_{\pi}(S_t = s_t, A_t = a_t|S_0 = s_0).
\]

This holds by definition for \( t = 0 \). Now assume the property is true for \( t' \leq t - 1 \). By construction,

\[
\mathbb{P}_{\pi}(S_t = s_t|S_0 = s_0) = \sum_{s_{t-1}} \sum_{a_{t-1}} p(s_t|s_{t-1}, a_{t-1}) \mathbb{P}_{\pi}(S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}|S_0 = s_0)
\]

\[
= \sum_{s_{t-1}} \sum_{a_{t-1}} p(s_t|s_{t-1}, a_{t-1}) \mathbb{P}_{\pi'}(S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}|S_0 = s_0)
\]

\[
= \mathbb{P}_{\pi}(S_t = s_t|S_0 = s_0).
\]

Hence,

\[
\mathbb{P}_{\pi'}(S_t = s_t, A_t = a_t|S_0 = s_0) = \pi'(a_t|s_t) \mathbb{P}_{\pi'}(S_t = s_t|S_0 = s_0)
\]

\[
= \mathbb{P}_{\pi}(A_t = a_t|S_t = s_t, S_0 = s_0) \mathbb{P}_{\pi}(S_T = s_t|S_0 = s_0)
\]

\[
= \mathbb{P}_{\pi'}(S_t = s_t, A_T = a_t|S_0 = s_0).
\]

It suffices then to notice that the quality criterion of \( \pi \) and \( \pi' \) depends on \( \pi \) only through respectively \( \mathbb{E}_{\pi}[r(S_t, A_t)|S_0 = s_0] \) or \( \mathbb{E}_{\pi}[r(S_t, A_t)|S_0 = s_0] \) which are equals. \( \square \)
2 Discounted Reward

2.1 Evaluation of a policy

Definition 1 (Value Function).

\[ v_\pi(s) = \mathbb{E}_\pi \left[ \sum_{t=0}^{+\infty} \gamma^t R_{t+1} \big| S_0 = s \right] = \sum_{t=0}^{+\infty} \gamma^t \mathbb{E}_\pi \left[ R_{t+1} \big| S_0 = s \right] \]

Definition 2 (Bellman Operator).

\[ T_\pi v(s) = \mathbb{E}_\pi \left[ R \big| S \right] + \gamma \sum_{s'} P_\pi \left( S = s' \big| S_0 = s \right) v(s') \]

\[ T_\pi v = r_\pi + \gamma P_\pi v \]

Proposition 2 (Value Function Characterization). Let \( \pi \) be a stationary Markov policy, if \( 0 < \gamma < 1 \) then \( v_\pi \) is the only solution of \( v = T_\pi v \),

\[ v = r_\pi + \gamma P_\pi v, \]

and \( v_\pi = (\text{Id} - \gamma P_\pi)^{-1} r_\pi \).

Proof. By definition, if \( v \) is a solution of \( v = T_\pi v \) then \( (\text{Id} - \gamma P_\pi)v = r_\pi \). As \( P_\pi \) is a stochastic matrix, \( \| P_\pi \| \leq 1 \) and thus

\[ \sum_{k=0}^{\infty} \gamma^k P_\pi \]

is well defined. One verify easily that this is an inverse of \( I - \gamma P_\pi \) and such a \( v \) exists, is unique and equal to

\[ \sum_{k=0}^{\infty} \gamma^k P_\pi r_\pi. \]

Now,

\[ v_\pi(s) = \sum_{t=0}^{+\infty} \gamma^t \mathbb{E}_\pi \left[ R_{t+1} \big| S_0 = s \right] = \sum_{t=0}^{+\infty} \gamma^t \sum_{s'} \mathbb{P}_\pi \left( S_t = s' \big| S_0 = s \right) \mathbb{E}_\pi \left[ R \big| S = s' \right] \]

\[ = \sum_{t=0}^{+\infty} \gamma^t \sum_{s'} \left( P_\pi^t \right)_{s,s'} r_\pi(s') \]
\[
\sum_{t=0}^{+\infty} \gamma^t (P^t \pi) (s)
\]

and thus \( v = v_\pi \). \(\square\)

**Proposition 3** (Bellman Operator Property). The operator \( T_\pi \) satisfies the following contraction property

\[
\|T_\pi v - T_\pi v'\|_\infty \leq \gamma \|v - v'\|_\infty
\]

Furthermore, \( v \leq v' \) implies \( T_\pi v \leq T_\pi v' \) and \( T_\pi (v + \delta \mathbb{1}) = T_\pi v + \gamma \delta \mathbb{1} \)

**Proof.** For any \( s \),

\[
|T_\pi(v) - T_\pi(v')(s)| = |\gamma P_\pi(v - v')(s)| \\
\leq \gamma \|v - v'\|_\infty
\]

because \( P_\pi \) is a stochastic matrix.

It suffices to use the positivity of a stochastic matrix and the fact that \( \mathbb{1} \) is an eigenvector for the eigenvalue \( 1 \) to obtain the two remaining properties. \(\square\)

**Proposition 4** (Policy Prediction). For any \( v_0 \), define \( v_{n+1} = T_\pi v_n \) then

\[
\lim_{n \to \infty} v_n = v_\pi
\]

and

\[
\|v_n - v_\pi\|_\infty \leq \gamma^n \|v_0 - v_\pi\|_\infty
\]

Furthermore,

\[
\|v_n - v_\pi\|_\infty \leq \frac{\gamma}{1 - \gamma} \|v_n - v_{n-1}\|_\infty
\]

Finally, if \( v_0 \geq T_\pi v_0 \) (respectively \( v_0 \leq T_\pi v_0 \)) then \( v_0 \geq v_\pi \) (respectively \( v_0 \leq v_\pi \)) and \( v_n \) converges monotonously to \( v_\pi \).

**Proof.** For the first part of the proposition, we notice that \( v_\pi \) is the only fixed point of \( T_\pi \) which is a contraction. Hence, by the fixed point theorem, for any \( v_0 \), the sequence defined by \( v_{n+1} = T_\pi v_n \) converges toward \( v_\pi \).

A straightforward computation shows that

\[
\|v_n - v_\pi\|_\infty \leq \gamma \|v_{n-1} - v_\pi\|_\infty \leq \gamma^n \|v_0 - v_\pi\|_\infty.
\]

Along the same line,

\[
\|v_{n+k} - v_{n+k+1}\|_\infty \leq \gamma^{k+1} \|v_n - v_{n-1}\|_\infty.
\]

This implies that

\[
\|v_n - v_\pi\|_\infty \leq \sum_{i=0}^{k} \|v_{n+i} - v_{n+i+1}\|_\infty + \|v_{n+k+1} - v_\pi\|_\infty
\]
\[ \leq \gamma - \gamma^{k+2} \frac{1}{1 - \gamma} \|v_n - v_{n-1}\|_\infty + \gamma^{n+k+1} \|v_0 - v_\pi\|_\infty \]

which yields the result by taking the limit in \( k \).

\[ 2.2 \text{ Optimal Policy} \]

\[ 2.2.1 \text{ Characterization} \]

**Definition 3** (Optimal Reward).

\[ v^*(s) = \max_\pi v_\pi(s) \]

where the maximum can be taken indifferently in the set of history dependent policies or Markov policies.

**Definition 4** (Optimal Bellman Operator).

\[ T^* v(s) = \max_a E[R|S = s, A = a] + \gamma \sum_{s'} P(s'|S = s, A = a) v(s') \]

\[ = \max_a r(s, a) + \gamma \sum_{s'} p(s'|s, a) v(s') \]

\[ T^* v = \max_{\pi \in \mathcal{D}} r_\pi + \gamma P_\pi v \]

where \( S \) is the set of deterministic Markov policies and the max is component-wise.

**Proposition 5** (Optimal Bellman Operator and Markov Policies).

\[ T^*_\pi v(s) = \max_\pi T^*_\pi v(s) \]

or \( T^*_\pi v = \max_\pi \pi r_\pi + \gamma P_\pi v \)

**Proof.** \( \pi_\alpha = e_\alpha \) is such that \( T^*_\pi_\alpha (s) = E[R|s, a] + \gamma \sum_{s'} p(s'|s, a) v(s') \) so that \( \max_\pi T^*_\pi (s) \geq T^*_\pi s \).

Now, for any \( \pi \),

\[ T^*_\pi (s) = \sum_a \pi(a|s) \left( E[R|S = s, A = a] + \gamma \sum_{s'} p(s'|s, a) v(s') \right) \]

\[ \leq \max_a E[R|S = s, A = a] + \gamma \sum_{s'} p(s'|s, a) v(s') \]

\[ \leq T^*_s(s) \]

\[ \square \]
**Proposition 6** (Bellman Operator Property). The operator $\mathcal{T}_*$ satisfies the following contraction property

$$\|\mathcal{T}_* v - \mathcal{T}_* v'\|_\infty \leq \gamma \|v - v'\|_\infty$$

Furthermore, $v \leq v'$ implies $\mathcal{T}_* v \leq \mathcal{T}_* v'$ and $\mathcal{T}_*(v + \delta \mathbb{1}) = \mathcal{T}v + \gamma \delta \mathbb{1}$

Proof. For any $s$, if $\mathcal{T}_* v(s) \geq \mathcal{T}_* v'(s)$

$$|\mathcal{T}_* v - \mathcal{T}_* v'(s)| = \mathcal{T}_* v(s) - \mathcal{T}_* v'(s)$$

$$= \max_a r(s, a) + \gamma \sum_{s'} p(s'|s,a)v(s') - \left( \max_a r(s, a) + \gamma \sum_{s'} p(s'|s,a)v'(s') \right)$$

$$\leq \max_a \left( r(s, a) + \gamma \sum_{s'} p(s'|s,a)v(s') - \left( \max_a r(s, a) + \gamma \sum_{s'} p(s'|s,a)v'(s') \right) \right)$$

$$\leq \gamma \max_a \sum_{s'|s,a} p(s'|s,a)(v(s') - v'(s'))$$

$$\leq \gamma \|v - v'\|_\infty$$

Now, if $v \leq v'$, for any $a'$

$$r(s, a') + \gamma \sum_{s'} p(s'|s,a')v(s') \leq r(s, a') + \gamma \sum_{s'} p(s'|s,a')v'(s')$$

$$\leq \mathcal{T}_* v'(s)$$

hence $\mathcal{T}_* v \leq \mathcal{T}_* v'$.

Finally,

$$\mathcal{T}_*(v + \delta \mathbb{1})(s) = \max_a r(s, a) + \gamma \sum_{s'} p(s'|s,a)(v(s') + \delta)$$

$$= \max_a r(s, a) + \gamma \sum_{s'} p(s'|s,a)v(s') + \delta$$

$$= \mathcal{T}_*(v)(s) + \delta.$$

**Proposition 7** (Optimal Reward Characterization). $v_*$ is the unique solution of $V = \mathcal{T}_* V$.

Proof. Assume $v \geq \mathcal{T}_* v$ so that

$$v \geq \max_\pi r_\pi + \gamma P_\pi v.$$

Let $\pi = (\pi_0, \pi_1, \ldots)$ be a sequence of Markov policies,

$$v \geq r_{\pi_0} + \gamma P_{\pi_0} v$$

$$v \geq r_{\pi_0} + \gamma P_{\pi_0} (r_{\pi_1} + \gamma P_{\pi_1} v)$$
\[ v \geq \sum_{k=0}^{n} \gamma^k P^k \pi r_{\pi_k} + \gamma^{n+1} P^{n+1}_\pi v \]

where \( P^k_\pi = \prod_{k' < k} P_{\pi_{k'}} \). As \( v_\pi = \sum_{k=0}^{\infty} \gamma^k P^k_\pi r_{\pi_k} \), we verify that

\[ v - v_\pi \geq \gamma^{n+1} P^{n+1}_\pi v - \sum_{k=n+1}^{\infty} \gamma^k P^k_\pi r_{\pi_k}. \]

Taking the limit in \( k \) yields \( v \geq v_\pi \) and thus \( v \geq v_* \).

Now, if \( v \leq T_* v = \max_{\pi} r_\pi + \gamma P_\pi v \) then assuming the max is reached at \( \tilde{\pi} \)

\[ v \leq r_{\tilde{\pi}} + \gamma P_{\tilde{\pi}} v \leq \sum_{k=0}^{n} \gamma^k P^k_{\tilde{\pi}} r_{\tilde{\pi}} + \gamma^{n+1} P^{n+1}_{\tilde{\pi}} v \]

and thus \( v \leq v_{\tilde{\pi}} \leq v_* \).

We deduce thus that \( v = T_* v \) implies \( v = v_* \). It remains to prove that such a solution exists. This is a direct application of the fixed point theorem for the operator \( T_* \).

**Proposition 8.** Any policy \( \pi_* \) such that \( v_{\pi_*} = v_* \) is optimal.

**Proof.** This is a direct consequence of the previous theorem.

**Proposition 9.** Any stationary policy \( \pi_* \) verifying \( \pi_* \in \arg\max_{\pi} r_\pi + \gamma P_\pi v_* \) is optimal.

**Proof.** By definition,

\[ T_{\pi_*} v_* = r_{\pi_*} + P_{\pi_*} v_* = \max_{\pi} r_\pi + P_\pi v_* = T_* v_* = v_* \]

Hence \( v_{\pi_*} = v_* \) and the policy is optimal.

**2.2.2 Policy Improvement and Policy Iteration**

**Proposition 10** (One step look-head policy improvement). For any \( \pi, \pi_+ \) define by

\[ \pi_+ \in \arg\max_{\pi'} r_{\pi'} + \gamma P_{\pi'} v_\pi \]

satisfies

\[ v_{\pi_+} \leq v_\pi \]
Proof. By construction,
\[ r_{\pi_+} + \gamma P_{\pi_+} v_{\pi} \geq r_{\pi} + \gamma P_{\pi} v_{\pi} = v_{\pi} \]
and thus
\[ r_{\pi_+} - (I - \gamma P_{\pi_+}) v_{\pi} \geq 0. \]

It suffices to notice that \( v_{\pi_+} = (I - \gamma P_{\pi_+})^{-1} r_{\pi_+} \) so that
\[ v_{\pi_+} - v_{\pi} = (I - \gamma P_{\pi_+})^{-1} (r_{\pi_+} - (I - \gamma P_{\pi_+}) v_{\pi}) \geq 0 \]
where we have used the positivity of \( (I - \gamma P_{\pi_+})^{-1} = \sum \gamma^k P^k_{\pi_+} \).

\[ \square \]

**Proposition 11.** Let \( \Delta = B - \text{Id} \), the policy iteration scheme satisfies
\[ v_{n+1} = v_n + \sum_{k=0}^{\infty} \gamma^k P^k_{\pi_{n+1}} \Delta v_n. \]

**Proof.** As proved before,
\[ v_{n+1} = (\text{Id} - \gamma P_{\pi_{n+1}})^{-1} r_{\pi_{n+1}}. \]

Now by construction,
\[ Bv_n = T_{\pi_{n+1}} v_n = r_{\pi_{n+1}} + \gamma P_{\pi_{n+1}} v_n \]
and thus
\[ r_{\pi_{n+1}} = \Delta v_n + (\text{Id} - \gamma P_{\pi_{n+1}}) v_n. \]

This implies immediately
\[ v_{n+1} = v_n + (\text{Id} - \gamma P_{\pi_{n+1}})^{-1} \Delta v_n \]
\[ = v_n + \sum_{k=0}^{\infty} \gamma^k P^k_{\pi_{n+1}} \Delta v_n \]
\[ \square \]

2.2.3 Value Iteration

**Proposition 12.** For any \( v_0 \), define \( v_{n+1} = T_* v_n \) then
\[ \lim_{n \to \infty} v_n = v_* \]
and
\[ \|v_n - v_*\|_\infty \leq \gamma^n \|v_0 - v_*\|_\infty \]

Furthermore,
\[ \|v_n - v_*\|_\infty \leq \gamma \frac{1}{1 - \gamma} \|v_n - v_{n-1}\|_\infty \]

Finally, if \( v_0 \geq T_* v_0 \) (respectively \( v_0 \leq T_* v_0 \)) then \( v_0 \geq v_* \) (respectively \( v_0 \leq v_* \)) and \( v_n \) converges monotonously to \( v_* \).
Proof. For the first part of the proposition, we notice that $v_*$ is the only fixed point of $T$ which is a contraction. Hence, by the fixed point theorem, for any $v_0$, the sequence defined by $v_{n+1} = T v_n$ converges toward $v_*$.

A straightforward computation shows that
\[
\|v_n - v_*\|_\infty \leq \gamma \|v_{n-1} - v_*\|_\infty \leq \gamma^n \|v_0 - v_*\|_\infty.
\]

Along the same line,
\[
\|v_{n+k} - v_{n+k+1}\|_\infty \leq \gamma^{k+1} \|v_n - v_{n-1}\|_\infty.
\]

This implies that
\[
\|v_n - v_*\|_\infty \leq \gamma - \frac{\gamma^{k+2}}{1 - \gamma} \|v_n - v_{n-1}\|_\infty + \gamma \|v_{n+k+1} - v_*\|_\infty.
\]

which yields the result by taking the limit in $k$. 

Proposition 13. For any $v$ and any $\pi \in \text{argmax}_\pi T_\pi v$,
\[
\|v_\pi - v_*\|_\infty \leq \frac{2 \gamma}{1 - \gamma} \|v - v_*\|_\infty
\]

If $v = T v'$ then
\[
\|v_\pi - v_*\|_\infty \leq \frac{2 \gamma}{1 - \gamma} \|v - v'\|_\infty
\]

Proof. By definition of $\pi$, $T_\pi v = T v$, hence
\[
\|v_\pi - v_*\|_\infty \leq \|v_\pi - T_\pi v\|_\infty + \|T_\pi v - v_*\|_\infty
\]
\[
\leq \|T_\pi v_\pi - T_\pi v\|_\infty + \|T_\pi v - T_\pi v_*\|_\infty
\]
\[
\leq \gamma \|v_\pi - v\|_\infty + \gamma \|v - v_*\|_\infty
\]
\[
\leq \gamma \|v_\pi - v_*\|_\infty + 2 \gamma \|v - v_*\|_\infty
\]

and thus
\[
\|v_\pi - v_*\|_\infty \leq \frac{2 \gamma}{1 - \gamma} \|v - v_*\|_\infty
\]

For the second inequality,
\[
\|v_\pi - v_*\|_\infty \leq \|v_\pi - v\|_\infty + \|v - v_*\|_\infty
\]

Now
\[
\|v_\pi - v\|_\infty \leq \|T_\pi v_\pi - T_\pi v\|_\infty + \|T_\pi v - T_\pi v'\|_\infty
\]
≤ \gamma \|v_\pi - v\|_\infty + \gamma \|v - v'\|_\infty

and thus

\|v_\pi - v\|_\infty ≤ \frac{\gamma}{1 - \gamma} \|v - v'\|_\infty

Along the same line,

\|v - v_*\|_\infty ≤ \|v - \mathcal{T}_*v\|_\infty + \|\mathcal{T}_*v - v_*\|_\infty

≤ \|\mathcal{T}_*v' - \mathcal{T}_*v\|_\infty + \|\mathcal{T}_*v - \mathcal{T}_*v_*\|_\infty

≤ \gamma \|v - v'\|_\infty + \gamma \|v - v_*\|_\infty

and thus

\|v - v_*\|_\infty ≤ \frac{\gamma}{1 - \gamma} \|v - v'\|_\infty

Combining those two bounds yields the result.

\[\square\]

### 2.2.4 Modifier Policy Iteration

**Proposition 14** (MPI). Let \(v_0\) such that \(\mathcal{T}_*v_0 ≥ v_0\), define for any \(n\) and any \(m\)

- \(\pi_{n+1} \in \arg\max r_\pi + P_\pi v_n\)
- \(v_{n,0} = \mathcal{T}_*v_n = \mathcal{T}_{\pi_{n+1}}v_n\)
- \(v_{n,m} = \mathcal{T}_{\pi_{n+1}}v_{n,m-1}\)
- \(v_{n+1} = v_{m_{n}}\)

then \(v_{n+1} ≥ v_n\) and

\[\lim_{n→∞} v_n = v_*\]

At any step,

\[\|v_{\pi_{n+1}} - v_*\|_\infty ≤ \frac{2}{1 - \gamma} \|v_n - v_{n,0}\|_\infty\]

Furthermore,

\[\|v_{n+1} - v_*\|_\infty ≤ \left(\frac{\gamma - \gamma^{m_{n+1}}}{1 - \gamma} \|P_{\pi_{n+1}} - P_\pi\| + \gamma^{m_{n+1}}\right) \|v_n - v_*\|_\infty\]

**Proposition 15.** Let \(\Delta = \mathcal{T}_* - \text{Id}\), let \(W^{(m)}_\pi v = \mathcal{T}^{m+1}_\pi v\),

\[W^{(m)}_\pi v = \sum_{k=0}^{m} \gamma^k P^{k}_\pi r_\pi + \gamma^{m+1} P^{m+1}_\pi v\]

= \[v_n + \sum_{k=0}^{m} \gamma^k P^{k}_\pi \Delta v\]
Proof. By definition,

\[
W^{(m)}_\pi v = T^{m+1}_\pi v
= r_\pi + \gamma P_\pi T^m v
= \sum_{k=0}^{m} \gamma^k P_\pi r_\pi + \gamma^{m+1} P^m v
= \sum_{k=0}^{m} \gamma^k P_\pi (r_\pi + \gamma P_\pi v) + v
= v + \sum_{k=0}^{m} \gamma^k P_\pi \Delta v
\]

\[\square\]

Proposition 16. Define \(W^{(m)}_*\) by

\[
W^{(m)}_* v(s) = \max_{\pi} W^{(m)}_\pi v(s).
\]

then \(W^{(m)}_*\) is a contraction:

\[
\|W^{(m)}_* v - W^{(m)}_* v\|_\infty \leq \gamma^{m+1} \|v - v\|_\infty.
\]

Furthermore, \(W^{(m)}_* v_* = v_*\).

Proof. Assume without loss of generality that \(W^{(m)}_* v(s) - W^{(m)}_* v'(s) \geq 0\) and let \(\hat{\pi} \in \arg\max W^{(m)}_\pi v(s)\),

\[
W^{(m)}_* v(s) - W^{(m)}_* v'(s) = \max_{\pi} W^{(m)}_\pi v(s) - \max_{\pi} W^{(m)}_\pi v'(s)
\leq W^{(m)}_{\hat{\pi}} v(s) - W^{(m)}_{\hat{\pi}} v'(s)
\leq \gamma^{m+1} P^{m+1}_{\hat{\pi}} (v - v')(s)
\leq \gamma^{m+1} \|v - v\|_\infty
\]

By construction \(\Delta v_* = T_* v_* - v_* = 0\) and thus \(W^{(m)}_* v_* = v_*\). We deduce immediately that \(W^{(m)}_* v_* = \sup_\pi W^{(m)}_\pi v_* = v_*\)

\[\square\]

Proposition 17. If \(u \geq v\) then for any \(\pi\), \(W^{m}_\pi u \geq W^{m}_\pi v\)

If \(u \geq v\) and \(\Delta u \geq 0\) then for any \(\pi\) \(W^{m}_\pi u \geq T_* v\).

If \(\Delta u \geq 0\) and \(\pi_u\) such that \(T_* u = T_{\pi_u} u\) then \(W^{(m)}_{\pi_u} u \geq 0\)

Proof. By definition,

\[
W^{m}_\pi u - W^{m}_\pi v \geq W^{m}_\pi (u - v)
\geq \gamma^{m+1} P^{m+1}_\pi (u - v) \geq 0
\]
Now,

$$W^{(m)}_\pi u = u + \sum_{k=0}^{m} \gamma^k P^k \Delta u$$

$$\geq u + \Delta u = T_* u$$

$$\geq T_* v$$

By construction

$$\Delta W^{(m)}_\pi u = T_* W^{(m)}_\pi u - W^{(m)}_\pi u$$

$$\geq T_* W^{(m)}_\pi u - W^{(m)}_\pi u$$

$$\geq \Delta u - T_* u + u$$

$$\geq \Delta u + (\gamma P_\pi - \text{Id}) \left(W^{(m)}_\pi u - u\right) \geq \Delta u + (\gamma P_\pi - \text{Id}) \sum_{k=0}^{m} \gamma^k P^k \Delta u$$

$$\geq \gamma^m P^m \Delta u \geq 0$$

**Proof of MPI.** Let $u_0 = v_0 = w_0$.

By construction $T_{\pi_{n+1}} v_n = T_* v_n$ and one verify easily that $v_{n+1} = T^{m_{n+1}}_{\pi_{n+1}} v_n = W^{(m_{n+1})}_{\pi_{n+1}} v_n$.

Define now, $u_{n+1} = T_* u_n$ and $w_{n+1} = W^{(m_n)}_* w_n$. We can prove by recursion that $\Delta v_n \geq 0$, $v_{n+1} \geq v_n$ and $u_n \leq v_n \leq w_n$.

By assumption, $\Delta v_0 \geq 0$ so that $v_1 = W^{(m_1)}_* v_0 \geq T_* v_0 \geq v_0$.

Assume the property holds for $n-1$ then using the previous lemmas one obtains immediately $\Delta v_n \geq 0$ and

$$u_n = T_* u_{n-1} \leq v_n = W^{(m_{n-1})}_{\pi_{n-1}} v_{n-1} \leq w_n = W^{(m_{n-1})}_* w_{n-1}$$

Finally,

$$v_n = W^{(m_n-1)}_{\pi_{n}} v_{n-1}$$

$$= v_{n-1} + \sum_{k=0}^{m_{n-1}} \gamma^k P_{\pi_{n}} \Delta v_{n-1}$$

$$\geq v_{n-1}.$$

Now, we have already proved that $u_n = T_* u_0$ tends to $v_*$ with

$$\|u_n - v_*\|_\infty \leq \gamma^n \|v_0 - v_*\|_\infty$$

It suffices now to prove that $w_n$ also converges toward $v_*$ to obtain the convergence of $v_n$. We verify that

$$\|w_n - v_*\|_\infty = \|W^{(m_{n-1})}_* w_{n-1} - W^{(m_{n-1})}_* v_*\|_\infty$$
\[
\gamma^{m_n-1} \|w_{n-1} - v_*\|_\infty \\
\gamma \sum_{k=0}^{m_n-1} \|v_0 - v_*\|_\infty
\]

which implies the convergence of \(w_n\).

We have

\[
\|v_{\pi_{n+1}} - v_*\|_\infty \leq \|v_{\pi_{n+1}} - v_n\|_\infty + \|v_n - v_*\|_\infty
\]

Notice that \(v_{n,0} = T_{\pi_{n+1}} v_n = T_* v_n\) so that

\[
\|v_{\pi_{n+1}} - v_n\|_\infty \leq \|v_{\pi_{n+1}} - v_{n,0}\|_\infty + \|v_{n,0} - v_n\|_\infty
\]

\[
\leq \gamma \|v_{\pi_{n+1}} - v_n\|_\infty + \|v_{n,0} - v_n\|_\infty
\]

Along the same line,

\[
\|v_* - v_n\|_\infty \leq \|v_* - v_{n,0}\|_\infty + \|v_{n,0} - v_n\|_\infty
\]

\[
\leq \|T_* v_* - T_* v_n\|_\infty + \|v_{n,0} - v_n\|_\infty
\]

\[
\leq \gamma \|v_* - v_n\|_\infty + \|v_{n,0} - v_n\|_\infty
\]

Combining those two inequalities yields

\[
\|v_{\pi_{n+1}} - v_*\|_\infty \leq \frac{2}{1 - \gamma} \|v_n - v_{0,n}\|_\infty
\]

As shown before,

\[
0 \leq v_* - v_{n+1} \leq v_* - v_n - \sum_{k=0}^{m_n} \gamma^k P_{\pi_{n+1}}^k \Delta v_n
\]

Now, let \(\pi_*\) such that \(T_{\pi_*} v_* = B v_*\),

\[
\Delta_n = \Delta v_n - \Delta v_* = T_* v_n - v_n - (T_* v_* - v_*)
\]

\[
\leq T_{\pi_*} v_n - v_n - (T_{\pi_*} v_* - v_*)
\]

\[
\leq (\gamma P_{\pi_*} - \text{Id})(v_n - v_*)
\]

Thus

\[
0 \leq v_* - v_{n+1} \leq v_* - v_n - \sum_{k=0}^{m_n} \gamma^k P_{\pi_{n+1}}^k (\gamma P_{\pi_*} - \text{Id})(v_n - v_*)
\]

\[
\leq \sum_{k=1}^{m_n} \gamma^k P_{\pi_{n+1}}^k (v_n - v_*) - \sum_{k=0}^{m_n} \gamma^{k+1} P_{\pi_{n+1}} P_{\pi_*}(v_n - v_*)
\]

\[
\leq \sum_{k=0}^{m_n} \gamma^{k+1} P_{\pi_{n+1}} P_{\pi_*}(v_n - v_*) - \gamma^{m_n+1} P_{\pi_{n+1}} P_{\pi_*}(v_n - v_*)
\]

12
\[
\leq \sum_{k=0}^{m_n} \gamma^{k+1} \| P_{\pi_{n+1}} - P_{\pi_n} \| v_n - v_s \|_\infty + \gamma^{m_n+1} \| v_n - v_s \|_\infty \\
\leq \left( \frac{\gamma - \gamma^{m_n+1}}{1 - \gamma} \right) \| P_{\pi_{n+1}} - P_{\pi_n} \| + \gamma^{m_n+1} \| v_n - v_s \|_\infty
\]

2.3 Asynchronous Dynamic Programming

**Proposition 18.** Assume \( T_{\pi_n} v_0 \geq v_0 \) and at any step \( n \)

- Define a subset \( S_n \) of the states and
- Either
  - keep the policy \( \pi_{n+1} = \pi_n \) and update the value function following
    \[
    v_{n+1}(s) = \begin{cases} 
    T_{\pi_n} v_n(s) & \text{if } s \in S_n \\
    v_n(s) & \text{otherwise}
    \end{cases}
    \]
  - keep the value function \( s_{n+1} = s_n \) and update the policy following
    \[
    \pi_{n+1}(s) = \begin{cases} 
    \arg\max_a r(s,a) + \gamma P_{\pi_n} v_n(s) & \text{if } s \in S_n \\
    \pi_n(s) & \text{otherwise}
    \end{cases}
    \]

Assume that for any state \( s \) and any \( n \) there exist \( n' > n \) such that \( s \in S_{n'} \)
and one performs a value update at step \( n' \) and \( n'' > n \) such that \( s \in S_{n''} \) and
one performs a policy update at step \( n'' \) then \( s_n \) tends monotonously to \( s^* \).

**Proof.** We start by proving by recursion that \( T_{\pi_n} v_n \geq v_n \) implies

\[
T_{\pi_{n+1}} v_{n+1} \geq v_{n+1} \geq v_n \quad \text{and} \quad T_{\pi_n} v_n
\]

Note that that \( T_{\pi_0} v_0 \geq v_0 \) is an assumption.

Assume now that \( T_{\pi_n} v_n \geq v_n \), then either at step \( n \) we update the value
function or the policy.

If we update the value function, \( \pi_{n+1} = \pi_n \) and thus

\[
v_{n+1}(s) = \begin{cases} 
T_{\pi_n} v_n(s) & \text{if } s \in S_n \\
v_n(s) & \text{otherwise}
\end{cases}
\]

As \( T_{\pi_n} v_n(s) \geq v_n(s) \), we deduce \( T_{\pi_n} v_n \geq v_{n+1} \geq v_n \). It suffices to notice
that \( v_{n+1} \geq v_n \) implies

\[
T_{\pi_{n+1}} v_{n+1} = T_{\pi_n} v_{n+1} \geq T_{\pi_n} v_n
\]

to obtain

\[
T_{\pi_{n+1}} v_{n+1} \geq v_{n+1} \geq v_n.
\]
Now, if we update the policy, \( v_{n+1} = v_n \) and

\[
T_{\pi_{n+1}} v_n(s) = \begin{cases} T_s v_n(s) & \text{if } s \in S_n \\ T_{\pi_n} v_n(s) & \text{otherwise} \end{cases}
\]

which implies \( T_{\pi_{n+1}} v_n \geq T_{\pi_n} v_n \) and thus as \( v_{n+1} = v_n \)

\[
T_{\pi_{n+1}} v_{n+1} \geq T_{\pi_n} v_n \geq v_n = v_{n+1}.
\]

We deduce thus that

\[
T_s v_{n+1} \geq T_{\pi_{n+1}} v_{n+1} \geq v_{n+1} \geq v_n.
\]

which implies if we take the limit in \( k \)

\[
v_s \geq v_{n+1} \geq v_n.
\]

Hence \( v_n \) converges toward a limit \( \tilde{v} \) satisfying

\[
v_n \leq \tilde{v} \leq T_s \tilde{v} \leq v_s.
\]

Assume now that there exists \( s \) such that \( \tilde{v}(s) < T_s \tilde{v}(s) \). By continuity of \( T_s \), there exists \( n \) such that for all \( n' \geq n \)

\[
\tilde{v}(s) < T_s v_n'(s)
\]

Let \( n' \geq n \) such that one updates the policy of \( s \) and \( n'' \) the smallest integer larger than \( n'' \) where one updates the value of \( s \).

\[
v_{n''+1}(s) = T_{\pi_{n''}} v_{n''}(s)
\]

\[
\geq T_{\pi_{n'+1}} v_{n'+1}(s)
\]

\[
\geq T_{\pi_{n'+1}} v_n'(s)
\]

\[
\geq T_s v_n'(s) > \tilde{v}(s)
\]

which is impossible. \( \square \)

### 2.4 Approximate Dynamic Programming

**Proposition 19.** If in a Generalized Policy Improvement, for all \( k \)

\[
\| v_k - v_{\pi_k} \|_\infty \leq \epsilon
\]

and

\[
\| T_{\pi_{k+1}} v_k - T_s v_k \|_\infty \leq \delta
\]

then

\[
\limsup_n \max_s (v_s(s) - v_{\pi_k}(s)) \leq \frac{\delta + 2\gamma\epsilon}{(1 - \gamma)^2}
\]
Proof. By construction,

\[ v_{\pi_k}(s) - v_{\pi_{k+1}}(s) = T_{\pi_k} v_{\pi_k}(s) - T_{\pi_{k+1}} v_{\pi_{k+1}} \]
\[ = T_{\pi_k} v_{\pi_k}(s) - T_{\pi_k} v_k(s) + T_{\pi_k} v_k(s) - T_{\pi_{k+1}} v_{\pi_{k+1}} \]
\[ \leq \gamma \epsilon + T_s v_k(s) - T_{\pi_{k+1}} v_{\pi_{k+1}} + \delta \]
\[ \leq \gamma \epsilon + T_{\pi_{k+1}} v_k(s) + T_{\pi_{k+1}} v_k(s) - T_{\pi_{k+1}} v_{\pi_{k+1}} + \delta \]
\[ \leq 2 \gamma \epsilon + \delta + \gamma \max_{s'} (v_{\pi_k}(s') - v_{\pi_{k+1}}(s')) \]

and thus

\[ \max_{s'} (v_{\pi_k}(s') - v_{\pi_{k+1}}(s')) \leq \frac{2 \gamma \epsilon + \delta}{1 - \gamma} \]

Now,

\[ v_*(s) - v_{\pi_{k+1}}(s) = v_*(s) - T_{\pi_{k+1}} v_{\pi_{k+1}}(s) \]
\[ = v_*(s) - T_{\pi_{k+1}} v_{\pi_k}(s) + T_{\pi_{k+1}} v_{\pi_k}(s) - T_{\pi_{k+1}} v_{\pi_{k+1}}(s) \]
\[ \leq v_*(s) - T_{\pi_{k+1}} v_{\pi_k}(s) + \gamma \epsilon + \frac{2 \gamma \epsilon + \delta}{1 - \gamma} \]
\[ \leq v_*(s) - T_s v_k(s) + \gamma \epsilon + \frac{2 \gamma \epsilon + \delta}{1 - \gamma} \]
\[ \leq v_*(s) - T_s v_k(s) + \gamma \epsilon + \delta + \gamma \frac{2 \gamma \epsilon + \delta}{1 - \gamma} \]
\[ \leq T_* v_*(s) - T_* v_{\pi_k}(s) + 2 \gamma \epsilon + \delta + \gamma \frac{2 \gamma \epsilon + \delta}{1 - \gamma} \]
\[ \leq \max_s (v_*(s) - v_{\pi_k}(s)) + 2 \gamma \epsilon + \delta + \gamma \frac{2 \gamma \epsilon + \delta}{1 - \gamma} \]

thus

\[ \max_s (v_*(s) - v_{\pi_{k+1}}(s)) \leq \max_s \left( v_*(s) - v_{\pi_k}(s) \right) + 2 \gamma \epsilon + \delta + \gamma \frac{2 \gamma \epsilon + \delta}{1 - \gamma} \]

and

\[ \limsup_s (v_*(s) - v_{\pi_k}(s)) \leq \limsup \gamma \max_s (v_*(s) - v_{\pi_k}(s)) + 2 \gamma \epsilon + \delta + \gamma \frac{2 \gamma \epsilon + \delta}{1 - \gamma} \]

which implies

\[ \limsup_s (v_*(s) - v_{\pi_k}(s)) \leq \frac{2 \gamma \epsilon + \delta}{(1 - \gamma)^2} \]

\[
\square
\]
3 Finite Horizon

Proposition 20. If $v_0 = r_{\pi,T-1}$ and $v_n = T_{\pi,T-n}v_{n-1} = r_{\pi,T-n} + P_{\pi,T-n}v_{n-1}$ then

$$v_n(s) = E_{\pi} \left[ \sum_{t=T-n}^{T-1} R_{t+1} | S_{t-n-1} = s \right] = v_{\pi,T-n}(s)$$

If $v_0 = r_*$ and $v_{n+1} = T_* v_n$ then

$$v_n(s) = \max_{\pi} E_{\pi} \left[ \sum_{t=T-n}^{T-1} R_{t+1} | S_{t-n-1} = s \right] = v_{*T-n}(s)$$

Proof. If $n = 0$ then by definition $v_{\pi,T}(s) = E_{\pi} [R_T | S_{T-1} = s] = r_{\pi,T-1}(s)$.

Now,

$$v_{\pi,T-n}(s) = E_{\pi} \left[ \sum_{t=T-n}^{T-1} R_{t+1} | S_{T-n-1} = s \right]$$

$$= r_{\pi,T-n-1}(s) + E_{\pi} \sum_{t=T-n}^{T-1} R_{t+1} | S_{T-n-1} = s$$

$$= r_{\pi,T-n-1}(s) + \sum_a \sum_{s'} p(s'|s,a) \pi(a|s) E_{\pi} \left[ \sum_{t=T-n}^{T-1} R_{t+1} | S_{t-n} = s' \right]$$

$$= v_{\pi,T-n}(s) + P_{\pi,T-n-1}v_{\pi,T-n-1}(s)$$

Along the same line, if $n = 0$ then by definition $v_{*T}(s) = \max_{\pi} E_{\pi} [R_T | S_{T-1} = s] = \max_{\pi} v_{\pi,T}(s) = r_*(s)$.

Now,

$$v_{*,T-n}(s) = \max_{\pi} E_{\pi} \left[ \sum_{t=T-n}^{T-1} R_{t+1} | S_{T-n-1} = s \right]$$

$$= \max_{\pi} \left( r_*(s) + E \left[ \sum_{t=T-n}^{T-1} R_{t+1} | S_{T-n-1} = s \right] \right)$$

$$= \max_{\pi} \left( r_{*,T-n-1}(s) + \sum_a \sum_{s'} p(s'|s,a) \pi(a|s) E_{\pi} \left[ \sum_{t=T-n}^{T-1} R_{t+1} | S_{t-n} = s' \right] \right)$$

$$= \max_{\pi} r_{*,T-n-1}(s) + P_{*,T-n-1} \max_{\pi} v_{*,T-n-1}(s)$$

$$= T_* v_{*,T-n-1}(s)$$

\(\square\)
4 Non Discounted Total Reward

Definition 5. Let \( \bar{s} \) be the absorbing state, we define the expected absorption time starting from \( s \) \( \tau_\pi(s) \) by

\[
\tau_\pi(s) = E_\pi \left[ \inf_{S_t=\bar{s}} t \middle| S_0 = s \right].
\]

If \( \tau_\pi \) is finite, we say that \( \pi \) is proper.

Definition 6. We define the maximum expected absorption time starting from \( s \) by \( \tau_*(s) \) by

\[
\tau_*(s) = \max_\pi \tau_\pi(s).
\]

Proposition 21. If \( \tau_\pi < +\infty \) then

\[
\tau_\pi = 1 + P_\pi \tau_\pi = T_\pi \tau_\pi
\]

If \( \tau_* < +\infty \) then

\[
\tau_* = \max_\pi 1 + P_\pi \tau_* = T_\tau \tau_
\]

Proof. It suffices to notice that \( \tau_\pi(s) = E_\pi \left[ \sum_{t=0}^{+\infty} R_{t+1} \right] \) with \( R_t = 0 \) if \( s_t = \bar{s} \) and 1 otherwise.

Proposition 22. \( T_\pi \) is a contraction of factor \( \max \tau_\pi(s) - 1 / \tau_\pi(s) \) with respect to the norm \( \| \cdot \|_{\infty,1/\tau_\pi} \).

\( T_\pi \) and \( T_* \) are contraction of factor \( \max \tau_*(s) - 1 / \tau_*(s) \) with respect to the norm \( \| \cdot \|_{\infty,1/\tau_*} \).

Proof.

\[
|T_\pi v(s) - T_\pi v'(s)| \leq |P_\pi(v - v')(s)| \\
\leq P_\pi(\tau \times |v - v'|)(s) \\
\leq P_\pi(\tau \times |v - v'|)_{\infty,1/\tau} \\
\leq \tau(s) \frac{1 + P_\pi(\tau(s) - 1)}{\tau(s)} \|v - v'|_{\infty,1/\tau} \\
\leq \tau(s) \frac{1 + P_\pi(\tau(s) - 1)}{\tau(s)} \|v - v'|_{\infty,1/\tau}
\]

which yields the result for both \( \tau = \tau_\pi \) and \( \tau = \tau_* \).

Now, assume without loss of generality that \( T_* v(s) \geq T_* v'(s) \),

\[
|T_* v(s) - T_* v'(s)|
\]
\[
= \max_{\pi} T_{\pi} v(s) - \max_{\pi} T_{\pi} v'(s)
\]
\[
\leq \max_{\pi} (T_{\pi} v(s) - T_{\pi} v'(s))
\]
\[
\leq \tau(s) \frac{1 + P_{\tau}(s) - 1}{\tau(s)} ||v - v'||_{\infty, 1/\tau}
\]
which yields the result for \( \tau = \tau_* \).

\[\square\]

5 Bandits

5.1 Regret

Definition 7. A k-armed bandit is defined by a collection of k random variable \( R(a), a \in \{1, \ldots, k\} \).

The best arm is \( a_* \) is such that \( \mathbb{E}[R(a_*)] \geq \max_a \mathbb{E}[R(a)] \).

For any policy \( \pi \), the regret is defined by

\[
r_{T, \pi} = T \mathbb{E}[R(a_*)] - \mathbb{E} \left[ \sum_{t=1}^{T} R(A_t) \right]
\]

where \( A_t \) is the arm chosen at time \( t \) following the policy \( \pi \).

Proposition 23. Let \( T_i(a) = \sum_{s=1}^{t} 1_{A_s=i} \) and \( \Delta(a) = \mathbb{E}[R(a_*)] - \mathbb{E}[R(a)] \) then

\[
r_{n, \pi} = \sum_{a=1}^{k} \Delta(a) \mathbb{E}[T_i(a)]
\]

Proof. By definition,

\[
r_{T, \pi} = n \mathbb{E}[R(a_*)] - \mathbb{E} \left[ \sum_{t=1}^{T} R(A_t) \right]
\]
\[
= \mathbb{E} \left[ \sum_{t=1}^{T} (\mathbb{E}[R(a_*)] - R(A_t)) \right]
\]
\[
= \mathbb{E} \left[ \sum_{t=1}^{T} \sum_{a=1}^{k} 1_{A_t=i} \mathbb{E}[R(a_*)] - R(A_t) \right]
\]
\[
= \sum_{a=1}^{k} \mathbb{E} \left[ \sum_{t=1}^{T} 1_{A_t=i} \mathbb{E}[R(a_*)] - R(A_t) \right]
\]
\[
= \sum_{a=1}^{k} \mathbb{E} \left[ \sum_{t=1}^{T} 1_{A_t=i} \Delta(a) \right]
\]
\[
= \sum_{a=1}^{k} \mathbb{E}[T_i(a)] \Delta(a)
\]

\[\square\]
5.2 Concentration of subgaussian random variables

Definition 8. A random variable $X$ is said to be $\sigma$-subgaussian if
$$E[\exp \lambda X] \leq \exp(\lambda^2 \sigma^2/2)$$

Proposition 24. If $X$ is $\sigma$-subgaussian then for any $\epsilon > 0$
$$P(X \geq \epsilon) \leq \exp \left( \frac{-\epsilon^2}{2\sigma^2} \right)$$

Proof.
$$P(X \geq \epsilon) = P(\exp(\lambda X) \geq \exp(\lambda \epsilon)) \leq \frac{E[\exp(\lambda X)]}{\exp(\lambda \epsilon)} \leq \exp(\lambda^2 \sigma^2/2 - \lambda \epsilon) \leq \exp \left( \frac{-\epsilon^2}{2\sigma^2} \right)$$
where the last inequality is obtained by optimizing in $\lambda$.

Proposition 25. If $X$ is $\sigma$-subgaussian and $Y$ is $\sigma'$-subgaussian conditionally to $X$ then
- $E[X] = 0$ and $\text{Var}[X] \leq \sigma^2$
- $cX$ is $|c|\sigma$-subgaussian.
- $X + Y$ is $\sqrt{\sigma^2 + (\sigma')^2}$-subgaussian.

Proof.
$$E[\exp \lambda X] = \sum_k \frac{\lambda^k}{k!} E[X^k]$$
while
$$\exp(\lambda^2 \sigma^2/2) = \sum_k \frac{\lambda^{2k} \sigma^{2k}}{2^k k!}$$
By looking at the term in front of $\lambda^1$ and $\lambda^2$, we obtain
$$\lambda E[X] \leq 0 \quad \text{and} \quad \frac{\lambda^2}{2!} E[X^2] \leq \frac{\lambda^2 \sigma^2}{2 \times 1!}$$
which implies
$$E[X] = 0 \quad \text{and} \quad \text{Var}[X] \leq \sigma^2.$$
By definition,
\[ E[\exp(\lambda eX)] \leq \exp(\lambda^2 e^2 \sigma^2 / 2) \]
hence the \(|c|\sigma\)-subgaussianity of \(cX\).

Now,
\[
E[\exp(\lambda(X + Y))] \leq E[E[\exp(\lambda(X + Y))|X]] \\
\leq E[E[\exp(\lambda X) \exp(\lambda Y)|X]] \\
\leq E\exp(\lambda X) \exp(\lambda^2 (\sigma')^2 / 2) \\
\leq \exp(\lambda^2 (\sigma^2 + (\sigma')^2) / 2)
\]

\[\square\]

**Proposition 26.** If \(X_i - \mu\) are iid \(\sigma\)-subgaussian variable,
\[
\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} X_i \geq \mu + \epsilon\right) \leq \exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right) \quad \text{and} \quad \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} X_i \leq \mu - \epsilon\right) \leq \exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right)
\]

Proof. It suffices to notice that \(\frac{1}{n} \sum_{i=1}^{n} X_i - \mu\) and \(\mu - \frac{1}{n} \sum_{i=1}^{n} X_i\) are \(\sigma/\sqrt{n}\)-subgaussian. \(\square\)

### 5.3 Explore Then Commit strategy

**Definition 9.** The simple current mean estimate \(Q_t(a)\) is defined by
\[
Q_t(a) = \frac{1}{T_t(a)} \sum_{s=1}^{t} 1_{A_s = a} R_s(a)
\]

**Proposition 27.** Assume we play the arm successively during \(Km\) steps and then play the arm which maximize the current mean estimate \(Q_t(a)\) then if the \(R(a) - \mathbb{E}[R(a)]\) is 1-subgaussian
\[
\begin{align*}
\mathbb{E}[r_{T,\pi}] &\leq \min(m, T/K) \sum_{a=1}^{k} \Delta(a) + \max(T - mK, 0) \sum_{a=1}^{k} \Delta(a) \exp(-m\Delta(a)^2 / 4)
\end{align*}
\]

Furthermore,
\[
\mathbb{P}(a_T = a_*) \geq 1 - \sum_{a \neq a_*} \exp(-m\Delta(a)^2 / 4)
\]

Proof. We have
\[
r_{T,\pi} = \sum_{a=1}^{k} \Delta(a) \mathbb{E}[T_T(a)]
\]
we can thus focus on $\mathbb{E}[T_T(a)]$.

Now

$$
\mathbb{E}[T_T(a)] \leq \min(m, n/K) + \max(n - mK, 0) \mathbb{P}(a_{mK+1} = a)
$$

$$
\leq \min(m, n/K) + \max(n - mK, 0) \mathbb{P}(Q_t(a) \geq \max_{a' \neq a} Q_t(a'))
$$

$$
\leq \min(m, n/K) + \max(n - mK, 0) \mathbb{P}(a_{mK+1} = a)
$$

$$
\leq \min(m, n/K) + \max(n - mK, 0) \mathbb{P}(Q_m(a) \geq Q_n(a_*))
$$

$$
\leq \min(m, n/K) + \max(n - mK, 0) \mathbb{P}(Q_{mK+1}(a) - \mathbb{E}[R(a)] - (Q_{mK+1}(a_*) - \mathbb{E}[R(a_*)]) \geq \Delta(a))
$$

It suffices then to notice that $Q_{mK+1}(a) - \mathbb{E}[R(a)] - (Q_{mK+1}(a_*) - \mathbb{E}[R(a_*)])$ is $\sqrt{2/m}$-subgaussian to obtain

$$
\mathbb{E}[T_T(a)] \leq \min(m, n/K) + \max(n - mK, 0) \mathbb{P}(Q_{mK+1}(a) \geq Q_{mK+1}(a_*))
$$

$$
\leq \min(m, n/K) + \max(n - mK, 0) \exp(-m\Delta(a)^2/4)
$$

Now

$$
\mathbb{P}(a_T = a_*) = 1 - \sum_{a \neq a_*} \mathbb{P}(a_T = a)
$$

$$
\leq 1 - \sum_{a \neq a_*} \exp(-m\Delta(a)^2/4)
$$

5.4 $\epsilon$-greedy strategy

**Proposition 28.** Let $\pi$ be an $\epsilon_t$-greedy strategy,

$$
r_{T,\pi} \geq \sum_{t=1}^{T} \frac{\epsilon_t}{k} \sum_{a=1}^{k} \Delta(a)
$$

**Proof.** By definition of an $\epsilon$-greedy strategy

$$
\mathbb{E}[T_i(a)] \geq \sum_{t=1}^{T} \frac{\epsilon_t}{k}
$$

hence the first result. 

**Proposition 29.** Let $\pi$ be an $\epsilon_t$-greedy strategy,

$$
\mathbb{P}(A_T = a_*) \geq 1 - \epsilon_T - \sum_{a \neq a_*} \frac{4(\Delta(a)^2)}{\Delta(a)^2 + \Sigma_T} e^{-6k}\Delta(a)^2/(4k)
$$

with $\Sigma_T = \sum_{s=1}^{T} \epsilon_s$. 

21
Furthermore,

\[ P(a_\ast = \arg\max Q_{T,a}) \geq 1 - \sum \exp(-\Sigma_{T, (6k)}) - \sum_{a \neq a_\ast} \frac{4}{\Delta(a)^2} e^{-\Delta(a)^2 \Sigma_{T, (4k)}} \]

If \( \epsilon_t = c/t \),

\[ r_{T,\pi} \leq \sum_{a \neq a_\ast} \Delta(a) \left( \frac{c \log(T) + 1}{k} + C + \frac{4}{\Delta(a) C'} \right) \]

as soon as \( c/(6k) > 1 \) and \( c \min_{a \neq a_\ast} \Delta(a)/4k < 1 \).

If \( \epsilon_t = c \log(t)/t \) then

\[ r_{T,\pi} \leq \sum_{a \neq a_\ast} \Delta(a) \left( \frac{c \log(T) \log(T) + 1}{k} + C + \frac{4}{\Delta(a) C'} \right) \]

Proof. By definition of \( \pi \),

\[ P(A_T = a) \leq \frac{\epsilon_t}{k} + (1 - \frac{\epsilon_t}{k} P(Q_T(a) \geq Q_T(a_\ast)) \]

and

\[ P(Q_T(a) \geq Q_T(a_\ast)) \leq P(Q_T(a) \geq \mu(a) + \Delta(a)/2) + P(Q_T(a_\ast) \leq \mu(a_\ast) - \Delta(a)/2) \]

By symmetry, it suffices to bound

\[ P(Q_T(a) \geq \mu(a) + \Delta/2) \leq \sum_{t=1}^{T} P(T_t(a) = t, Q_T(a) \geq \mu(a) + \Delta/2) \]

\[ \leq \sum_{t=1}^{T} P(T_T(a) = t, \frac{1}{t} \sum_{k=1}^{t} R_k(a) \geq \mu(a) + \Delta/2) \]

\[ \leq \sum_{t=1}^{T} P(T_T(a) = t, \frac{1}{t} \sum_{k=1}^{t} R_k(a) \geq \mu(a) + \Delta/2) P \left( \frac{1}{t} \sum_{k=1}^{t} R_k(a) \geq \mu(a) + \Delta/2 \right) \]

\[ \leq \sum_{t=1}^{T} P(T_T(a) = t, \frac{1}{t} \sum_{k=1}^{t} R_k(a) \geq \mu(a) + \Delta/2) e^{-\Delta^2 t/2} \]

\[ \leq \sum_{t=1}^{T} P(T_T(a) = t, \frac{1}{t} \sum_{k=1}^{t} R_k(a) \geq \mu(a) + \Delta/2) + \sum_{t=T_0+1}^{T} e^{-\Delta^2 t/2} \]

Let \( T^R_T(a) \) be the number of time the arm \( a \) has been chosen at random before time \( T \)

\[ \leq \sum_{t=1}^{T_0} P \left( T^R_T(a) \leq t, \frac{1}{t} \sum_{k=1}^{t} R_k(a) \geq \mu(a) + \Delta/2 \right) + \frac{2}{\Delta^2} e^{-\Delta^2 T_0/2} \]

22
\[
\leq \sum_{t=1}^{T_0} \mathbb{P}(T^R_t(a) \leq t) + \frac{2}{\Delta^2} e^{-\Delta^2 T_0/2}
\]

Now the Bernstein inequality yields

\[
\mathbb{P}(T^R_t(a) \leq \mathbb{E}[T^R_t(a)] - \lambda) \leq \exp\left(-\frac{\lambda^2}{2 \text{Var}[T^R_t(a)] + \lambda/2}\right)
\]

with

\[
\mathbb{E}[T^R_t(a)] = \sum_{s=1}^{t} \frac{\epsilon_s}{k}
\]

\[
\text{Var}[T^R_t(a)] = \sum_{s=1}^{t} \frac{\epsilon_s (1 - \epsilon_s/k)}{k}
\]

Choosing \( T_0 = \frac{1}{2} \sum_{s=1}^{T} \epsilon_s \) leads

\[
\mathbb{P}(T^R_t(a) \leq T_0) = \mathbb{P}(T^R_t(a) \leq 2T_0 - T_0)
\]

\[
\leq \exp\left(-\frac{T_0^2/2}{\sigma^2 + T_0/2}\right)
\]

\[
\leq \exp\left(-\frac{T_0^2/2}{T_0 + T_0/2}\right)
\]

\[
\leq \exp(-T_0/3)
\]

which implies

\[
\mathbb{P}(Q_T(a) \geq \mu(a) + \Delta/2) \leq T_0 \exp(-T_0/3) + \frac{2}{\Delta^2} e^{-\Delta^2 T_0/2}
\]

and thus

\[
\mathbb{P}(a = \arg\max Q_T(a)) \leq 2(1 - \frac{\epsilon_T}{k}) \left(\Sigma_T/(2k) \exp(-\Sigma_T/(6k)) + \frac{2}{\Delta(a)^2} e^{-\Delta(a)^2\Sigma_T/(4k)}\right)
\]

\[
\leq \frac{\epsilon_T}{k} + \Sigma_T + \frac{4}{\Delta(a)^2} e^{-\Delta(a)^2\Sigma_T/(4k)}\]

with \( \Sigma_T = \sum_{s=1}^{T} \epsilon_s \) which goes to 0 as soon as \( \Sigma_T \) tends to \(+\infty\) We deduce then that

\[
\mathbb{P}(A_T = a) \leq \frac{\epsilon_T}{k} + \frac{\epsilon_T}{k} + \frac{4}{\Delta(a)^2} e^{-\Delta(a)^2\Sigma_T/(4k)}
\]

23
which goes to 0 if furthermore $\epsilon_t$ tends to 0.

Finally,

$$\mathbb{E} [T_t(a)] = \sum_{t=1}^{T} \mathbb{P} (A_t = a)$$

$$\leq \sum_{t=1}^{T} \left( \frac{\epsilon_t}{k} + \sum_{t=1}^{T} \frac{\epsilon_k}{k} \exp(-\Sigma_t/(6k) + \frac{4}{\Delta(a)^2} e^{-\Delta(a)^2 \Sigma_t/(4k)}) \right)$$

Hence

$$r_{T, \pi} \leq \sum_{a \neq a_*} \left( \Delta(a) \left( \frac{\Sigma_T}{k} + \sum_{t=1}^{T} \frac{\Sigma_t}{k} e^{-\Sigma_t/(6k)} \right) + \frac{4}{\Delta(a)} \sum_{t=1}^{T} e^{-\Delta(a)^2 \Sigma_t/(4k)} \right)$$

Assume that $\epsilon_t = c/t$ so that $\Sigma_t \leq c(\ln(t) + 1)$ then the previous inequality becomes

$$r_{T, \pi} \leq \sum_{a \neq a_*} \left( \Delta(a) \left( c \log(T) + 1 + \sum_{t=1}^{T} c \log(t) + 1 \right) e^{-c(\log(t) + 1)/(6k)} \right) + \frac{4}{\Delta(a)} \sum_{t=1}^{T} e^{-\Delta(a)^2 c(\log(t) + 1)/(4k)}$$

as soon as $c/(6k) > 1$ and $c \min_{a \neq a_*} \Delta(a)/4k < 1$.

If $\epsilon_t = c \log(t)/t$ then

$$r_{T, \pi} \leq \sum_{a \neq a_*} \left( \Delta(a) \left( c \log(T)(\log(T) + 1) + C \right) + \frac{4}{\Delta(a)} C' \right)$$

5.5 UCB strategy

**Proposition 30.** Assume we use a UCB strategy with a variance term $\sqrt{\frac{c \log t}{T_t(a)}}$ then

$$r_n(t) \leq C_c \sum_a \Delta(a) + \sum_a \frac{4c \ln t}{\Delta(a)}.$$

with $C_c < +\infty$ as soon as $c > 3/2$.

Furthermore

$$\mathbb{P} (A_t = a_*) \geq 1 - 2kt^{-2c+2}$$

as soon as $t \geq \max_a \frac{4c \ln t}{\Delta(a)^2}$. 

24
Proof. By construction,

\[ T_t(a) = \sum_{s=1}^{t} 1_{A_s = a} \]

\[ \leq \sum_{s=1}^{t} 1_{Q_s(a) + c_s a = \max Q_s(a') + c_s a'} \]

\[ \leq T_0(a) + \sum_{s=T_0+1}^{t} 1_{Q_s(a) + c_s a = \max Q_s(a') + c_s a', T_s(a) \geq T_0(a)} \]

\[ \leq T_0(a) + \sum_{s=T_0+1}^{t} 1_{Q_s(a) + c_s a \geq Q_s(a_*) + c_s a_*, T_s(a) \geq T_0(a)} \]

\[ \leq T_0(a) + \sum_{s=T_0+1}^{t} 1_{\max_{T_0(a) \leq s' \leq s} \frac{1}{s} \sum_{j=1}^{s'} R(a_{(j)}(j)} + \sqrt{\frac{\ln s}{s'}}} \]

\[ \leq T_0(a) + \sum_{s=T_0+1}^{t} \sum_{s'=1}^{s-1} \sum_{s''=T_0(a)}^{s-1} 1_{\mu(a_s) \leq \mu(a) + 2\sqrt{\frac{\ln s}{s'}}} + \frac{1}{s} \sum_{j=1}^{s''} R(a_{(j)}(j)} \geq \mu(a) + \sqrt{\frac{\ln s}{s'}}} \]

\[ \leq T_0(a) + \sum_{s=T_0+1}^{t} \sum_{s'=1}^{s-1} \sum_{s''=T_0(a)}^{s-1} 1_{\mu(a_s) \leq \mu(a) + 2\sqrt{\frac{\ln s}{s'}}} + \frac{1}{s} \sum_{j=1}^{s''} R(a_{(j)}(j)} \geq \mu(a) + \sqrt{\frac{\ln s}{s'}}} + 2e^{-2c \ln s} \]

\[ \mathbb{E} [T_t(a)] \leq T_0(a) + \sum_{s=T_0+1}^{t} \sum_{s'=1}^{s-1} \sum_{s''=T_0(a)}^{s-1} 1_{\Delta(a) \leq 2\sqrt{\frac{\ln s}{s'}}} + 2s^{-2c} \]

choosing \( T_0(a) = \frac{4c \ln t}{\Delta(a)^2} \)

\[ \leq \frac{4c \ln t}{\Delta(a)^2} + \sum_{s=T_0+1}^{t} 2s^{-2c+2} \]

\[ \leq \frac{4c \ln t}{\Delta(a)^2} + C_c \]

as soon as \( c > 3/2 \).

One deduce thus

\[ r_n(t) \leq C_c \sum_a \Delta(a) + \sum_a \frac{4c \ln t}{\Delta(a)}. \]

Note that we have shown

\[ \mathbb{P} (A_t = a) \leq 2t^{-2c}. \]
as soon as $t \geq \frac{4e \ln t}{\Delta^{(n)}(a)^2}$. Thus
\[ \mathbb{P}(A_t = a_*) \geq 1 - 2kt^{-2c+2} \]
as soon as $t \geq \max_{n} \frac{4e \ln t}{\Delta^{(n)}(a)^2}$. \qed

6 Stochastic Approximation

6.1 Convergence of a mean

**Proposition 31.** Assume $X_i$ are i.i.d. such that $\mathbb{E}[X_i|\mathcal{F}_{i-1}] = \mu$ and $\text{Var}[X_i|\mathcal{F}_{i-1}] \leq \sigma^2$, let
\[ M_n = M_{n-1} + \alpha_n(X_n - M_{n-1}) \]
with $1 \geq \alpha_i \geq 0$ then

- if $\sum_{i=1}^{n} \alpha_i \to +\infty$ and $\sum_{i=1}^{n} \alpha_i^2 < +\infty$, $M_n \to \mu$ in quadratic norm.
- $\alpha_i = \alpha$ then $\lim \sup \|M_n - \mu\|^2 \leq \alpha \sigma^2$

**Proof.** By definition,
\[ M_n = M_{n-1} + \alpha_n(X_n - M_{n-1}) = (1 - \alpha_n)M_{n-1} + \alpha_nX_n = \prod_{i=1}^{n}(1 - \alpha_i)M_0 + \sum_{k=1}^{n} \prod_{i=k+1}^{n}(1 - \alpha_i)\alpha_kX_k \]
thus
\[ \mathbb{E}[\|M_n - \mu\|^2] = \prod_{i=1}^{n}(1 - \alpha_i)\|M_0 - \mu\|^2 + \sum_{k=1}^{n} \prod_{i=k+1}^{n}(1 - \alpha_i)^2\alpha_k^2\sigma^2 \]
Thus it suffices to prove that
\[ \prod_{i=1}^{n}(1 - \alpha_i) \to 0 \quad \text{and} \quad \sum_{k=1}^{n} \prod_{i=k+1}^{n}(1 - \alpha_i)^2\alpha_k^2 \to 0 \]
For the first part, we use $(1 - x) \leq e^{-x}$ for $0 \leq x \leq 1$ to obtain
\[ \prod_{i=1}^{n}(1 - \alpha_i) \leq e^{-\sum_{i=1}^{n}\alpha_i} \]
which goes to 0 if $\sum_{i=1}^{n} \alpha_i \to +\infty$.
For the second one,
\[ \sum_{k=1}^{n} \prod_{i=k+1}^{n}(1 - \alpha_i)^2\alpha_k^2 \leq \sum_{k=1}^{n} \prod_{i=k+1}^{n}(1 - \alpha_i)^2\alpha_k^2 + \sum_{k=m+1}^{n} \prod_{i=k+1}^{n}(1 - \alpha_i)^2\alpha_k^2 \]
Choosing $m = n/2$ yields

$$
\mathbb{E} \left[ \| M_n - \mu \|^2 \right] \leq e^{-\sum_{i=1}^{n} \alpha_i} \| M_0 - \mu \|^2 + e^{-2 \sum_{k=m/2}^{n} \alpha_k} \sum_{k=1}^{n/2} \alpha_k^2 \sigma^2 + \max_{k \geq n/2} \alpha_k \sigma^2
$$

If we assume that $\sum_{k=1}^{n} \alpha_i \rightarrow +\infty$ and $\sum_{k=1}^{m} \alpha_k^2 < +\infty$ then all the term in the right hand side goes to 0.

If we assume $\alpha_k = \alpha$ then

$$
\mathbb{E} \left[ \| M_n - \mu \|^2 \right] \leq e^{-n \alpha} \| M_0 - \mu \|^2 + ne^{-n \alpha} \alpha^2 \sigma^2 + \alpha \sigma^2
$$

which is yields the result. \qed

6.2 Generic Stochastic Approximation

**Definition 10** (Generic Stochastic Algorithm). Let $H_t$ be a sequence of approximation of an operator $h$, let $\alpha_i(t)$ be a set of non negative sequences, for any initial value $X_0$, we define the following iterative scheme

$$
X_{t+1,i} = X_t,i + \alpha_i(t) H_t(X_t)_i.
$$

**Definition 11.** $h$ and $H_t$ are compatible if

$$
H_t(x) = h(x) + \epsilon_t(x) + \delta_t(x)
$$

with

$$
\mathbb{E} [\epsilon_t(x)|\mathcal{F}_t] = 0 \quad \text{and} \quad \text{Var} [\epsilon_t(x)|\mathcal{F}_t] \leq c_0 (1 + \|x\|^2)
$$

and with probability 1

$$
\|\delta_n(x)\|^2 \leq c_n (1 + \|x\|)^2
$$

with $c_n \rightarrow 0$ and either

- it exists a non negative $V \in C^1$ with L-Lipschitz gradient satisfying

$$
\langle \nabla V(x), h(x) \rangle \leq -c \|\nabla V(x)\|^2
$$

$$
\mathbb{E} [\| H_t(x) \|^2] \leq c'_0 (1 + \|\nabla V(x)\|^2),
$$
• or \( h \) is a contraction for the norm considered.

**Proposition 32** (Generic Stochastic Approximation). Assume that for any \( i \), we have almost surely

\[
\sum_{i=1}^{T} \alpha_i \rightarrow +\infty \quad \text{and} \quad \sum_{i=1}^{T} \alpha_i^2 < +\infty
\]

Then providing \( h \) and \( H_t \) are compatible,

\[
h(X_n) \rightarrow 0.
\]

**Proof.** See Neuro-Dynamic programming from Bertsekas and Tsitsiklis. \( \square \)

### 6.3 TD(\( \lambda \)) and linear approximation

**Proposition 33.** Provided there is a unique stationary distribution \( \mu \) on the states, that the basis function are linearly independent and

\[
\sum_{i=1}^{T} \alpha \rightarrow +\infty \quad \text{and} \quad \sum_{i=1}^{T} \alpha^2 < +\infty
\]

For any \( \lambda \in (0,1) \), the TD(\( \lambda \)) algorithm with linear approximation converges with probability one. The limit \( w_{*,\lambda} \) is the unique solution of

\[
\Pi_\mu T_\pi^{(\lambda)} X w_{*,\lambda} = X w_{*,\lambda}.
\]

Furthermore,

\[
\|X w_{*,\lambda} - v_\pi\|_{2,\mu} \leq \frac{1 - \lambda \gamma}{1 - \gamma} \|\Pi_\mu v_\pi - v_\pi\|_{2,\mu}
\]

**Proof.** See Tsitsiklis and Van Roy. \( \square \)

**Proof.** Assume \( A \) is invertible and let \( w_{TD} = A^{-1} b \)

\[
E[w_{t+1} - w_{TD} | w_t] = w_t + \alpha(b - Aw_t) - w_{TD} = (Id - \alpha A)(w_t - w_{TD})
\]

If we prove that \( A \) is positive definite then \( A \) will be invertible and the asymptotic algorithm will converge provided \( \alpha \) is small enough.

In the continuous task setting,

\[
A = \sum_{s} \mu(s) \sum_{a} \pi(a | s) \sum_{r,s'} p(r,s' | s,a) x(s)(x(s) - \gamma x(s'))
\]

\[
= \sum_{s} \mu(s) \sum_{a} \pi(a | s) \sum_{s'} p_\pi(s' | s) x(s)(x(s) - \gamma x(s'))
\]

28
\[
\begin{align*}
  &= \sum_s \mu(s)x(s) \left( x(s) - \gamma \sum_{s'} p_\pi(s'|s)x(s') \right)^t \\
  &= X^t D (\text{Id} - \gamma P_\pi) X
\end{align*}
\]

where \( D \) is a diagonal matrix having \( \mu(s) \) on the diagonal.

As \( P_\pi \) is a stochastic matrix, the row sums of \( D (\text{Id} - \gamma P_\pi) \) are non-negative.

Recall that \( \mu \) is such that \( \mu^t P_\pi = \mu^t \) and thus

\[
1^t D (\text{Id} - \gamma P_\pi) = \mu^t (\text{Id} - \gamma P_\pi) \\
= \mu^t - \gamma \mu^t P_\pi \\
= (1 - \gamma) \mu^t > 0
\]