Reinforcement Learning
Book of Proofs

E. Le Pennec

May 27, 2023
Contents

1 History Dependent or Markov Policies 5

2 Discounted Reward 7
   2.1 Evaluation of a policy ................................................. 7
   2.2 Optimal Policy .......................................................... 9
      2.2.1 Characterization .................................................. 9
      2.2.2 Policy Improvement and Policy Iteration ......................... 12
      2.2.3 Value Iteration ................................................... 13
      2.2.4 Modifier Policy Iteration ......................................... 15
   2.3 Asynchronous Dynamic Programming ................................. 19
   2.4 Approximate Dynamic Programming .................................. 21

3 Finite Horizon 23

4 Non Discounted Total Reward 25

5 Bandits 27
   5.1 Regret ................................................................. 27
   5.2 Concentration of subgaussian random variables .................... 28
   5.3 Explore Then Commit strategy ....................................... 29
   5.4 $\epsilon$-greedy strategy .............................................. 30
   5.5 UCB strategy .......................................................... 34

6 Stochastic Approximation 37
   6.1 Convergence of a mean ............................................... 37
   6.2 Generic Stochastic Approximation .................................. 38
   6.3 TD(\lambda) and linear approximation ................................ 41
1 History Dependent or Markov Policies

Proposition 1.1  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 Markov 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

$$P_{\pi'}(S_t = s_t, A_t = a_t|S_0 = s_0) = P_{\pi}(S_t = s_t, A_t = a_t|S_0 = s_0).$$

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

$$P_{\pi}(S_t = s_t|S_0 = s_0) = \sum_{s_{t-1}} \sum_{a_{t-1}} p(s_t|s_{t-1}, a_{t-1}) P_{\pi}(S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}|S_0 = s_0)$$

$$= \sum_{s_{t-1}} \sum_{a_{t-1}} p(s_t|s_{t-1}, a_{t-1}) P_{\pi'}(S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}|S_0 = s_0)$$

$$= P_{\pi'}(S_t = s_t|S_0 = s_0).$$

Hence,

$$P_{\pi'}(S_t = s_t, A_t = a_t|S_0 = s_0) = \pi'(a_t|s_t) P_{\pi'}(S_t = s_t|S_0 = s_0)$$

$$= \mathbb{E}[\pi'(A_t = a_t|H_t, S_t = s_t, S_0 = s_0)] P_{\pi'}(S_T = s_T|S_0 = s_0)$$

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

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. \qed
2 Discounted Reward

2.1 Evaluation of a policy

Definition 2.1.1 Value Function

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

Definition 2.1.2 Bellman Operator

\[
\mathcal{T}_\pi v(s) = \mathbb{E}_\pi[R|s] + \gamma \sum_{s'} \mathbb{P}_\pi(s'|s) v(s') \\
\mathcal{T}_\pi v = r_\pi + \gamma P_\pi v
\]

Proposition 2.1.3 Value Function Characterization

Let \( \pi \) be a stationary Markov policy, if \( 0 < \gamma < 1 \) then \( v_\pi \) is the only solution of \( v = \mathcal{T}_\pi v \),

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

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

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

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

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

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

Now,
\[
v_\pi(s) = \sum_{t=0}^{+\infty} \gamma^t E_\pi[R_{t+1}|S_0 = s]
\]
\[
= \sum_{t=0}^{+\infty} \gamma^t \sum_{s'} P_\pi(S_t = s'|S_0 = s) E_\pi[R|S = s']
\]
\[
= \sum_{t=0}^{+\infty} \gamma^t \sum_{s'} (P^t_\pi)_{s,s'} r_\pi(s')
\]
\[
= \sum_{t=0}^{+\infty} \gamma^t (P^t_\pi r_\pi)(s)
\]
and thus \(v = v_\pi\).

\[\Box\]

**Proposition 2.1.4 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 a eigenvector for the eigenvalue 1 to obtain the two remaining properties. \[\Box\]

**Proposition 2.1.5 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 \(v_\pi\).
2.2 Optimal Policy

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

$$\|v_n - v_\pi\|_\infty \leq \gamma^n \|v_0 - v_\pi\|_\infty.$$ 

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

$$\|v_n - v_\pi\|_\infty \leq \sum_{i=0}^{k} \|v_{n+i} - v_{n+i+1}\|_\infty + \|v_{n+k+1} - v_\pi\|_\infty$$

$$\leq \frac{\gamma - \gamma^{k+2}}{1 - \gamma} \|v_n - v_{n-1}\|_\infty + \gamma^{n+k+1} \|v_0 - v_\pi\|_\infty$$

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

Finally, note that as

$$v_{n+2} = r_\pi + \gamma P_\pi v_{n+1}$$

and $P_\pi$ is made of non negative elements, $v_{n+1} \leq v_n$ implies

$$v_{n+2} = r_\pi + \gamma P_\pi v_n = v_{n+1}.$$ 

Thus $v_1 = T_\pi v_0 \leq v_0$ implies that $v_n$ is a decreasing sequence whose limit is $v_\pi$, yielding the result. The increasing case is obtained with a similar proof.

2.2 Optimal Policy

2.2.1 Characterization

**Definition 2.2.1 Optimal Reward**

$$v_\star(s) = \max_\pi v_\pi(s)$$

where the maximum can be taken indifferently in the set of history dependent policies or Markov policies.
Definition 2.2.2  
**Optimal Bellman Operator**

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

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

**Proposition 2.2.3  
Optimal Bellman Operator and Markov Policies**

\[ T_v(s) = \max_{\pi \in S} T^\pi_v(s) \]

or \( T_v = \max_{\pi \in S} r^\pi + \gamma P^\pi_v \) where \( S \) is the set of deterministic Markov policies and the max is componentwise.

**Proof.** \( \pi_a = e_a \) is such that \( T^\pi_v(s) = \mathbb{E}[R|S = s, A = a] + \gamma \sum_{s'} p(s'|s, a)v(s') \) so that \( \max_{\pi} T^\pi_v(s) \geq T_v(s) \).

Now, for any \( \pi \),

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

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

\[ \leq T_v(s) \]

\[ \square \]

**Proposition 2.2.4  
Bellman Operator Property**

The operator \( T_v \) satisfies the following contraction property

\[ \| T_v - T_v' \|_\infty \leq \gamma \| v - v' \|_\infty \]

Furthermore, \( v \leq v' \) implies \( T_v \leq T_v' \) and \( T_v(v + \delta 1) = T_v + \gamma \delta 1 \).
2.2 Optimal Policy

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

\[
|\mathcal{T}_s v - \mathcal{T}_s v'(s)| = \mathcal{T}_s v(s) - \mathcal{T}_s v'(s) \\
= \max_a r(s, a) + \gamma \sum_{s'} p(s'|s, a) v(s') - \left( \max_a r(s, a) + \gamma \sum_{s'} p(s'|s, a) v'(s') \right) \\
\leq \max_a \left( r(s, a) + \gamma \sum_{s'} p(s'|s, a) v(s') - \left( r(s, a) + \gamma \sum_{s'} p(s'|s, a) v'(s') \right) \right) \\
\leq \gamma \max_a \sum_{s'|s, a} p(s'|s, a)(v(s') - v'(s')) \\
\leq \gamma \|v - v'\|_{\infty}
\]

Now, if \( v \leq v' \), for any \( a' \)

\[
r(s, a') + \gamma \sum_{s'} p(s'|s, a') v(s') \leq r(s, a') + \gamma \sum_{s'} p(s'|s, a') v'(s') \\
\leq \mathcal{T}_s v'(s)
\]

hence \( \mathcal{T}_s v \leq \mathcal{T}_s v' \).

Finally,

\[
\mathcal{T}_s (v + \delta \mathbb{1})(s) = \max_a r(s, a) + \gamma \sum_{s'} p(s'|s, a) (v(s') + \delta) \\
= \max_a r(s, a) + \gamma \sum_{s'} p(s'|s, a) v(s') + \delta \\
= \mathcal{T}_s (v)(s) + \delta.
\]

**Proposition 2.2.5**

\( v_* \) is the unique solution of \( V = \mathcal{T}_s V \).

**Proof.** Assume \( v \geq \mathcal{T}_s v \) 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,

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

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

\[
v - v_{\pi} \geq \gamma^{n+1} P_{\pi_{n+1}} v - \sum_{k=n+1}^{\infty} \gamma^k P_{\pi_{r_{\pi_k}}}.
\]
2 Discounted Reward

Taking the limit in $n$ yields $v \geq v_\pi$ and thus $v \geq v_\ast$.

Now, if $v \leq \mathcal{T}_* v = \max_\pi r_\pi + \gamma P_\pi v$ then assuming the max is reached at $\tilde{\pi}$

$$v \leq r_{\tilde{\pi}} + \gamma P_{\tilde{\pi}} v \leq \sum_{k=0}^{n} \gamma^k P_t r_{\tilde{\pi}} + \gamma^{n+1} P_{\tilde{\pi}}^{n+1} v$$

and thus $v \leq v_\tilde{\pi} \leq v_\ast$.

We deduce thus that $v = \mathcal{T}_* v$ implies $v = v_\ast$. It remains to prove that such a solution exists. This is a direct application of the fixed point theorem for the operator $\mathcal{T}_*$.

\[ \square \]

Proposition 2.2.6

Any policy $\pi_*$ such that $v_{\pi_*} = v_\ast$ is optimal.

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

\[ \square \]

Proposition 2.2.7

Any stationary policy $\pi_*$ verifying $\pi_* \in \arg\max_\pi r_\pi + \gamma P_\pi v_\ast$ is optimal.

Proof. By definition,

$$\mathcal{T}_{\pi_*} v_\ast = r_{\pi_*} + P_{\pi_*} v_\ast = \max_\pi r_\pi + P_\pi v_\ast = \mathcal{T}_* v_\ast = v_\ast.$$ 

Hence $v_{\pi_*} = v_\ast$ and the policy is optimal.

\[ \square \]

2.2.2 Policy Improvement and Policy Iteration

Proposition 2.2.8

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

$$v_{\pi_+} \geq v_\pi$$

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 \((I - \gamma P_{\pi})^{-1} = \sum \gamma^k P_{\pi}^k\).

Proposition 2.2.9

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

Proof. As proved before,
\[
v_{n+1} = (\text{Id} - \gamma P_{\pi_{n+1}})^{-1} r_{\pi_{n+1}}.
\]
Now by construction,
\[
T_s v_n = T_{\pi_{n+1}} v_n = r_{\pi_{n+1}} + \gamma P_{\pi_{n+1}} v_n
\]
and thus
\[
r_{\pi_{n+1}} = \Delta v_n + (\text{Id} - \gamma P_{\pi_{n+1}}) v_n.
\]
This implies immediately
\[
v_{n+1} = v_n + (\text{Id} - \gamma P_{\pi_{n+1}})^{-1} \Delta v_n
\]
\[
= v_n + \sum_{k=0}^{\infty} \gamma^k P_{\pi_{n+1}} \Delta v_n
\]

Proposition 2.2.10

For any \(v_0\), define \(v_{n+1} = T_s v_n\) then
\[
\lim_{n \to \infty} v_n = v_s
\]
and
\[
\|v_n - v_s\| \leq \gamma^n \|v_0 - v_s\| \leq \gamma^n \|v_0 - v_s\| \infty
\]
Furthermore,
\[
\|v_n - v_s\| \leq \frac{\gamma}{1 - \gamma} \|v_n - v_{n-1}\| \infty
\]
Finally, if \(v_0 \geq T_s v_0\) (respectively \(v_0 \leq T_s v_0\)) then \(v_0 \geq v_s\) (respectively \(v_0 \leq v_s\)) and \(v_n\) converges monotonously to \(v_s\).
2 Discounted Reward

**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

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

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

**Proposition 2.2.11**

For any \( v \) and any \( \pi \in \arg\max \pi \ T^*_\pi v \),

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

If \( v = T^*_s v' \) then

\[
\|v_\pi - v^*\|_\infty \leq \frac{2\gamma}{1 - \gamma} \|v' - v^*\|_\infty.
\]

**Proof.** By definition of \( \pi \), \( T^*_\pi v = T^*_s v \), hence

\[
\|v_\pi - v^*\|_\infty \leq \|v_\pi - T^*_\pi v\|_\infty + \|T^*_s v - v^*\|_\infty
\leq \|T^*_\pi v - T^*_\pi v\|_\infty + \|T^*_s v - T^*_s v^*\|_\infty
\leq \gamma \|v_\pi - v\|_\infty + \gamma \|v - v^*\|_\infty
\leq \gamma \|v_\pi - v^*\|_\infty + 2\gamma \|v - v^*\|_\infty
\]

and thus

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

For the second inequality,

\[
\|v_\pi - v^*\|_\infty \leq \|v_\pi - v\|_\infty + \|v - v^*\|_\infty.
\]
2.2 Optimal Policy

Now
\[ \| v_\pi - v \|_\infty \leq \| T_\pi v_\pi - T_\pi v \|_\infty + \| T_\pi v - T_* v \|_\infty \]
\[ \leq \gamma \| v_\pi - v \|_\infty + \gamma \| v - v' \|_\infty \]
and thus
\[ \| v_\pi - v \|_\infty \leq \frac{\gamma}{1 - \gamma} \| v - v' \|_\infty \]

Along the same line,
\[ \| v - v_* \|_\infty \leq \| v - T_* v \|_\infty + \| T_* v - v_* \|_\infty \]
\[ \leq \| T_* v' - T_* v \|_\infty + \| T_* v - T_* v_* \|_\infty \]
\[ \leq \gamma \| v - v' \|_\infty + \gamma \| v - v_* \|_\infty \]
and thus
\[ \| v - v_* \|_\infty \leq \frac{\gamma}{1 - \gamma} \| v - v' \|_\infty \]

. Combining those two bounds yields the result. □

2.2.4 Modifier Policy Iteration

**Proposition 2.2.12** MPI

Let \( v_0 \) such that \( T_* v_0 \geq v_0 \), define for any \( n \) and any \( m_n \)

- \( \pi_{n+1} \in \arg\max r_\pi + P_\pi v_n \)
- \( v_{n,0} = T_* v_n = T_{\pi_{n+1}} v_n \)
- \( v_{n,m} = T_{\pi_{n+1}} v_{n,m-1} \)
- \( v_{n+1} = v_{m_n} \)

then \( v_{n+1} \geq v_n \) and

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

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

Furthermore,
\[ \| v_{n+1} - v_* \|_\infty \leq \left( \frac{\gamma - \gamma^{m_n+1}}{1 - \gamma} \right) \| P_{\pi_{n+1}} - P_\pi \| + \gamma^{m_n+1} \| v_n - v_* \|_\infty \]
Proposition 2.2.13
Let $\Delta = T_* - \text{Id}$, let $W^{(m)}_\pi v = T^m_\pi v$,
\[ W^{(m)}_\pi v = \sum_{k=0}^{m} \gamma^k P^k_\pi r_\pi + \gamma^{m+1} P^{m+1}_\pi v \]
\[ = v_n + \sum_{k=0}^{m} \gamma^k P^k_\pi \Delta v \]

Proof. By definition,
\[ W^{(m)}_\pi v = T^m_\pi v \]
\[ = r_\pi + \gamma P_\pi T^m_\pi v \]
\[ = \sum_{k=0}^{m} \gamma^k P^k_\pi r_\pi + \gamma^{m+1} P^{m+1}_\pi v \]
\[ = \sum_{k=0}^{m} \gamma^k P^k_\pi (r_\pi + \gamma P_\pi v - v) + v \]
\[ = v + \sum_{k=0}^{m} \gamma^k P^k_\pi \Delta v \]

Proposition 2.2.14
Define $W^{(m_n)}_* v(s)$ by
\[ W^{(m_n)}_* v(s) = \max_\pi W^{(m_n)}_\pi v(s). \]
then $W^{(m_n)}_*$ is a contraction:
\[ ||W^{(m_n)}_* v - W^{(m_n)}_* v'||_\infty \leq \gamma^{m_n+1} ||v - v'||_\infty. \]
Furthermore, $W^{(m_n)}_* v_* = v_*$. 

Proof. Assume without loss of generality that $W^{(m_n)}_* v(s) - W^{(m_n)}_* v'(s) \geq 0$ and let $\tilde{\pi} \in \text{argmax} W^{(m_n)}_\pi v(s)$,
\[ W^{(m_n)}_* v(s) - W^{(m_n)}_* v'(s) = \max_\pi W^{(m_n)}_\pi v(s) - \max_\pi W^{(m_n)}_\pi v'(s) \]
\[ \leq W^{(m_n)}_{\tilde{\pi}} v(s) - W^{(m_n)}_{\tilde{\pi}} v'(s) \]
\[ \leq \gamma^{m_n+1} P^{m_n+1}_\tilde{\pi} (v - v')(s) \]
\[ \leq \gamma^{m_n+1} ||v - v'||_\infty \]
2.2 Optimal Policy

By construction $\Delta v_s = T_s v_s - v_s = 0$ and thus $W^{(m_n)}_\pi v_s = v_s$. We deduce immediately that $W^{(m_n)}_s v_s = \sup_\pi W^{(m_n)}_\pi v_s = v_s$.

**Proposition 2.2.15**

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 $\pi$ $W^{(m)}_\pi u \geq T^{(m)}_\pi v$.

If $\Delta u \geq 0$ and $\pi_u$ such that $T^{(m)}_\pi u = T^{(m)}_\pi u$ 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,

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

By construction

$$\Delta W^{(m)}_{\pi_u} u = T^{(m)}_\pi W^{(m)}_{\pi_u} u - W^{(m)}_{\pi_u} u \geq \Delta u - T^{(m)}_{\pi_u} u + u \geq \Delta u + (\gamma P_{\pi_u} - \text{Id}) \left( W^{(m)}_{\pi_u} u - u \right) \geq \Delta u + (\gamma P_{\pi_u} - \text{Id}) \sum_{k=0}^m \gamma^k P^k \pi_u \Delta u \geq \gamma^m P^{m+1} \pi_u \Delta u \geq 0$$

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

By construction $T^{(m+1)}_{\pi} v_n = T_s v_n$ and one verify easily that $v_{n+1} = T^{(m+1)}_{\pi} v_n = W^{(m_n)}_{\pi_{n+1}} v_n$.

Define now, $u_{n+1} = T_u u_n$ and $w_{n+1} = W^{(m_n)}_s w_n$. 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 $v_1 = W^{(m_1)}_\pi v_0 \geq T_s v_0 \geq v_0$.

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

$$u_n = T_u u_{n-1} \leq v_n = W^{(m_n)}_{\pi_{n-1}} v_{n-1} \leq w_n = W^{(m_{n-1})}_s w_{n-1}$$
2 Discounted Reward

Finally,
\[
v_n = W_{\pi_n}^{(m_n-1)} v_{n-1}
\]
\[
= v_{n-1} + \sum_{k=0}^{m_n-1} \gamma^k P_{\pi_n} \Delta v_{n-1}
\]
\[
\geq v_{n-1}.
\]

Now, we have already proved that \( u_n = T_s u_0 \) tends to \( v_s \) with
\[
\|u_n - v_s\|_\infty \leq \gamma^n \|v_0 - v_s\|_\infty
\]

It suffices now to prove that \( w_n \) also converges toward \( v_s \) to obtain the convergence of \( v_n \). We verify that
\[
\|w_n - v_s\|_\infty = \|W_{\pi_n}^{(m_n-1)} w_{n-1} - W_{\pi_n}^{(m_n-1)} v_s\|_\infty
\]
\[
\gamma^{m_n-1} \|w_{n-1} - v_s\|_\infty
\]
\[
\gamma \sum_{k=0}^{n-1} m_k \|v_0 - v_s\|_\infty
\]

which implies the convergence of \( w_n \).

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

Notice that \( v_{n,0} = T_{\pi_{n+1}} v_n = T_{\pi_n} v_n \) so that
\[
\|v_{\pi_{n+1}} - v_n\|_\infty \leq \|v_{\pi_{n+1}} - v_{n,0}\|_\infty + \|v_{n,0} - v_n\|_\infty
\]
\[
\leq \|T_{\pi_{n+1}} v_{\pi_{n+1}} - T_{\pi_{n+1}} v_n\|_\infty + \|v_{n,0} - v_n\|_\infty
\]
\[
\leq \gamma \|v_{\pi_{n+1}} - v_n\|_\infty + \|v_{n,0} - v_n\|_\infty
\]

Along the same line,
\[
\|v_s - v_n\|_\infty \leq \|v_s - v_{n,0}\|_\infty + \|v_{n,0} - v_n\|_\infty
\]
\[
\leq \|T_s v_s - T_s v_n\|_\infty + \|v_{n,0} - v_n\|_\infty
\]
\[
\leq \gamma \|v_s - v_n\|_\infty + \|v_{n,0} - v_n\|_\infty
\]

Combining those two inequalities yields
\[
\|v_{\pi_{n+1}} - v_s\|_\infty \leq \frac{2}{1 - \gamma} \|v_n - v_{0,n}\|_\infty
\]

As show before,
\[
0 \leq v_s - v_{n+1} \leq v_s - v_n - \sum_{k=0}^{m_n} \gamma^k P_{\pi_{n+1}} \Delta v_n
\]
Now, let \( \pi_* \) such that \( T_{\pi_*} v_* = B v_* \),
\[
\Delta_n = \Delta v_n - \Delta v_* = T_{\pi} v_n - v_n - (T_{\pi_*} v_* - v_*) \\
\leq T_{\pi} v_n - v_n - (T_{\pi_*} v_* - v_*) \\
\leq (\gamma P_{\pi_*} - \text{Id})(v_n - v_*)
\]
Thus
\[
0 \leq v_s - v_{n+1} \leq v_s - v_n - \sum_{k=0}^{m_n} \gamma^k P_{\pi_{n+1}}^k (\gamma P_{\pi_*} - \text{Id})(v_n - v_*) \\
\leq \sum_{k=1}^{m_n} \gamma^k P_{\pi_{n+1}}^k (v_n - v_*) - \sum_{k=0}^{m_n} \gamma^{k+1} P_{\pi_{n+1}} P_{\pi_*} (v_n - v_*) \\
\leq \sum_{k=0}^{m_n} \gamma^{k+1} P_{\pi_{n+1}}^k (P_{\pi_{n+1}} - P_{\pi_*})(v_n - v_*) - \gamma^{m_n+1} P_{\pi_{n+1}}^{m_n+1} (v_n - v_*) \\
\leq \sum_{k=0}^{m_n} \gamma^{k+1} \| P_{\pi_{n+1}} - P_{\pi_*} \| \| v_n - v_* \|_\infty + \gamma^{m_n+1} \| v_n - v_* \|_\infty \\
\leq \left( \frac{\gamma - \gamma^{m_n+1}}{1 - \gamma} \| P_{\pi_{n+1}} - P_{\pi_*} \| + \gamma^{m_n+1} \right) \| v_n - v_* \|_\infty
\]

\[ \square \]

### 2.3 Asynchronous Dynamic Programming

**Proposition 2.3.1**

Assume \( T_{\pi_0} v_0 \geq v_0 \) 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
    \[
    \pi_{n+1}(s) = \begin{cases} 
    \arg\max_a r(s, a) + \gamma P_{\pi_*} v_n(s) & \text{if } s \in S_n \\ 
    \pi_n(s) & \text{otherwise}
    \end{cases}
    \]

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 \( s_* \).
2 Discounted Reward

Proof. We start by proving by recursion that $T_{π_n}v_n ≥ v_n$ implies

$$T_{π_{n+1}}v_{n+1} ≥ v_{n+1} ≥ v_n \quad \text{and} \quad T_{π_n}v_n$$

Note that that $T_{π_0}v_0 ≥ v_0$ is an assumption.

Assume now that $T_{π_n}v_n ≥ v_n$, then either at step $n$ we update the value function or the policy.

If we update the value function, $π_{n+1} = π_n$ and thus

$$v_{n+1}(s) = \begin{cases} T_{π_n}v_n(s) & \text{if } s ∈ S_n \\ v_n(s) & \text{otherwise} \end{cases}$$

As $T_{π_n}v_n(s) ≥ v_n(s)$, we deduce $T_{π_{n+1}}v_{n+1} ≥ v_{n+1} ≥ v_n$. It suffices to notice that $v_{n+1} ≥ v_n$ implies

$$T_{π_{n+1}}v_{n+1} = T_{π_n}v_{n+1} ≥ T_{π_n}v_n$$

to obtain

$$T_{π_{n+1}}v_{n+1} ≥ v_{n+1} ≥ v_n.$$  

Now, if we update the policy, $v_{n+1} = v_n$ and

$$T_{π_{n+1}}v_{n+1}(s) = \begin{cases} T_{π_n}v_n(s) & \text{if } s ∈ S_n \\ T_{π_n}v_n(s) & \text{otherwise} \end{cases}$$

which implies $T_{π_{n+1}}v_n ≥ T_{π_n}v_n$ and thus as $v_{n+1} = v_n$

$$T_{π_{n+1}}v_{n+1} ≥ T_{π_n}v_n ≥ v_n = v_{n+1}.$$  

We deduce thus that

$$T_k v_{n+1} ≥ T_{π_{n+1}}v_{n+1} ≥ v_{n+1} ≥ v_n,$$

which implies if we take the limit in $k$

$$v_s ≥ v_{n+1} ≥ v_n.$$

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

$$v_n ≤ \tilde{v} ≤ T_0 \tilde{v} ≤ v_s.$$  

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 $n' ≥ n$

$$\tilde{v}(s) < T_s v_{n'}(s)$$

Let $n' ≥ n$ such that one updates the policy of $s$ and $n''$ the smallest integer larger than $n''$ where one updates the value of $s$.

$$v_{n''+1}(s) = T_{π_{n''}}v_{n''}(s) ≥ T_{π_{n'+1}}v_{n'+1}(s) ≥ T_{π_{n'+1}}v_{n'}(s) ≥ T_s v_{n'}(s) > \tilde{v}(s)$$

which is impossible. □
Proposition 2.4.1

If in a Generalized Policy Improvement, for all \( k \)

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

and

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

then

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

Proof. By construction,

\[
v_{\pi_k}(s) - v_{\pi_{k+1}}(s) = T_{\pi_k}v_{\pi_k}(s) - T_{\pi_{k+1}}v_{\pi_{k+1}}
\]

\[
= T_{\pi_k}v_{\pi_k}(s) - T_{\pi_k}v_k(s) + T_{\pi_k}v_k(s) - T_{\pi_{k+1}}v_{\pi_{k+1}}
\]

\[
\leq \gamma \epsilon + T_*v_k(s) - T_{\pi_{k+1}}v_{\pi_{k+1}}
\]

\[
\leq \gamma \epsilon + T_{\pi_{k+1}}v_k(s) - T_{\pi_{k+1}}v_{\pi_{k+1}} + \delta
\]

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

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

and thus

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

Now,

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

thus

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

and

\[ \limsup \max_s (v_k(s) - v_{\pi_{k}}(s)) \leq \limsup \gamma \max_s (v_k(s) - v_{\pi_{k}}(s)) + 2\gamma \epsilon + \delta + \gamma \frac{2\gamma \epsilon + \delta}{1 - \gamma} \]

which implies

\[ \limsup \max_s (v_k(s) - v_{\pi_{k}}(s)) \leq \frac{2\gamma \epsilon + \delta}{(1 - \gamma)^2} \]

\[ \Box \]
3 Finite Horizon

Proposition 3.1

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

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

If \( v_0 = r_* \) and \( v_{n+1} = T_*v_n \) then

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

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

Now,

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

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

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

Now,

\[
v_{s,T-n}(s) = \max_{\pi} \mathbb{E}_{\pi} \left[ \sum_{t=T-n}^{T-1} R_{t+1} \middle| S_{T-n-1} = s \right]
\]

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

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

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

\[
= T_{s} v_{s,T-n-1}(s)
\]
4 Non Discounted Total Reward

Definition 4.1
Let \( \tilde{s} \) be the absorbing state, we define the expected absorption time starting from \( s \) \( \tau_\pi(s) \) by
\[
\tau_\pi(s) = \mathbb{E}_\pi \left[ \inf_{S_t = \tilde{s}} t \mid S_0 = s \right].
\]
If \( \tau_\pi \) is finite, we say that \( \pi \) is proper.

Definition 4.2
We define the maximum expected absorption time starting from \( s \) by \( \tau_\ast(s) \) by
\[
\tau_\ast(s) = \max_\pi \tau_\pi(s).
\]

Proposition 4.3
If \( \tau_\pi < +\infty \) then
\[
\tau_\pi = 1 + P_\pi \tau_\ast = T_\pi \tau_\pi.
\]
If \( \tau_\ast < +\infty \) then
\[
\tau_\ast = \max_\pi 1 + P_\pi \tau_\ast = T \tau_\ast.
\]

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

Proposition 4.4
\( T_\pi \) is a contraction of factor \( \max \frac{\tau_\pi(s) - 1}{\tau_\pi(s)} \) with respect to the norm \( \| \cdot \|_{\infty, 1/\tau_\pi} \).
\( T_\pi \) and \( T_\ast \) are contraction of factor \( \max \frac{\tau_\ast(s) - 1}{\tau_\ast(s)} \) with respect to the norm \( \| \cdot \|_{\infty, 1/\tau_\ast} \).
4 Non Discounted Total Reward

Proof.

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

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 \( \tau = \tau_* \). \qed
5 Bandits

5.1 Regret

Definition 5.1.1
A $k$-armed bandit is defined by a collection of $k$ random variables $R(a)$, $a \in \{1, \ldots, k\}$. The best arm is $a_*$ 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 5.1.2
Let $T_t(a) = \sum_{s=1}^{t} 1_{A_s = a}$ and $\Delta(a) = \mathbb{E}[R(a_*)] - \mathbb{E}[R(a)]$ then

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

Proof. By definition,

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

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

$$= \mathbb{E}\left[\sum_{t=1}^{T} \sum_{a=1}^{k} 1_{A_t = a} (\mathbb{E}[R(a_*)] - R(a))\right]$$

$$= \sum_{a=1}^{k} \mathbb{E}\left[\sum_{t=1}^{T} 1_{A_t = a} (\mathbb{E}[R(a_*)] - R(a))\right]$$

$$= \sum_{a=1}^{k} \mathbb{E}\left[\sum_{t=1}^{T} 1_{A_t = a} \Delta(a)\right]$$

$$= \sum_{a=1}^{k} \mathbb{E}[T_t(a)] \Delta(a)$$
5.2 Concentration of subgaussian random variables

**Definition 5.2.1**
A random variable $X$ is said to be $\sigma$-subgaussian if
$$\mathbb{E}[\exp \lambda X] \leq \exp(\lambda^2 \sigma^2 / 2)$$

**Proposition 5.2.2**
If $X$ is $\sigma$-subgaussian then for any $\epsilon > 0$
$$\mathbb{P}(X \geq \epsilon) \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$$

*Proof.*

$$\mathbb{P}(X \geq \epsilon) = \mathbb{P}(\exp(\lambda X) \geq \exp(\lambda \epsilon))$$
$$\leq \frac{\mathbb{E}[\exp(\lambda X)]}{\exp(\lambda \epsilon)}$$
$$\leq \exp(\lambda^2 \sigma^2 / 2 - \lambda \epsilon)$$
$$\leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$$

where the last inequality is obtained by optimizing in $\lambda$.

**Proposition 5.2.3**
If $X$ is $\sigma$-subgaussian and $Y$ is $\sigma'$-subgaussian conditionally to $X$ then
- $\mathbb{E}[X] = 0$ and $\text{Var}[X] \leq \sigma^2$
- $cX$ is $|c|\sigma$-subgaussian.
- $X + Y$ is $\sqrt{\sigma^2 + (\sigma')^2}$-subgaussian.

*Proof.*

$$\mathbb{E}[\exp \lambda X] = \sum_k \frac{\lambda^k}{k!} \mathbb{E} \left[ X^k \right]$$
5.3 Explore Then Commit strategy

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

$$\lambda \mathbb{E}[X] \leq 0 \quad \text{and} \quad \frac{\lambda^2}{2!} \mathbb{E}[X^2] \leq \frac{\lambda^2 \sigma^2}{2 \times 1!}$$

which implies

$$\mathbb{E}[X] = 0 \quad \text{and} \quad \mathbb{V}[X] \leq \sigma^2.$$ By definition,

$$\mathbb{E}[\exp(\lambda c X)] \leq \exp(\lambda^2 c^2 \sigma^2 / 2)$$

hence the $|c| \sigma$-subgaussianity of $cX$.

Now,

$$\mathbb{E}[\exp(\lambda(X + Y))] \leq \mathbb{E}[\mathbb{E}[\exp(\lambda(X + Y))|X]] \leq \mathbb{E}[\mathbb{E}[\exp(\lambda X) \exp(\lambda Y)|X]] \leq 
\exp\left(\lambda^2 (\sigma^2 + (\sigma')^2)/2\right) \right)$$


\[\square\]

**Proposition 5.2.4**

*If $X_i \sim \mu$ 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.*

\\

5.3 Explore Then Commit strategy

**Definition 5.3.1**

*The simple current mean estimate $Q_t(a)$ is defined by*

$$Q_t(a) = \frac{1}{T_t(a)} \sum_{s=1}^{t} \mathbf{1}_{A_s = a} R_s(a)$$
Proposition 5.3.2
Assume we play the arm successively during \( K m \) 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

\[
    r_{T,\pi} \leq \min(m, T/K) \sum_{a=1}^{k} \Delta(a) + \max(T - mK, 0) \sum_{a=1}^{k} \Delta(a) \exp(-m\Delta(a)^2/4)
\]

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

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

\[
    \leq \min(m, n/K) + \max(n - mK, 0) \mathbb{P}
    \left( Q_T(a) \geq \max_{a' \neq a} Q_T(a') \right)
\]

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

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

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

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

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

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

\[
    \square
\]

5.4 \( \epsilon \)-greedy strategy
5.4 \( \epsilon \)-greedy strategy

Proposition 5.4.1
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

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

hence the first result. \( \square \)

Proposition 5.4.2
Let \( \pi \) be an \( \epsilon_t \)-greedy strategy,

\[
 P(A_T = a^*) \geq 1 - \epsilon_T - \Sigma_t \exp\left(-\Sigma_T/(6k)\right) - \sum_{a \neq a^*} \frac{4}{\Delta(a)^2} e^{-\Delta(a)^2 \Sigma_T/(4k)}
\]

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

Furthermore,

\[
 P(a^* = \arg\max Q_{T,a}) \geq 1 - \Sigma_t \exp\left(-\Sigma_T/(6k)\right) - \sum_{a \neq a^*} \frac{4}{\Delta(a)^2} e^{-\Delta(a)^2 \Sigma_T/(4k)}
\]

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

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

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

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

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

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

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

and

\[
 P(Q_T(a) \geq Q_T(a^*)) \leq P(Q_T(a) \geq \mu(a) + \Delta(a)/2) + P(Q_T(a^*) \leq \mu(a^*) - \Delta(a)/2).
\]
5 Bandits

By symmetry, it suffices to bound

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

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

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

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

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

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

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

Now the Bernstein inequality yields

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

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

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

\[ \leq \sum_{s=1}^{t} \frac{\epsilon_s}{k}, \]

32
5.4 \( \epsilon \)-greedy strategy

Choosing \( T_0 = \frac{1}{2} \sum \epsilon \) leads

\[
\mathbb{P}(T_1^R(a) \leq T_0) = \mathbb{P}(T_1^R(a) \leq 2T_0 - T_0)
\]

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

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

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

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

\[
\mathbb{P}(a = \text{argmax}_{T} Q_T(a)) \leq 2(1 - \frac{\epsilon_T}{k}) \left(\sum_T/(2k) \exp(-\Delta_T/(6k)) + \frac{2}{\Delta(a)^2} e^{-\Delta(a)^2 \Sigma_T/4}\right)
\]

\[
\leq \frac{\epsilon_T}{k} + \frac{\Sigma_T}{k} \exp(-\Sigma_T/(6k)) + \frac{4}{\Delta(a)^2} e^{-\Delta(a)^2 \Sigma_T/(4k)}
\]

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

\[
\mathbb{P}(A_T = a) \leq \frac{\epsilon_T}{k} + \frac{\Sigma_T}{k} \exp(-\Sigma_T/(6k)) + \frac{4}{\Delta(a)^2} e^{-\Delta(a)^2 \Sigma_T/(4k)}
\]

which goes to 0 if furthermore \( \epsilon_T \) tends to 0.

Finally,

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

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

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

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

\[
\leq \sum_{a \neq a^*} \left(\Delta(a) \left(\frac{\log(T) + 1}{k} + C\right) + \frac{4}{\Delta(a)} C'\right)
\]

33
5 Bandits

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

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

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

5.5 UCB strategy

**Proposition 5.5.1**

Assume we use a UCB strategy with a variance term \( \sqrt{\frac{c \log t}{T(a)}} \) 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 \( t \geq \max_a \frac{4c \ln t}{\Delta(a)/4} \).
5.5 UCB strategy

Proof. By construction,

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

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

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

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

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

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

\[ + \frac{1}{s''} \sum_{j=1}^{s''} j \cdot 1_{s''}^R(a_{(j)}(j) \leq \mu(a_{s'}) - \sqrt{\frac{\ln s}{s''}}} \]

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

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

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

\[ \mathbb{P}(A_t = a) \leq 2t^{-2c} \]
5 Bandits

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

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

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

6.1 Convergence of a mean

**Proposition 6.1.1**

Assume $X_i$ are i.i.d. such that $\mathbb{E}[X_i|F_{i-1}] = \mu$ and $\mathbb{V}ar[X_i|F_{i-1}] \leq \sigma^2$, 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 $\limsup \|M_n - \mu\|^2 \leq \alpha \sigma^2$

**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 $\sum_{i=1}^n \alpha_i \to +\infty$. 
Choosing \( m = n/2 \) yields

\[
E[\|M_n - \mu\|^2] \leq e^{-\sum_{k=1}^{m} \alpha_k} \|M_0 - \mu\|^2 + e^{-\sum_{k=m/2}^{n} \alpha_k} \|M_0 - \mu\|^2 + \alpha \sigma^2
\]

If we assume that \( \sum_{k=1}^{n} \alpha_k \to +\infty \) 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

\[
E[\|M_n - \mu\|^2] \leq e^{-n \alpha} \|M_0 - \mu\|^2 + n e^{-n \alpha} \sigma^2 + \alpha \sigma^2
\]

which is yields the result.

6.2 Generic Stochastic Approximation

Definition 6.2.1

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 6.2.2

\( h \) and \( H_t \) are compatible if

\[
H_t(x) = h(x) + \epsilon_t(x) + \delta_t(x)
\]

with

\[
E[\epsilon_t(x)|\mathcal{F}_t] = 0 \quad \text{and} \quad \text{Var}[\epsilon_t(x)|\mathcal{F}_t] \leq c_0(1 + \|x\|^2)
\]
and with probability 1
\[ \|\delta_n(x)\|^2 \leq c_n(1 + \|x\|)^2 \]
with \( c_n \to 0 \) and either
- it exists a non negative \( V \in C^1 \) with \( L \)-Lipschitz gradient satisfying
  \[ \langle \nabla V(x), h(x) \rangle \leq -c\|\nabla V(x)\|^2 \]
  \[ \mathbb{E}\left[\|H_t(x)\|^2\right] \leq c'_0(1 + \|\nabla V(x)\|^2), \]
- or \( h \) is a contraction for the norm considered.

**Proposition 6.2.3** Generic Stochastic Approximation
Assume that for any \( i \), we have almost surely
\[ \sum_{i=1}^{T} \alpha_i \to +\infty \quad \text{and} \quad \sum_{i=1}^{T} \alpha_i^2 < +\infty \]
Then providing \( h \) and \( H_t \) are compatible,
\[ h(X_n) \to 0. \]
Proof. See Neuro-Dynamic programming from Bertsekas and Tsitsiklis.
\[ \square \]

**Lemma 6.2.4**


From \( \theta_{k+1} = \theta_k + \alpha_k h_k(\theta_k) \) with \( h_k(\theta) = H(\theta) + \epsilon_k + \eta_k \)

to \( \frac{d\theta}{dt} = H(\dot{\theta}) \)
Sketch. 

- Difference between $\theta$ and a solution of the ODE with $\bar{\theta}(t_k) = \theta_k$ at $t_{k+1}$:

$$
\theta(t_{k+1}) - \bar{\theta}(t_{k+1}) = \int_{t_k}^{t_{k+1}} \left( \dot{\theta}(u) - \bar{\dot{\theta}}(u) \right) du
$$

$$
= \sum_{k' = k}^{k+l-1} \int_{t_{k'}}^{t_{k'+1}} \left( H(\theta(t_k)) + \epsilon_k + \eta_k - H(\bar{\theta}(u)) \right) du
$$

$$
= \sum_{k' = k}^{k+l-1} \int_{t_{k'}}^{t_{k'+1}} \left( H(\theta(t_k)) - H(\bar{\theta}(u)) \right) du
$$

\[ + \sum_{k' = k}^{k+l-1} \alpha_{k'} \epsilon_{k'} + \sum_{k' = k}^{k+l-1} \alpha_{k'} \eta_{k'} \]

- The last two term are going to be small by construction...

- Difference between $\theta$ and a solution of the ODE with $\bar{\theta}(t_k) = \theta_k$ at $t_{k+l}$:

$$
\theta(t_{k+l}) - \bar{\theta}(t_{k+l}) = \sum_{k' = k}^{k+l-1} \int_{t_{k'}}^{t_{k'+1}} \left( H(\theta(t_k)) - H(\bar{\theta}(u)) \right) du
$$

\[ + \sum_{k' = k}^{k+l-1} \alpha_{k'} \epsilon_{k'} + \sum_{k' = k}^{k+l-1} \alpha_{k'} \eta_{k'} \]

- The last two term are going to be small by construction:

$$
\mathbb{E} \left[ \sum_{k' = k}^{k+l-1} \alpha_{k'} \epsilon_{k'} \right] = 0 \quad \text{and} \quad \text{Var} \left[ \sum_{k' = k}^{k+l-1} \alpha_{k'} \epsilon_{k'} \right] < \sigma^2 \sum_{k' = k}^{k+l-1} \alpha_{k'}^2 \rightarrow 0
$$

\[ \| \sum_{k' = k}^{k+l-1} \alpha_{k'} \eta_{k'} \| \leq (t_{k+l-1} - t_k) \sup_{k' \geq k} \| \eta_{k'} \| \]

which is small if we assume that $t_{k+l-1} - t_k \leq \Delta$.

- We can now use a Lipchitz assumption on $H$ to obtain:

$$
\left\| \int_{t_{k'}}^{t_{k'+1}} \left( H(\theta(t_k')) - H(\bar{\theta}(u)) \right) du \right\| \leq L \int_{t_{k'}}^{t_{k'+1}} \| \theta(t_k') - \bar{\theta}(u) \| du
$$

$$
\leq L \alpha_{k'} \| \theta(t_k') - \bar{\theta}(t_k') \| + L \int_{t_{k'}}^{t_{k'+1}} \| \theta(t_k') - \bar{\theta}(u) \| du
$$

$$
\leq L \alpha_{k'} \| \theta(t_k') - \bar{\theta}(t_k') \| + L \| H \|_{\infty} \alpha_{k'}^2
$$

- Combining all the results leads to

$$
\| \theta(t_{k+l}) - \bar{\theta}(t_{k+l}) \| \leq L \sum_{k' = k}^{k+l-1} \alpha_{k'} \| \theta(t_k') - \bar{\theta}(t_k') \|
$$

\[ + L \| H \|_{\infty} \sum_{k' = k}^{k+l-1} \alpha_{k'}^2 + \left\| \sum_{k' = k}^{k+l-1} \alpha_{k'} \epsilon_{k'} \right\| + \sum_{k' = k}^{k+l-1} \alpha_{k'} \| \eta_{k'} \| \]
6.3 TD(λ) and linear approximation

• Using a discrete Gronwall Lemma, \( \forall l \leq l', z_l \leq L \sum_{l'=0}^{l-1} \alpha_{l'} z_{l'} + A \Rightarrow z_l \leq A e^{L \sum_{l'=0}^{l-1} \alpha_{l'}} \),
we obtain that if \( t_{k+l} - t_k \leq \Delta \)

\[
\| \theta(t_{k+1}) - \hat{\theta}(t_{k+1}) \| \leq \left( L \| H \|_\infty \sum_{k'=k}^{\infty} \alpha_{k'}^2 + \sup_{l' \leq l} \| \sum_{k'=k}^{l'-1} \alpha_{k'} e_{k'} \| + L \sup_{k' \geq k} \| \eta_{k'} \| \right) e^{L \Delta} \rightarrow 0 \text{ when } k \rightarrow \infty
\]

\[ \square \]

6.3 TD(λ) and linear approximation

**Proposition 6.3.1**

Provided there is a unique stationary distribution \( \mu \) on the states, that the basis function are linearily 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 \( w_{TD} = A^{-1} b \)

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

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

In the continuous task setting,

\[
A = \sum_s \mu(s) \sum_a \pi(a|s) \sum_{r,s'} p(r,s'|s,a) x(s)(x(s) - \gamma x(s'))^t
\]

\[
= \sum_s \mu(s) \sum_a \pi(a|s) \sum_{s'} p_\pi(s'|s)x(s)(x(s) - \gamma x(s'))^t
\]

\[
= \sum_s \mu(s)x(s) \left( x(s) - \gamma \sum_{s'} p_\pi(s'|s)x(s') \right)^t
\]

\[
= X^T D(\text{Id} - \gamma P_\pi)X
\]

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

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

\[
= \mu^T - \gamma \mu^T P_\pi
\]

\[
= (1 - \gamma) \mu^T > 0
\]