

Reinforcement Learning Book of Proofs

E. Le Pennec

January 25, 2025

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	Modified 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	ϵ -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(λ) and linear approximation	41

1 History Dependent or Markov Policies

Proposition 1.1 Equivalence of History Dependent and Markov Policies

Let π be a stochastic history dependent policy. For each state $s_0 \in S$, there exists a Markov stochastic policy π' 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,

$$\begin{aligned} \mathbb{P}_\pi(S_t = s_t|S_0 = s_0) &= \sum_{s_{t-1}} \sum_{a_{t-1}} p(s_t|s_{t-1}, A_{t-1}) \mathbb{P}_\pi(S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}|S_0 = s_0) \\ &= \sum_{s_{t-1}} \sum_{a_{t-1}} p(s_t|s_{t-1}, a_{t-1}) \mathbb{P}_{\pi'}(S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}|S_0 = s_0) \\ &= \mathbb{P}_{\pi'}(S_t = s_t|S_0 = s_0). \end{aligned}$$

Hence,

$$\begin{aligned} \mathbb{P}_{\pi'}(S_t = s_t, A_t = a_t|s_0) &= \pi'(a_t|s_t) \mathbb{P}_{\pi'}(S_t = s_t|S_0 = s_0) \\ &= \mathbb{E}_\pi[\mathbb{P}_\pi(A_t = a_t|H_t, S_t = s_t, S_0 = s_0)] \mathbb{P}_\pi(S_t = s_t|S_0 = s_0) \\ &= \mathbb{E}_\pi[\mathbb{P}_\pi(S_t = s_t, A_t = a_t, H_t|S_0 = s_0)]. \end{aligned}$$

It suffices then to notice that the quality criterion of π and π' depends on π 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 2.1.1

Value Function

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_\pi \left[\sum_{t=0}^{+\infty} \gamma^t R_{t+1} \middle| S_0 = s \right] \\ &= \sum_{t=0}^{+\infty} \gamma^t \mathbb{E}_\pi [R_{t+1} | S_0 = s] \end{aligned}$$

Definition 2.1.2

Bellman Operator

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

Proposition 2.1.3

Value Function Characterization

Let π be a stationary Markov policy, if $0 < \gamma < 1$ then v_π 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_π 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,

$$\begin{aligned}
 v_\pi(s) &= \sum_{t=0}^{+\infty} \gamma^t \mathbb{E}_\pi[R_{t+1} | S_0 = s] \\
 &= \sum_{t=0}^{+\infty} \gamma^t \sum_{s'} \mathbb{P}_\pi(S_t = s' | S_0 = s) \mathbb{E}_\pi[R | S = s'] \\
 &= \sum_{t=0}^{+\infty} \gamma^t \sum_{s'} (P_\pi^t)_{s,s'} r_\pi(s') \\
 &= \sum_{t=0}^{+\infty} \gamma^t (P_\pi^t r_\pi)(s)
 \end{aligned}$$

and thus $v = v_\pi$. □

Proposition 2.1.4

Bellman Operator Property

The operator $\mathcal{T}^\pi$ satisfies the following contraction property

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

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

Proof. For any s ,

$$\begin{aligned}
 |\mathcal{T}^\pi(v) - \mathcal{T}^\pi(v')(s)| &= |\gamma P_\pi(v - v')(s)| \\
 &\leq \gamma \|v - v'\|_\infty
 \end{aligned}$$

because P_π 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. □

Proposition 2.1.5

Policy Prediction

For any v_0 , define $v_{n+1} = \mathcal{T}^\pi v_n$ then

$$\lim_{n \rightarrow \infty} v_n = v_\pi$$

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 $v_0 \geq \mathcal{T}^\pi v_0$ (respectively $v_0 \leq \mathcal{T}^\pi v_0$) then $v_0 \geq v_\pi$ (respectively $v_0 \leq v_\pi$) and v_n converges monotonously to v_π .

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

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

$$\begin{aligned} \|v_n - v_\pi\|_\infty &\leq \sum_{i=0}^k \|v_{n+i} - v_{n+i+1}\|_\infty + \|v_{n+k+1} - v_\pi\|_\infty \\ &\leq \frac{\gamma - \gamma^{k+2}}{1 - \gamma} \|v_n - v_{n-1}\|_\infty + \gamma^{n+k+1} \|v_0 - v_\pi\|_\infty \end{aligned}$$

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

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

Thus $v_1 = \mathcal{T}^\pi v_0 \leq v_0$ implies that v_n is a decreasing sequence whose limit is v_* , yielding the result. The increasing case is obtained with a similar proof. $\square$

2.2 Optimal Policy

2.2.1 Characterization

Definition 2.2.1

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 2.2.2**Optimal Bellman Operator**

$$\begin{aligned}
\mathcal{T}^*v(s) &= \max_a \mathbb{E}[R|S=s, A=a] + \gamma \sum_{s'} \mathbb{P}(S'=s'|S=s, A=a) v(s') \\
&= \max_a r(s, a) + \gamma \sum_{s'} p(s'|s, a) v(s')
\end{aligned}$$

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

$$\mathcal{T}^*v(s) = \max_{\pi \in \mathcal{S}} \mathcal{T}^\pi v(s)$$

or $\mathcal{T}^*v = \max_{\pi \in \mathcal{S}} r_\pi + \gamma P_\pi v$ where $\mathcal{S}$ is the set of deterministic Markov policies and the max is componentwise.

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

Now, for any π ,

$$\begin{aligned}
\mathcal{T}^\pi(s) &= \sum_a \pi(a|s) \left(\mathbb{E}[R|S=s, A=a] + \gamma \sum_{s'} p(s'|s, a) v(s') \right) \\
&\leq \max_a \mathbb{E}[R|S=s, A=a] + \gamma \sum_{s'} p(s'|s, a) v(s') \\
&\leq \mathcal{T}^*(s)
\end{aligned}$$

□

Proposition 2.2.4**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 $\mathcal{T}^*v(s) \geq \mathcal{T}^*v'(s)$

$$\begin{aligned}
 |\mathcal{T}^*v - \mathcal{T}^*v'(s)| &= \mathcal{T}^*v(s) - \mathcal{T}^*v'(s) \\
 &= \max_a r(s, a) + \gamma \sum_{s'} p(s'|s, a)v(s') - \left(\max_a r(s, a) + \gamma \sum_{s'} p(s'|s, a)v'(s') \right) \\
 &\leq \max_a \left(r(s, a) + \gamma \sum_{s'} p(s'|s, a)v(s') - \left(r(s, a) + \gamma \sum_{s'} p(s'|s, a)v'(s') \right) \right) \\
 &\leq \gamma \max_a \sum_{s'|s, a} p(s'|s, a)(v(s') - v'(s')) \\
 &\leq \gamma \|v - v'\|_\infty
 \end{aligned}$$

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

$$\begin{aligned}
 r(s, a') + \gamma \sum_{s'} p(s'|s, a')v(s') &\leq r(s, a') + \gamma \sum_{s'} p(s'|s, a')v'(s') \\
 &\leq \mathcal{T}^*v'(s)
 \end{aligned}$$

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

Finally,

$$\begin{aligned}
 \mathcal{T}^*(v + \delta \mathbb{1})(s) &= \max_a r(s, a) + \gamma \sum_{s'} p(s'|s, a)(v(s') + \delta) \\
 &= \max_a r(s, a) + \gamma \sum_{s'} p(s'|s, a)v(s') + \delta \\
 &= \mathcal{T}^*(v)(s) + \delta.
 \end{aligned}$$

Proposition 2.2.5

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

$$\begin{aligned}
 v &\geq r_{\pi_0} + \gamma P_{\pi_0}v \\
 v &\geq r_{\pi_0} + \gamma P_{\pi_0}(r_{\pi_1} + \gamma P_{\pi_1}v) \\
 v &\geq \sum_{k=0}^n \gamma^k P_{\pi}^t r_{\pi_k} + \gamma^{n+1} P_{\pi}^{n+1}v
 \end{aligned}$$

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

$$v - v_{\pi} \geq \gamma^{n+1} P_{\pi}^{n+1}v - \sum_{k=n+1}^{\infty} \gamma^k P_{\pi}^k r_{\pi_k}.$$

2 Discounted Reward

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

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

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

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

We deduce thus that $v = \mathcal{T}^*v$ 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 $\mathcal{T}^*$. $\square$

Proposition 2.2.6

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

Proof. This is a direct consequence of the previous theorem. $\square$

Proposition 2.2.7

Any stationary policy π_ verifying $\pi_* \in \operatorname{argmax}_\pi r_\pi + \gamma P_\pi v_*$ is optimal.*

Proof. By definition,

$$\begin{aligned} \mathcal{T}^{\pi_*} v_* &= r_{\pi_*} + P_{\pi_*} v_* \\ &= \max_\pi r_\pi + P_\pi v_* \\ &= \mathcal{T}^* v_* = v_*. \end{aligned}$$

Hence $v_{\pi_*} = v_*$ 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 π , π_+ define by

$$\pi_+ \in \operatorname{argmax}_{\pi'} r_{\pi'} + \gamma P_{\pi'} v_\pi$$

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 $(I - \gamma P_{\pi_+})^{-1} = \sum \gamma^k P_{\pi_+}^k$. $\square$

Proposition 2.2.9

Let $\Delta = \mathcal{T}^* - \text{Id}$, the policy iteration scheme satisfies

$$v_{n+1} = v_n + \sum_{k=0}^{\infty} \gamma^k P_{\pi_{n+1}}^k \Delta v_n.$$

Proof. As proved before,

$$v_{n+1} = (\text{Id} - \gamma P_{\pi_{n+1}})^{-1} r_{\pi_{n+1}}.$$

Now by construction,

$$\mathcal{T}^* v_n = \mathcal{T}^{\pi_{n+1}} v_n = r_{\pi_{n+1}} + \gamma P_{\pi_{n+1}} v_n$$

and thus

$$r_{\pi_{n+1}} = \Delta v_n + (\text{Id} - \gamma P_{\pi_{n+1}}) v_n.$$

This implies immediately

$$\begin{aligned} v_{n+1} &= v_n + (\text{Id} - \gamma P_{\pi_{n+1}})^{-1} \Delta v_n \\ &= v_n + \sum_{k=0}^{\infty} \gamma^k P_{\pi_{n+1}}^k \Delta v_n \end{aligned}$$

$\square$

2.2.3 Value Iteration

Proposition 2.2.10

For any v_0 , define $v_{n+1} = \mathcal{T}^* v_n$ then

$$\lim_{n \rightarrow \infty} v_n = v_*$$

and

$$\|v_n - v_*\|_{\infty} \leq \gamma^n \|v_0 - v_*\|_{\infty}$$

Furthermore,

$$\|v_n - v_*\|_{\infty} \leq \frac{\gamma}{1 - \gamma} \|v_n - v_{n-1}\|_{\infty}$$

Finally, if $v_0 \geq \mathcal{T}^* v_0$ (respectively $v_0 \leq \mathcal{T}^* v_0$) then $v_0 \geq v_*$ (respectively $v_0 \leq v_*$) and v_n converges monotonously to v_* .

2 Discounted Reward

Proof. For the first part of the proposition, we notice that v_* is the only fixed point of $\mathcal{T}^*$ which is a contraction. Hence, by the fixed point theorem, for any v_0 , the sequence defined by $v_{n+1} = \mathcal{T}^*v_n$ 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

$$\begin{aligned} \|v_n - v_*\|_\infty &\leq \sum_{i=0}^k \|v_{n+i} - v_{n+i+1}\|_\infty + \|v_{n+k+1} - v_*\|_\infty \\ &\leq \frac{\gamma - \gamma^{k+2}}{1 - \gamma} \|v_n - v_{n+1}\|_\infty + \gamma^{n+k+1} \|v_0 - v_*\|_\infty \end{aligned}$$

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

Proposition 2.2.11

For any v and any $\pi \in \operatorname{argmax}_\pi \mathcal{T}^\pi v$,

$$\|v_\pi - v_*\|_\infty \leq \frac{2\gamma}{1 - \gamma} \|v - v_*\|_\infty$$

If $v = \mathcal{T}^*v'$ then

$$\|v_\pi - v_*\|_\infty \leq \frac{2\gamma}{1 - \gamma} \|v - v'\|_\infty$$

Proof. By definition of π , $\mathcal{T}^\pi v = \mathcal{T}^*v$, hence

$$\begin{aligned} \|v_\pi - v_*\|_\infty &\leq \|v_\pi - \mathcal{T}^\pi v\|_\infty + \|\mathcal{T}^*v - v_*\|_\infty \\ &\leq \|\mathcal{T}^\pi v_\pi - \mathcal{T}^\pi v\|_\infty + \|\mathcal{T}^*v - \mathcal{T}^*v_*\|_\infty \\ &\leq \gamma \|v_\pi - v\|_\infty + \gamma \|v - v_*\|_\infty \\ &\leq \gamma \|v_\pi - v_*\|_\infty + 2\gamma \|v - v_*\|_\infty \end{aligned}$$

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

$$\begin{aligned}\|v_\pi - v\|_\infty &\leq \|\mathcal{T}^\pi v_\pi - \mathcal{T}^\pi v\|_\infty + \|\mathcal{T}^* v - \mathcal{T}^* v'\|_\infty \\ &\leq \gamma \|v_\pi - v\|_\infty + \gamma \|v - v'\|_\infty\end{aligned}$$

and thus

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

Along the same line,

$$\begin{aligned}\|v - v_*\|_\infty &\leq \|v - \mathcal{T}^* v\|_\infty + \|\mathcal{T}^* v - v_*\|_\infty \\ &\leq \|\mathcal{T}^* v' - \mathcal{T}^* v\|_\infty + \|\mathcal{T}^* v - \mathcal{T}^* v_*\|_\infty \\ &\leq \gamma \|v - v'\|_\infty + \gamma \|v - v_*\|_\infty\end{aligned}$$

and thus

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

. Combining those two bounds yields the result. $\square$

2.2.4 Modified Policy Iteration

Proposition 2.2.12

MPI

Let v_0 such that $\mathcal{T}^* v_0 \geq v_0$, define for any n and any m_n

- $\pi_{n+1} \in \operatorname{argmax}_\pi r_\pi + P_\pi v_n$
- $v_{n,0} = \mathcal{T}^* v_n = \mathcal{T}^{\pi_{n+1}} v_n$
- $v_{n,m} = \mathcal{T}^{\pi_{n+1}} v_{n,m-1}$
- $v_{n+1} = v_{m_n}$

then $v_{n+1} \geq v_n$ and

$$\lim_{n \rightarrow \infty} v_n = v_*.$$

At any step,

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

Furthermore,

$$\|v_{n+1} - v_*\|_\infty \leq \left(\frac{\gamma - \gamma^{m_n+1}}{1-\gamma} \|P_{\pi_{n+1}} - P_{\pi_*}\| + \gamma^{m_n+1} \right) \|v_n - v_*\|_\infty$$

Proposition 2.2.13

Let $\Delta = \mathcal{T}^* - \text{Id}$, let $W_\pi^{(m)}v = (\mathcal{T}^\pi)^{m+1}v$,

$$\begin{aligned} W_\pi^{(m)}v &= \sum_{k=0}^m \gamma^k P_\pi^k r_\pi + \gamma^{m+1} P_\pi^{m+1}v \\ &= v_n + \sum_{k=0}^m \gamma^k P_\pi^k \Delta v \end{aligned}$$

Proof. By definition,

$$\begin{aligned} W_\pi^{(m)}v &= (\mathcal{T}^\pi)^{m+1}v \\ &= r_\pi + \gamma P_\pi (\mathcal{T}^\pi)^m v \\ &= \sum_{k=0}^m \gamma^k P_\pi^k r_\pi + \gamma^{m+1} P_\pi^{m+1}v \\ &= \sum_{k=0}^m \gamma^k P_\pi^k (r_\pi + \gamma P_\pi v - v) + v \\ &= v + \sum_{k=0}^m \gamma^k P_\pi^k \Delta v \end{aligned}$$

□

Proposition 2.2.14

Define $W_*^{(m_n)}$ by

$$W_*^{(m_n)}v(s) = \max_{\pi} W_\pi^{(m_n)}v(s).$$

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 $\tilde{\pi} \in \arg\max_{\pi} W_\pi^{(m_n)}v(s)$,

$$\begin{aligned} W_*^{(m_n)}v(s) - W_*^{(m_n)}v'(s) &= \max_{\pi} W_\pi^{(m_n)}v(s) - \max_{\pi} W_\pi^{(m_n)}v'(s) \\ &\leq W_{\tilde{\pi}}^{(m_n)}v(s) - W_{\tilde{\pi}}^{(m_n)}v'(s) \\ &\leq \gamma^{m_n+1} P_{\tilde{\pi}}^{m_n+1}(v - v')(s) \\ &\leq \gamma^{m_n+1} \|v - v'\|_\infty \end{aligned}$$

By construction $\Delta v_* = \mathcal{T}^* v_* - v_* = 0$ and thus $W_\pi^{(m_n)} v_* = v_*$. We deduce immediately that $W_*^{(m_n)} v_* = \sup_\pi W_\pi^{(m_n)} v_* = v_*$ $\square$

Proposition 2.2.15

If $u \geq v$ then for any π , $W_*^m u \geq W_\pi^m v$

If $u \geq v$ and $\Delta u \geq 0$ then for any π $W_\pi u \geq \mathcal{T}^* v$.

If $\Delta u \geq 0$ and π_u such that $\mathcal{T}^* u = \mathcal{T}^{\pi_u} u$ then $W_{\pi_u}^{(m)} u \geq 0$

Proof. By definition,

$$\begin{aligned} W_*^m u - W_\pi^m v &\geq W_\pi^m u - W_\pi^m v \\ &\geq W_\pi^m (u - v) \\ &\geq \gamma^{m_n+1} P_\pi^{m_n+1} (u - v) \geq 0 \end{aligned}$$

Now,

$$\begin{aligned} W_\pi^{(m)} u &= u + \sum_{k=0}^m \gamma^k P_\pi^k \Delta u \\ &\geq u + \Delta u = \mathcal{T}^* u \\ &\geq \mathcal{T}^* v \end{aligned}$$

By construction

$$\begin{aligned} \Delta W_{\pi_u}^{(m)} u &= \mathcal{T}^* W_{\pi_u}^{(m)} u - W_{\pi_u}^{(m)} u \\ &\geq \mathcal{T}^{\pi_u} W_{\pi_u}^{(m)} u - W_{\pi_u}^{(m)} u \\ &\geq \Delta u - \mathcal{T}^{\pi_u} u + u \\ &\geq \Delta u + (\gamma P_{\pi_u} - \text{Id}) (W_{\pi_u}^{(m)} u - u) \geq \Delta u + (\gamma P_{\pi_u} - \text{Id}) \sum_{k=0}^m \gamma^k P_{\pi_u}^k \Delta u \\ &\geq \gamma^m P_{\pi_u}^m \Delta u \geq 0 \end{aligned}$$

$\square$

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

By construction $\mathcal{T}^{\pi_{n+1}} v_n = \mathcal{T}^* v_n$ and one verify easily that $v_{n+1} = (\mathcal{T}^{\pi_{n+1}})^{m_n+1} v_n = W_{\pi_{n+1}}^{(m_n)} v_n$.

Define now, $u_{n+1} = \mathcal{T}^* u_n$ 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 $v_1 = W_{\pi_1}^{(m_n)} v_0 \geq \mathcal{T}^* v_0 \geq v_0$.

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

$$u_n = \mathcal{T}^* u_{n-1} \leq v_n = W_{\pi_n}^{(m_{n-1})} v_{n-1} \leq w_n = W_*^{(m_{n-1})} w_{n-1}$$

2 Discounted Reward

Finally,

$$\begin{aligned} v_n &= W_{\pi_n}^{(m_{n-1})} v_{n-1} \\ &= v_{n-1} + \sum_{k=0} m_{n-1} \gamma^k P_{\pi_n} \Delta v_{n-1} \\ &\geq v_{n-1}. \end{aligned}$$

Now, we have already proved that $u_n = \mathcal{T}^* u_0$ 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

$$\begin{aligned} \|w_n - v_*\|_\infty &= \|W_*^{(m_{n-1})} w_{n-1} - W_*^{(m_{n-1})} v_*\|_\infty \\ &= \gamma^{m_{n-1}} \|w_{n-1} - v_*\|_\infty \\ &= \gamma^{\sum_{k=0}^{n-1} m_k} \|v_0 - v_*\|_\infty \end{aligned}$$

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 $v_{n,0} = \mathcal{T}^{\pi_{n+1}} v_n = \mathcal{T}^* v_n$ so that

$$\begin{aligned} \|v_{\pi_{n+1}} - v_n\|_\infty &\leq \|v_{\pi_{n+1}} - v_{n,0}\|_\infty + \|v_{n,0} - v_n\|_\infty \\ &\leq \|\mathcal{T}^{\pi_{n+1}} v_{\pi_{n+1}} - \mathcal{T}^{\pi_{n+1}} v_n\|_\infty + \|v_{n,0} - v_n\|_\infty \\ &\leq \gamma \|v_{\pi_{n+1}} - v_n\|_\infty + \|v_{n,0} - v_n\|_\infty \end{aligned}$$

Along the same line,

$$\begin{aligned} \|v_* - v_n\|_\infty &\leq \|v_* - v_{n,0}\|_\infty + \|v_{n,0} - v_n\|_\infty \\ &\leq \|\mathcal{T}^* v_* - \mathcal{T}^* v_n\|_\infty + \|v_{n,0} - v_n\|_\infty \\ &\leq \gamma \|v_* - v_n\|_\infty + \|v_{n,0} - v_n\|_\infty \end{aligned}$$

Combining those two inequalities yields

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

As show before,

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

2.3 Asynchronous Dynamic Programming

Now, let π_* such that $\mathcal{T}^{\pi_*}v_* = Bv_*$,

$$\begin{aligned}\Delta_n &= \Delta v_n - \Delta v_* = \mathcal{T}^*v_n - v_n - (\mathcal{T}^*v_* - v_*) \\ &\leq \mathcal{T}^{\pi_*}v_n - v_n - (\mathcal{T}^{\pi_*}v_* - v_*) \\ &\leq (\gamma P_{\pi_*} - \text{Id})(v_n - v_*)\end{aligned}$$

Thus

$$\begin{aligned}0 \leq v_* - v_{n+1} &\leq v_* - v_n - \sum_{k=0}^{m_n} \gamma^k P_{\pi_{n+1}}^k (\gamma P_{\pi_*} - \text{Id})(v_n - v_*) \\ &\leq \sum_{k=1}^{m_n} \gamma^k P_{\pi_{n+1}}^k (v_n - v_*) - \sum_{k=0}^{m_n} \gamma^{k+1} P_{\pi_{n+1}} P_{\pi_*} (v_n - v_*) \\ &\leq \sum_{k=0}^{m_n} \gamma^{k+1} P_{\pi_{n+1}}^k (P_{\pi_{n+1}} - P_{\pi_*})(v_n - v_*) - \gamma^{m_n+1} P_{\pi_{n+1}}^{m_n+1} (v_n - v_*) \\ &\leq \sum_{k=0}^{m_n} \gamma^{k+1} \|P_{\pi_{n+1}} - P_{\pi_*}\| \|v_n - v_*\|_\infty + \gamma^{m_n+1} \|v_n - v_*\|_\infty \\ &\leq \left(\frac{\gamma - \gamma^{m_n+1}}{1 - \gamma} \|P_{\pi_{n+1}} - P_{\pi_*}\| + \gamma^{m_n+1} \right) \|v_n - v_*\|_\infty\end{aligned}$$

□

2.3 Asynchronous Dynamic Programming

Proposition 2.3.1

Assume $\mathcal{T}^{\pi_0}v_0 \geq v_0$ 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

$$v_{n+1}(s) = \begin{cases} \mathcal{T}^{\pi_n}v_n(s) & \text{if } s \in S_n \\ v_n(s) & \text{otherwise} \end{cases}$$

- keep the value function $s_{n+1} = s_n$ and update the policy following

$$\pi_{n+1}(s) = \begin{cases} \operatorname{argmax}_a r(s, a) + \gamma P_{\pi_a} v_n(s) & \text{if } s \in S_n \\ \pi_n(s) & \text{otherwise} \end{cases}$$

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 $\mathcal{T}^{\pi_n} v_n \geq v_n$ implies

$$\mathcal{T}^{\pi_{n+1}} v_{n+1} \geq v_{n+1} \geq v_n \quad \text{and} \quad \mathcal{T}^{\pi_n} v_n$$

Note that that $\mathcal{T}^{\pi_0} v_0 \geq v_0$ is an assumption.

Assume now that $\mathcal{T}^{\pi_n} v_n \geq v_n$, 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

$$v_{n+1}(s) = \begin{cases} \mathcal{T}^{\pi_n} v_n(s) & \text{if } s \in S_n \\ v_n(s) & \text{otherwise} \end{cases}$$

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

$$\mathcal{T}^{\pi_{n+1}} v_{n+1} = \mathcal{T}^{\pi_n} v_{n+1} \geq \mathcal{T}^{\pi_n} v_n$$

to obtain

$$\mathcal{T}^{\pi_{n+1}} v_{n+1} \geq v_{n+1} \geq v_n.$$

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

$$\mathcal{T}^{\pi_{n+1}} v_n(s) = \begin{cases} \mathcal{T}^* v_n(s) & \text{if } s \in S_n \\ \mathcal{T}^{\pi_n} v_n(s) & \text{otherwise} \end{cases}$$

which implies $\mathcal{T}^{\pi_{n+1}} v_n \geq \mathcal{T}^{\pi_n} v_n$ and thus as $v_{n+1} = v_n$

$$\mathcal{T}^{\pi_{n+1}} v_{n+1} \geq \mathcal{T}^{\pi_n} v_n \geq v_n = v_{n+1}.$$

We deduce thus that

$$\mathcal{T}^* v_{n+1} \geq \mathcal{T}^{\pi_{n+1}} v_{n+1} \geq v_{n+1} \geq v_n.$$

which implies if we take the limit in k

$$v_* \geq v_{n+1} \geq v_n.$$

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

$$v_n \leq \tilde{v} \leq \mathcal{T}^* \tilde{v} \leq v_*.$$

Assume now that there exists s such that $\tilde{v}(s) < \mathcal{T}^* \tilde{v}(s)$. By continuity of $\mathcal{T}^*$, there exists n such that for all $n' \geq n$

$$\tilde{v}(s) < \mathcal{T}^* v_{n'}(s)$$

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 .

$$\begin{aligned} v_{n''+1}(s) &= \mathcal{T}^{\pi_{n''}} v_{n''}(s) \\ &\geq \mathcal{T}^{\pi_{n'+1}} v_{n'+1}(s) \\ &\geq \mathcal{T}^{\pi_{n'+1}} v_{n'}(s) \\ &\geq \mathcal{T}^* v_{n'}(s) > \tilde{v}(s) \end{aligned}$$

which is impossible. □

2.4 Approximate Dynamic Programming

Proposition 2.4.1

If in a Generalized Policy Improvement, for all k

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

and

$$\|\mathcal{T}^{\pi_{k+1}} v_k - \mathcal{T}^* v_k\|_{\infty} \leq \delta$$

then

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

Proof. By construction,

$$\begin{aligned} v_{\pi_k}(s) - v_{\pi_{k+1}}(s) &= \mathcal{T}^{\pi_k} v_{\pi_k}(s) - \mathcal{T}^{\pi_{k+1}} v_{\pi_{k+1}} \\ &= \mathcal{T}^{\pi_k} v_{\pi_k}(s) - \mathcal{T}^{\pi_k} v_k(s) + \mathcal{T}^{\pi_k} v_k(s) - \mathcal{T}^{\pi_{k+1}} v_{\pi_{k+1}} \\ &\leq \gamma\epsilon + \mathcal{T}^* v_k(s) - \mathcal{T}^{\pi_{k+1}} v_{\pi_{k+1}} \\ &\leq \gamma\epsilon + \mathcal{T}^{\pi_{k+1}} v_k(s) - \mathcal{T}^{\pi_{k+1}} v_{\pi_{k+1}} + \delta \\ &\leq \gamma\epsilon + \mathcal{T}^{\pi_{k+1}} v_k(s) + \mathcal{T}^{\pi_{k+1}} v_{\pi_k}(s) - \mathcal{T}^{\pi_{k+1}} v_{\pi_k}(s) - \mathcal{T}^{\pi_{k+1}} v_{\pi_{k+1}} + \delta \\ &\leq 2\gamma\epsilon + \delta + \gamma \max_{s'} (v_{\pi_k}(s') - v_{\pi_{k+1}}(s')) \end{aligned}$$

and thus

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

Now,

$$\begin{aligned} v_*(s) - v_{\pi_{k+1}}(s) &= v_*(s) - \mathcal{T}^{\pi_{k+1}} v_{\pi_{k+1}}(s) \\ &= v_*(s) - \mathcal{T}^{\pi_{k+1}} v_{\pi_k}(s) + \mathcal{T}^{\pi_{k+1}} v_{\pi_k}(s) - \mathcal{T}^{\pi_{k+1}} v_{\pi_{k+1}}(s) \\ &\leq v_*(s) - \mathcal{T}^{\pi_{k+1}} v_{\pi_k}(s) + \gamma \frac{2\gamma\epsilon + \delta}{1 - \gamma} \\ &\leq v_*(s) - \mathcal{T}^{\pi_{k+1}} v_k(s) + \gamma\epsilon + \gamma \frac{2\gamma\epsilon + \delta}{1 - \gamma} \\ &\leq v_*(s) - \mathcal{T}^* v_k(s) + \gamma\epsilon + \delta + \gamma \frac{2\gamma\epsilon + \delta}{1 - \gamma} \\ &\leq \mathcal{T}^* v_*(s) - \mathcal{T}^* v_{\pi_k}(s) + 2\gamma\epsilon + \delta + \gamma \frac{2\gamma\epsilon + \delta}{1 - \gamma} \\ &\leq \gamma \max_s (v_*(s) - v_{\pi_k}(s)) + 2\gamma\epsilon + \delta + \gamma \frac{2\gamma\epsilon + \delta}{1 - \gamma} \end{aligned}$$

2 Discounted Reward

thus

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

and

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

which implies

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

□

3 Finite Horizon

Proposition 3.1

If $v_0 = r_{\pi, T-1}$ and $v_n = \mathcal{T}^{\pi, T-n} v_{n-1} = r_{\pi, T-n} + P_{\pi, T-n} v_{n-1}$ then

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

If $v_0 = r_*$ and $v_{n+1} = \mathcal{T}^* v_n$ then

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

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,

$$\begin{aligned} v_{\pi, T-n}(s) &= \mathbb{E}_{\pi} \left[\sum_{t=T-n-1}^{T-1} R_{t+1} \middle| S_{t-n-1} = s \right] \\ &= r_{\pi, T-n-1}(s) + \mathbb{E}_{\pi} \left[\sum_{t=T-n}^{T-1} R_{t+1} \middle| S_{t-n-1} = s \right] \\ &= r_{\pi, T-n-1}(s) + \sum_a \sum_{s'} p(s'|s, a) \pi(a|s) \mathbb{E}_{\pi} \left[\sum_{t=T-n}^{T-1} R_{t+1} \middle| S_{t-n} = s' \right] \\ &= r_{\pi, T-n-1}(s) + P_{\pi, T-n-1} v_{\pi, T-n-1}(s) \end{aligned}$$

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,

$$\begin{aligned}
v_{*,T-n}(s) &= \max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=T-n-1}^{T-1} R_{t+1} \middle| S_{T-n-1} = s \right] \\
&= \max_{\pi} \left(r_{\pi}(s) + \mathbb{E} \left[\sum_{t=T-n}^{T-1} R_{t+1} \middle| S_{T-n-1} = s \right] \right) \\
&= \max_{\pi} \left(r_{\pi,T-n-1}(s) + \sum_a \sum_a p(s'|s, a) \pi(a|s) \mathbb{E} \left[\sum_{t=T-n}^{T-1} R_{t+1} \middle| S_{t-n} = s' \right] \right) \\
&= \max_{\pi} r_{\pi,T-n-1}(s) + P_{\pi,T-n-1} \max_{\pi} v_{\pi,T-n-1}(s) \\
&= \mathcal{T}^* v_{*,T-n-1}(s)
\end{aligned}$$

□

4 Non Discounted Total Reward

Definition 4.1

Let s_{abs} 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 τ_π is finite, we say that π is proper.

Definition 4.2

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

$$\tau_*(s) = \max_{\pi} \tau_\pi(s)$$

Proposition 4.3

If $\tau_\pi < +\infty$ then

$$\tau_\pi = 1 + P_\pi \tau_\pi = \mathcal{T}^\pi \tau_\pi$$

If $\tau_* < +\infty$ then

$$\tau_* = \max_{\pi} 1 + P_\pi \tau_* = \mathcal{T} \tau_*$$

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

$\mathcal{T}^\pi$ is a contraction of factor $\max \frac{\tau_\pi(s)-1}{\tau_\pi(s)}$ with respect to the norm $\|\cdot\|_{\infty, 1/\tau_\pi}$

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

4 Non Discounted Total Reward

Proof.

$$\begin{aligned}
|\mathcal{T}^\pi v(s) - \mathcal{T}^\pi v'(s)| &\leq |P_\pi(v - v')(s)| \\
&\leq P_\pi(\tau \times \frac{|v - v'|}{\tau})(s) \\
&\leq P_\pi \tau(s) \|v - v'\|_{\infty, 1/\tau} \\
&\leq \tau(s) \frac{1 + P_\pi \tau(s) - 1}{\tau(s)} \|v - v'\|_{\infty, 1/\tau} \\
&\leq \tau(s) \frac{1 + P_* \tau(s) - 1}{\tau(s)} \|v - v'\|_{\infty, 1/\tau}
\end{aligned}$$

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

Now, assume without loss of generality that $\mathcal{T}^* v(s) \geq \mathcal{T}^* v'(s)$,

$$\begin{aligned}
|\mathcal{T}^* v(s) - \mathcal{T}^* v'(s)| &= \max_{\pi} \mathcal{T}^\pi v(s) - \max_{\pi} \mathcal{T}^\pi v'(s) \\
&\leq \max_{\pi} (\mathcal{T}^\pi v(s) - \mathcal{T}^\pi v'(s)) \\
&\leq \tau(s) \frac{1 + P_* \tau(s) - 1}{\tau(s)} \|v - v'\|_{\infty, 1/\tau}
\end{aligned}$$

which yields the result for $\tau = \tau_*$. □

5 Bandits

5.1 Regret

Definition 5.1.1

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

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

For any policy π , 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 π .

Proposition 5.1.2

Let $T_t(a) = \sum_{s=1}^t \mathbf{1}_{A_s=a}$ 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,

$$\begin{aligned} r_{T,\pi} &= n\mathbb{E}[R(a_*)] - \mathbb{E}\left[\sum_{t=1}^T R(A_t)\right] \\ &= \mathbb{E}\left[\sum_{t=1}^T (\mathbb{E}[R(a_*)] - R(A_t))\right] \\ &= \mathbb{E}\left[\sum_{t=1}^T \sum_{a=1}^k \mathbf{1}_{A_t=a} (\mathbb{E}[R(a_*)] - R(a))\right] \\ &= \sum_{a=1}^k \mathbb{E}\left[\sum_{t=1}^T \mathbf{1}_{A_t=a} (\mathbb{E}[R(a_*)] - R(a))\right] \\ &= \sum_{a=1}^k \mathbb{E}\left[\sum_{t=1}^T \mathbf{1}_{A_t=a} \Delta(a)\right] \\ &= \sum_{a=1}^k \mathbb{E}[T_t(a)] \Delta(a) \end{aligned}$$

□

5.2 Concentration of subgaussian random variables

Definition 5.2.1

A random variable X is said to be σ -subgaussian if

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

Proposition 5.2.2

If X is σ -subgaussian then for any $\epsilon > 0$

$$\mathbb{P}(X \geq \epsilon) \leq \exp\left(\frac{-\epsilon^2}{2\sigma^2}\right)$$

Proof.

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

where the last inequality is obtained by optimizing in λ .

□

Proposition 5.2.3

If X is σ -subgaussian and Y is σ' -subgaussian conditionnaly 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.

$$\mathbb{E}[\exp \lambda X] = \sum_k \frac{\lambda^k}{k!} \mathbb{E}[X^k]$$

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 λ^1 and λ^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

$$\mathbb{E}[X] = 0 \quad \text{and} \quad \text{Var}[X] \leq \sigma^2.$$

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,

$$\begin{aligned} \mathbb{E}[\exp(\lambda(X + Y))] &\leq \mathbb{E}[\mathbb{E}[\exp(\lambda(X + Y)) | X]] \\ &\leq \mathbb{E}[\mathbb{E}[\exp(\lambda X) \exp(\lambda Y) | X]] \\ &\leq \mathbb{E}[\exp(\lambda X) \exp(\lambda^2 (\sigma')^2 / 2)] \\ &\leq \exp(\lambda^2 (\sigma^2 + (\sigma')^2) / 2) \end{aligned}$$

□

Proposition 5.2.4

If $X_i - \mu$ are iid σ -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 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

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

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

$$\begin{aligned} \mathbb{E}[T_T(a)] &\leq \min(m, n/K) + \max(n - mK, 0) \mathbb{P}(a_{mK+1} = a) \\ &\leq \min(m, n/K) + \max(n - mK, 0) \mathbb{P}\left(Q_t(a) \geq \max_{a' \neq a} Q_t(a')\right) \\ &\leq \min(m, n/K) + \max(n - mK, 0) \mathbb{P}(a_{mK+1} = a) \\ &\leq \min(m, n/K) + \max(n - mK, 0) \mathbb{P}(Q_m(a) \geq Q_m(a_*)) \\ &\leq \min(m, n/K) + \max(n - mK, 0) \mathbb{P}(Q_{mK+1}(a) - \mathbb{E}[R(a)] - (Q_{mK+1}(a_*) - \mathbb{E}[R(a_*)]) \geq \Delta(a)) \end{aligned}$$

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

$$\begin{aligned} \mathbb{E}[T_T(a)] &\leq \min(m, n/K) + \max(n - mK, 0) \mathbb{P}(Q_{mK+1}(a) \geq Q_{mK+1}(a_*)) \\ &\leq \min(m, n/K) + \max(n - mK, 0) \exp(-m\Delta(a)^2/4) \end{aligned}$$

Now

$$\begin{aligned} \mathbb{P}(a_T = a_*) &= 1 - \sum_{a \neq a_*} \mathbb{P}(a_T = a) \\ &\leq 1 - \sum_{a \neq a_*} \exp(-m\Delta(a)^2/4) \end{aligned}$$

□

5.4 ϵ -greedy strategy

Proposition 5.4.1

Let π be an ϵ_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 ϵ -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 π be an ϵ_t -greedy strategy,

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

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

Furthermore,

$$\mathbb{P}(a_* = \operatorname{argmax} Q_{T,a}) \geq 1 - \sum_t \exp(-\Sigma_T/(6k)) - \sum_{a \neq a_*} \frac{4}{\Delta(a)^2} e^{-\Delta(a)^2 \Sigma_T/(4k)}$$

If $\epsilon_t = c/t$,

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

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

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

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

Proof. By definition of π ,

$$\mathbb{P}(A_T = a) \leq \frac{\epsilon_t}{k} + (1 - \frac{\epsilon_t}{k}) \mathbb{P}(Q_T(a) \geq Q_T(a_*))$$

and

$$\mathbb{P}(Q_T(a) \geq Q_T(a_*)) \leq \mathbb{P}(Q_T(a) \geq \mu(a) + \Delta(a)/2) + \mathbb{P}(Q_T(a_*) \leq \mu(a_*) - \Delta(a)/2).$$

5 Bandits

By symmetry, it suffices to bound

$$\begin{aligned}
\mathbb{P}(Q_T(a) \geq \mu(a) + \Delta/2) &\leq \sum_{t=1}^T \mathbb{P}(T_t(a) = t, Q_T(a) \geq \mu(a) + \Delta/2) \\
&\leq \sum_{t=1}^T \mathbb{P}\left(T_T(a) = t, \frac{1}{t} \sum_{k=1}^t R_k(a) \geq \mu(a) + \Delta/2\right) \\
&\leq \sum_{t=1}^T \mathbb{P}\left(T_T(a) = t \middle| \frac{1}{t} \sum_{k=1}^t R_k(a) \geq \mu(a) + \Delta/2\right) \mathbb{P}\left(\frac{1}{t} \sum_{k=1}^t R_k(a) \geq \mu(a) + \Delta/2\right) \\
&\leq \sum_{t=1}^T \mathbb{P}\left(T_T(a) = t \middle| \frac{1}{t} \sum_{k=1}^t R_k(a) \geq \mu(a) + \Delta/2\right) e^{-\Delta^2 t/2} \\
&\leq \sum_{t=1}^{T_0} \mathbb{P}\left(T_T(a) = t \middle| \frac{1}{t} \sum_{k=1}^t R_k(a) \geq \mu(a) + \Delta/2\right) + \sum_{t=T_0+1}^T e^{-\Delta^2 t/2}
\end{aligned}$$

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

$$\begin{aligned}
&\leq \sum_{t=1}^{T_0} \mathbb{P}\left(T_T^R(a) \leq t \middle| \frac{1}{t} \sum_{k=1}^t R_k(a) \geq \mu(a) + \Delta/2\right) + \frac{2}{\Delta^2} e^{-\Delta^2 T_0/2} \\
&\leq \sum_{t=1}^{T_0} \mathbb{P}\left(T_T^R(a) \leq t\right) + \frac{2}{\Delta^2} e^{-\Delta^2 T_0/2}
\end{aligned}$$

Now the Bernstein inequality yields

$$\mathbb{P}\left(T_t^R(a) \leq \mathbb{E}[T_t^R(a)] - \lambda\right) \leq \exp\left(-\frac{\lambda^2/2}{\mathbb{V}\text{ar}[T_t^R(a)] + \lambda/2}\right)$$

with

$$\begin{aligned}
\mathbb{E}[T_t^R(a)] &= \sum_{s=1}^t \frac{\epsilon_s}{k} \\
\mathbb{V}\text{ar}[T_t^R(a)] &= \sum_{s=1}^t \frac{\epsilon_s}{k} \left(1 - \frac{\epsilon_s}{k}\right) \\
&\leq \sum_{s=1}^t \frac{\epsilon_s}{k},
\end{aligned}$$

. Choosing $T_0 = \frac{1}{2} \frac{\Sigma_T}{k} = \frac{1}{2} \sum_{s=1}^T \frac{\epsilon_s}{k} = \frac{1}{2} \mathbb{E} [T_T^R(a)] \leq \frac{1}{2} \mathbb{V}ar [T_T^R(a)]$ leads

$$\begin{aligned} \mathbb{P}(T_T^R(a) \leq T_0) &= \mathbb{P}(T_T^R(a) \leq 2T_0 - T_0) \\ &\leq \exp\left(-\frac{T_0^2/2}{\sigma^2 + T_0/2}\right) \\ &\leq \exp\left(-\frac{T_0^2/2}{T_0 + T_0/2}\right) \\ &\leq \exp(-T_0/3) \end{aligned}$$

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

$$\begin{aligned} \mathbb{P}(a = \operatorname{argmax} Q_T(a)) &\leq 2(1 - \frac{\epsilon_T}{k}) \left(\Sigma_T/(2k) \exp(-\Sigma_T/(6k)) + \frac{2}{\Delta(a)^2} e^{-\Delta(a)^2 \Sigma_T/4} \right) \\ &\leq \frac{\epsilon_T}{k} + \frac{\Sigma_T}{k} \exp(-\Sigma_T/(6k)) + \frac{4}{\Delta(a)^2} e^{-\Delta(a)^2 \Sigma_T/(4k)} \end{aligned}$$

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

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

which goes to 0 if furthermore ϵ_T tends to 0

Finally,

$$\begin{aligned} \mathbb{E}[T_T(a)] &= \sum_{t=1}^T \mathbb{P}(A_t = a) \\ &\leq \sum_{t=1}^T \left(\frac{\epsilon_t}{k} + \frac{\Sigma_t}{k} \exp(-\Sigma_t/(6k)) + \frac{4}{\Delta(a)^2} e^{-\Delta(a)^2 \Sigma_t/(4k)} \right) \end{aligned}$$

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

$$\begin{aligned} r_{T,\pi} &\leq \sum_{a \neq a_*} \left(\Delta(a) \left(c \frac{\log(T) + 1}{k} + \sum_{t=1}^T c \frac{\log(t) + 1}{k} e^{-c(\log(t)+1)/(6k)} \right) + \frac{4}{\Delta(a)} \sum_{t=1}^T e^{-\Delta(a)^2 c(\log(t)+1)/(4k)} \right) \\ &\leq \sum_{a \neq a_*} \left(\Delta(a) \left(c \frac{\log(T) + 1}{k} + C \right) + \frac{4}{\Delta(a)} C' \right) \end{aligned}$$

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

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

□

5.5 UCB strategy

Proposition 5.5.1

Assume we use a UCB strategy with a variance term $\sqrt{\frac{c \log t}{T_t(a)}}$ 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 $t \geq \max_a \frac{4c \ln t}{\Delta(a)^2}$.

Proof. By construction,

$$\begin{aligned}
T_t(a) &= \sum_{s=1}^t \mathbf{1}_{A_s=a} \\
&\leq \sum_{s=1}^t \mathbf{1}_{Q_s(a)+c_s(a)=\max Q_s(a')+c_s(a')} \\
&\leq T_0(a) + \sum_{s=T_0+1}^t \mathbf{1}_{Q_s(a)+c_s(a)=\max Q_s(a')+c_s(a'), T_s(a) \geq T_0(a)} \\
&\leq T_0(a) + \sum_{s=T_0+1}^t \mathbf{1}_{Q_s(a)+c_s(a) \geq Q_s(a_*)+c_s(a_*), T_t(a) \geq T_0(a)} \\
&\leq T_0(a) + \sum_{s=T_0+1}^t \mathbf{1}_{\max_{T_0(a) \leq s'' \leq t} \frac{1}{s''} \sum_{j=1}^{s''} R(a)_{(j)} + \sqrt{\frac{c \ln s}{s''}} \geq \min_{s' \leq t} \frac{1}{s'} \sum_{j=1}^{s'} R(a_*)_{(j)} + \sqrt{\frac{c \ln s}{s'}}} \\
&\leq T_0(a) + \sum_{s=T_0+1}^t \sum_{s'=1}^{s-1} \sum_{s''=T_0(a)}^{s-1} \mathbf{1}_{\frac{1}{s''} \sum_{j=1}^{s''} R(a)_{(j)} + \sqrt{\frac{c \ln s}{s''}} \geq \frac{1}{s'} \sum_{j=1}^{s'} R(a_*)_{(j)} + \sqrt{\frac{c \ln s}{s'}}} \\
&\leq T_0(a) + \sum_{s=T_0+1}^t \sum_{s'=1}^{s-1} \sum_{s''=T_0(a)}^{s-1} \mathbf{1}_{\mu(a_*) \leq \mu(a) + 2\sqrt{\frac{c \ln s}{s''}}} + \mathbf{1}_{\frac{1}{s''} \sum_{j=1}^{s''} R(a)_{(j)} \geq \mu(a) + \sqrt{\frac{c \ln s}{s''}}} \\
&\quad + \mathbf{1}_{\frac{1}{s'} \sum_{j=1}^{s'} R(a_*)_{(j)} \leq \mu(a_*) - \sqrt{\frac{c \ln s}{s'}}} \\
&\leq T_0(a) + \sum_{s=T_0+1}^t \sum_{s'=1}^{s-1} \sum_{s''=T_0(a)}^{s-1} \mathbf{1}_{\mu(a_*) \leq \mu(a) + 2\sqrt{\frac{c \ln s}{s''}}} + 2e^{-2c \ln s} \\
\mathbb{E}[T_t(a)] &\leq T_0(a) + \sum_{s=T_0+1}^t \sum_{s'=1}^{s-1} \sum_{s''=T_0(a)}^{s-1} \mathbf{1}_{\Delta(a) \leq 2\sqrt{\frac{2c \ln t}{s''}}} + 2s^{-2c}
\end{aligned}$$

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

$$\begin{aligned}
&\leq \frac{4c \ln t}{\Delta(a)^2} + \sum_{s=T_0+1}^t 2s^{-2c+2} \\
&\leq \frac{4c \ln t}{\Delta(a)^2} + C_c
\end{aligned}$$

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

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

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

□

6 Stochastic Approximation

6.1 Convergence of a mean

Proposition 6.1.1

Assume X_i are i.i.d. such that $\mathbb{E}[X_i|\mathcal{F}_{i-1}] = \mu$ and $\mathbb{V}\text{ar}[X_i|\mathcal{F}_{i-1}] \leq \sigma^2$, let

$$M_n = M_{n-1} + \alpha_n(X_n - M_{n-1})$$

with $1 \geq \alpha_i \geq 0$ then

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

Proof. By definition,

$$\begin{aligned} M_n &= M_{n-1} + \alpha_n(X_n - M_{n-1}) \\ &= (1 - \alpha_n)M_{n-1} + \alpha_n X_n \\ &= \prod_{i=1}^n (1 - \alpha_i) M_0 + \sum_{k=1}^n \prod_{i=k+1}^n (1 - \alpha_i) \alpha_k X_k \end{aligned}$$

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

$$\prod_{i=1}^n (1 - \alpha_i) \rightarrow 0 \quad \text{and} \quad \sum_{k=1}^n \prod_{i=k+1}^n (1 - \alpha_i)^2 \alpha_k^2 \rightarrow 0$$

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 $\sum_{i=1}^n \alpha_i \rightarrow +\infty$.

6 Stochastic Approximation

For the second one,

$$\begin{aligned}
\sum_{k=1}^n \prod_{i=k+1}^n (1 - \alpha_i)^2 \alpha_k^2 &\leq \sum_{k=1}^m \prod_{i=k+1}^n (1 - \alpha_i)^2 \alpha_k^2 + \sum_{k=m+1}^n \prod_{i=k+1}^n (1 - \alpha_i)^2 \alpha_k^2 \\
&\leq \sum_{k=1}^m \prod_{i=m}^n (1 - \alpha_i)^2 \alpha_k^2 + \max_{k \geq m+1} \alpha_k \sum_{k=m+1}^n \left(\prod_{i=k+1}^n (1 - \alpha_i) - \prod_{i=k}^n (1 - \alpha_i) \right) \\
&\leq e^{-2 \sum_{k=m}^n \alpha_i} \sum_{k=1}^m \alpha_k^2 + \max_{k \geq m+1} \alpha_k \left(1 - \prod_{i=m+1}^n (1 - \alpha_i) \right) \\
&\leq e^{-2 \sum_{k=m}^n \alpha_i} \sum_{k=1}^m \alpha_k^2 + \max_{k \geq m+1} \alpha_k
\end{aligned}$$

Choosing $m = n/2$ yields

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

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

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

which 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

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

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 6.2.3

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. □

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 $\frac{d\tilde{\theta}}{dt} = H(\tilde{\theta})$

6 Stochastic Approximation

Sketch. • Difference between θ and a solution of the ODE with $\tilde{\theta}(t_k) = \theta_k$ at t_{k+l} :

$$\begin{aligned}\theta(t_{k+l}) - \tilde{\theta}(t_{k+l}) &= \int_{t_k}^{t_{k+l}} (\theta'(u) - \tilde{\theta}'(u)) du \\ &= \sum_{k'=k}^{k+l-1} \int_{t_{k'}}^{t_{k'+1}} (H(\theta(t_k)) + \epsilon_k + \eta_k - H(\tilde{\theta}(u))) du \\ &= \sum_{k'=k}^{k+l-1} \int_{t_{k'}}^{t_{k'+1}} (H(\theta(t_k)) - H(\tilde{\theta}(u))) du \\ &\quad + \sum_{k'=k}^{k+l-1} \alpha_{k'} \epsilon_{k'} + \sum_{k'=k}^{k+l-1} \alpha_{k'} \eta_{k'}\end{aligned}$$

- The last two term are going to be small by construction. . .
- Difference between θ and a solution of the ODE with $\tilde{\theta}(t_k) = \theta_k$ at t_{k+l} :

$$\begin{aligned}\theta(t_{k+l}) - \tilde{\theta}(t_{k+l}) &= \sum_{k'=k}^{k+l-1} \int_{t_{k'}}^{t_{k'+1}} (H(\theta(t_k)) - H(\tilde{\theta}(u))) du \\ &\quad + \sum_{k'=k}^{k+l-1} \alpha_{k'} \epsilon_{k'} + \sum_{k'=k}^{k+l-1} \alpha_{k'} \eta_{k'}\end{aligned}$$

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

$$\begin{aligned}\mathbb{E} \left[\sum_{k'=k}^{k+l-1} \alpha_{k'} \epsilon_{k'} \right] &= 0 \quad \text{and} \quad \text{Var} \left[\sum_{k'=k}^{k+l-1} \alpha_{k'} \epsilon_{k'} \right] < \sigma^2 \sum_{k'=k}^{k+l-1} \alpha_{k'}^2 \rightarrow 0 \\ \left\| \sum_{k'=k}^{k+l-1} \alpha_{k'} \eta_{k'} \right\| &\leq (t_{k+l-1} - t_k) \sup_{k' \geq k} \|\eta_{k'}\|\end{aligned}$$

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:

$$\begin{aligned}\left\| \int_{t_{k'}}^{t_{k'+1}} (H(\theta(t_{k'})) - H(\tilde{\theta}(u))) du \right\| &\leq L \int_{t_{k'}}^{t_{k'+1}} \|\theta(t_{k'}) - \tilde{\theta}(u)\| du \\ &\leq L \alpha_{k'} \|\theta(t_{k'}) - \tilde{\theta}(t_{k'})\| + L \int_{t_{k'}}^{t_{k'+1}} \|\theta(t_{k'}) - \tilde{\theta}(u)\| du \\ &\leq L \alpha_{k'} \|\theta(t_{k'}) - \tilde{\theta}(t_{k'})\| + L \|H\|_{\infty} \alpha_{k'}^2\end{aligned}$$

- Combining all the results leads to

$$\begin{aligned}\|\theta(t_{k+l}) - \tilde{\theta}(t_{k+l})\| &\leq L \sum_{k'=k}^{k+l-1} \alpha_{k'} \|\theta(t_{k'}) - \tilde{\theta}(t_{k'})\| \\ &\quad + L \|H\|_{\infty} \sum_{k'=k}^{k+l-1} \alpha_{k'}^2 + \left\| \sum_{k'=k}^{k+l-1} \alpha_{k'} \epsilon_{k'} \right\| + \sum_{k'=k}^{k+l-1} \alpha_{k'} \|\eta_{k'}\|\end{aligned}$$

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 $t_{k+l} - t_k \leq \Delta$

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

□

6.3 TD(λ) and linear approximation

Proposition 6.3.1

Provided there is a unique stationary distribution μ 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(λ) algorithm with linear approximation converges with probability one. The limit $\mathbf{w}_{*,\lambda}$ is the unique solution of

$$\Pi_{\mu} \mathcal{T}^{\pi,(\lambda)} \mathbb{X} \mathbf{w}_{*,\lambda} = \mathbb{X} \mathbf{w}_{*,\lambda}.$$

Furthermore,

$$\|\mathbb{X} \mathbf{w}_{*,\lambda} - v_{\pi}\|_{2,\mu} \leq \frac{1 - \lambda\gamma}{1 - \gamma} \|\Pi_{\mu} v_{\pi} - v_{\pi}\|_{2,\mu}$$

Proof. See Tsitsiklis and Van Roy. □

Proof. Assume $\mathbf{A}$ is invertible and let $\mathbf{w}_{TD} = \mathbf{A}^{-1} \mathbf{b}$

$$\begin{aligned} \mathbb{E}[\mathbf{w}_{t+1} - \mathbf{w}_{TD} | \mathbf{w}_t] &= \mathbf{w}_t + \alpha(\mathbf{b} - \mathbf{A} \mathbf{w}_t) - \mathbf{w}_{TD} \\ &= (\text{Id} - \alpha \mathbf{A})(\mathbf{w}_t - \mathbf{w}_{TD}) \end{aligned}$$

If we prove that $\mathbf{A}$ is positive definite then $\mathbf{A}$ will be invertible and the asymptotic algorithm will converge provided α is small enough.

6 Stochastic Approximation

In the continuous task setting,

$$\begin{aligned}
\mathbf{A} &= \sum_s \mu(s) \sum_a \pi(a|s) \sum_{r,s'} p(r, s'|s, a) \mathbf{x}(s) (\mathbf{x}(s) - \gamma \mathbf{x}(s'))^t \\
&= \sum_s \mu(s) \sum_a \pi(a|s) \sum_{s'} p_\pi(s'|s) \mathbf{x}(s) (\mathbf{x}(s) - \gamma \mathbf{x}(s'))^t \\
&= \sum_s \mu(s) \mathbf{x}(s) \left(\mathbf{x}(s) - \gamma \sum_{s'} p_\pi(s'|s) \mathbf{x}(s') \right)^t \\
&= \mathbf{X}^t \mathbf{D} (\text{Id} - \gamma P_\pi) \mathbf{X}
\end{aligned}$$

where $\mathbf{D}$ is a diagonal matrix having $\mu(s)$ on the diagonal.

As P_π is a stochastic matrix, the row sums of $\mathbf{D}(\text{Id} - \gamma P_\pi)$ are non negative. Recall that μ is such that $\mu^t P_\pi = \mu^t$ and thus

$$\begin{aligned}
\mathbf{1}^t \mathbf{D} (\text{Id} - \gamma P_\pi) &= \mu^t (\text{Id} - \gamma P_\pi) \\
&= \mu^t - \gamma \mu^t P_\pi \\
&= (1 - \gamma) \mu^t > 0
\end{aligned}$$

□