Markov Decision Problems

MDP $M = (S, A, T, R, \gamma)$

- S → States
- A → Actions
- T → Transition Function
- R → Reward Function
- γ → Discount Factor

* $R(s,a,s')$ = Reward for $s \xrightarrow{a} s'$ $\quad R \in [-R_{max}, R_{max}]$ $\quad R_{max} \geq 0$

$T(s,a,s')$ = Probability of reaching s' by starting at s and taking action a.

* Thus $T(s,a,\cdot)$ = Probability distribution over S.

Agent Environment Interaction:

At $t=0$,

$s^1 \sim T(s^0, a^0, \cdot)$

$r^0 = R(s^0, a^0, s^1)$

$$\text{Agent} \underset{r^t,\, s^{t+1}}{\overset{a^t}{\rightleftarrows}} \text{Environment}$$

Resulting Trajectory = $s^0\, a^0\, r^0\, s^1\, a^1\, r^1\, s^2 \ldots$

Assume a^t is picked based on s^t alone

Policy $\pi : S \to A$

$\pi \to$ Markovian, deterministic and stationary.

Policy:

$\Pi \to$ Denote the set of all policies.

$|\Pi| = k^n$ $\quad n \to$ states, $k \to$ actions

Q. Which $\pi \in \Pi$ is a good policy?

Value function "Larger is better"

$V^\pi(s)$: Value of State S under Policy π

For $s \in S$, $V^\pi(s) \stackrel{def}{=} E_\pi[r^0 + \gamma r^1 + \gamma^2 r^2 + \gamma^3 r^3 + \ldots \mid s^0 = s]$

$\gamma = [0,1) \Rightarrow$ Discount Factor

Large $\gamma \Rightarrow$ Farther lookahead

★★ Every MDP is guaranteed to have an optimal policy π^* such that $\forall \pi \in \Pi, \forall s \in S: V^{\pi^*}(s) \geq V^\pi(s)$

MDP Planning Problem:

Given $M = (S, A, T, R, \gamma)$ find a policy π^* from the set of all policies Π such that $\forall s \in S, \forall \pi \in \Pi: V^{\pi^*}(s) \geq V^\pi(s)$

* Every MDP is guaranteed to have a deterministic, markovian, stationary optimal policy
* An MDP can have more than one optimal policy.

⇒ Value function of every optimal policy is the same unique V^*.

Alternative Formulations:

→ $R(s,a,s') \Rightarrow$ Random variable bounded in $[-R_{max}, R_{max}]$

→ $R(s,a,s') \to R(s,a)$, $R(s')$

→ Combined T and R.

→ Minimum "COST" instead of maximising "Reward"

★ **Finite Horizon Reward**: $E_\pi[r^0 + r^1 + r^2 \ldots r^{T-1} \mid s^0 = s]$

★ **Average Reward**: $E_\pi[\lim_{m \to \infty} (r^0 + r^1 + \ldots r^{m-1})/m \mid s^0 = s]$

Episodic Tasks:

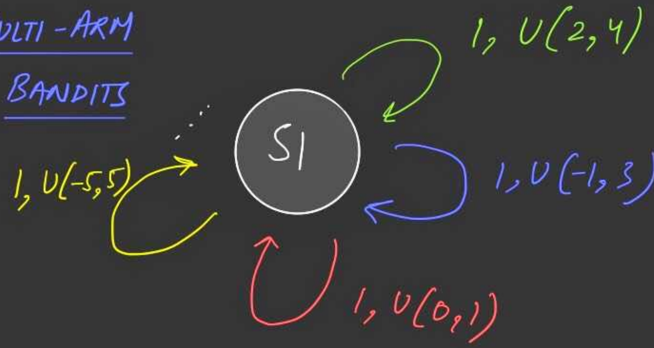
→ Have special sink / terminal state S_T from which there are no outgoing transitions or rewards

* From every non-terminal state and for every policy there is a non-zero probability of reaching terminal state in finite number of steps

Total reward: $V^\pi(s) \stackrel{def}{=} E_\pi[r^0 + r^1 + r^2 + \ldots \mid s^0 = s]$ → Only be used on episodic tasks

POLICY EVALUATION :

MULTI-ARM BANDITS



$$V^{\pi}(s) \stackrel{def}{=} \mathbb{E}_{\pi}\left[r^0 + \gamma r^1 + \gamma^2 r^2 \dots \mid s^0 = s\right]$$

$$= \sum_{s' \in S} T(s, \pi(s), s')\, \mathbb{E}_{\pi}\left[r^0 + \gamma r^1 + \dots \mid s^0 = s, s^1 = s'\right]$$

$$= \sum_{s' \in S} T(s, \pi(s), s') R(s, \pi(s), s') + \gamma \sum_{s' \in S} T(s, \pi(s), s')\, \mathbb{E}_{\pi}\left[r^1 + \gamma r^2 + \dots \mid s^1 = s'\right]$$

$$= \sum_{s' \in S} T(s, \pi(s), s') \left\{ R(s, \pi(s), s') + \gamma V^{\pi}(s') \right\} \Rightarrow \text{BELLMAN'S EQUATIONS}$$

n-equations, n-unknowns — $V^{\pi}(s_1), V^{\pi}(s_2) \dots V^{\pi}(s_2)$

* If $\gamma < 1 \Rightarrow$ UNIQUE Solution

POLICY EVALUATION : Computing V^{π} for a policy π

ACTION VALUE FUNCTION :

$$Q^{\pi}(s,a) \stackrel{def}{=} \mathbb{E}\left[r^0 + \gamma r^1 + \gamma^2 r^2 \dots \mid s^0 = s; a^0 = a; a^t = \pi(s^t) \text{ for } t \geq 1\right]$$

$$Q^{\pi}(s,a) = \sum_{s' \in S} T(s, a, s') \left\{ R(s, a, s') + \gamma V^{\pi}(s') \right\}$$

* All optimal policies have the same action value function Q^*

ALGORITHMS : ① Bellman Optimality ② Value Iteration ③ Linear Programming Formulation

CONTRACTION MAPPING :

A mapping $T: X \to X$ is called a contraction mapping with a contraction factor L if $\forall u, v \in X$,

$$\|Tv - Tu\| \leq L \|v - u\|$$

Fixed point : $Tx^* = x^*$

BANACH'S FIXED POINT THEOREM :

Let $(X, \|\cdot\|)$ be a Banach space and let $T: X \to X$ be a contraction mapping with $L \in [0,1)$ then:

① T has a unique fixed point $x^* \in X$

② For $x \in X$, $m \geq 0$: $\|T^m x - x^*\| \leq L^m \|x - x^*\|$

BELLMAN OPTIMALITY OPERATOR

$B^* : (S \to \mathbb{R}) \to (S \to \mathbb{R})$ for an MDP (S, A, T, R, γ) is defined as :

For $F : S \to \mathbb{R}$ and $s \in S$:

$$(B^*(F))(s) \stackrel{def}{=} \max_{a \in A} \sum_{s' \in S} T(s,a,s') \{R(s,a,s') + \gamma F(s')\}$$

Since $S = \{s_1, s_2 \dots s_n\}$, we may equivalently view B^* as a mapping from $\mathbb{R}^n \to \mathbb{R}^n$

* max norm $\|\cdot\|_\infty$ of $F = (f_1, f_2, \dots, f_n) \in \mathbb{R}^n \Rightarrow \|F\|_\infty = \max\{|f_1|, |f_2|, \dots, |f_n|\}$

Already established $\longrightarrow$ $(\mathbb{R}^n, \|\cdot\|_\infty)$ is a Banach Space.

CLAIM : B^* is a contraction mapping in the $(\mathbb{R}^n, \|\cdot\|_\infty)$ Banach space with contraction factor γ.

PRE REQUISITE : $|\max_a f(a) - \max_a g(a)| \leqslant \max_a |f(a) - g(a)|$

PROOF : $\|B^*(F) - B^*(G)\|_\infty \leqslant \max_{s \in S} |(B^*(F))(s) - (B^*(G))(s)|$

$$= \max_{s \in S} \left| \max_{a \in A} \sum T(s,a,s')\{R(s,a,s') + \gamma F(s')\} - \max_{a \in A} \sum_{s' \in S} T(s,a,s')\{R(s,a,s') + \gamma G(s')\} \right|$$

$$= \gamma \max_{s \in S} \max_{a \in A} \left| \sum_{s' \in S} T(s,a,s')\{F(s') - G(s')\} \right|$$

$$= \gamma \max_{s \in S} \max_{a \in A} \sum_{s' \in S} T(s,a,s') \|F - G\|_\infty = \gamma \|F - G\|_\infty \Rightarrow \text{Fixed point for } B^*$$

BELLMAN OPTIMALITY EQUATIONS $\longleftarrow$

$$\boxed{V^*(s) = \max_{a \in A} \sum_{s' \in S} T(s,a,s')\{R(s,a,s') + \gamma V^*(s')\}}$$

* Not linear like Bellman Equations

$\Rightarrow$ V^* is the value function for every policy $\pi^* : S \to A$ that satisfies for all $s \in S$:

$$\pi^*(s) = \operatorname{argmax}_{a \in A} \sum_{s' \in S} T(s,a,s') \{R(s,a,s') + \gamma V^*(s')\}$$

① VALUE ITERATION

$$V_0 \xrightarrow{B^*} V_1 \xrightarrow{B^*} V_2 \xrightarrow{B^*} \cdots$$

* Stop when $V_t \approx V_{t+1}$ (upto machine precision)

* $V^* \rightarrow Q^*$ $\quad Q^*(s,a) = \sum_{s' \in S} T(s,a,s') \{ R(s,a,s') + \gamma V^*(s') \}$

* $Q^* \rightarrow \pi^*$ $\quad \pi^*(s) = \underset{a \in A}{\operatorname{argmax}} \; Q^*(s,a)$

* $\pi^* \rightarrow V^*$ Solve Bellman Equations for π^*

② LINEAR PROGRAMMING FORMULATION

* Optimise linear function of variables
* Subject to linear constraints

Create n variables $V(s_1), V(s_2) \dots V(s_n) \longrightarrow V^*$ will be the solution

Linear constraint for max $\longrightarrow$ $V(s) \geq \sum_{s' \in S} T(s,a,s') \{ R(s,a,s') + \gamma V(s') \}$ for $s \in S$, $a \in A$

Vector Comparison

$X \succeq Y \Leftrightarrow \forall s \in S : X(s) \geq Y(s)$

$X \succ Y \Leftrightarrow X \succeq Y$ and $\exists s \in S : X(s) > Y(s)$

* Sometimes vectors are incomparable.

B^* Preserves $\succeq$

TO PROVE: $(B^*(X))(s) - (B^*(Y))(s) \geq 0$

$$\begin{aligned}
&(B^*(X))(s) - (B^*(Y))(s) \\
&= \max_{a \in A} \sum_{s' \in S} T(s,a,s') \{ R(s,a,s') + \gamma X(s') \} \\
&\quad - \max_{a \in A} \sum_{s' \in S} T(s,a,s') \{ R(s,a,s') + \gamma Y(s') \} \\
&\geq \gamma \min_{a \in A} \sum_{s' \in S} T(s,a,s') \{ X(s') - Y(s') \} \geq 0
\end{aligned}$$

* $\max_a f(a) - \max_a g(a) \geq \min_a (f(a) - g(a))$

* $V \succeq B^*(V)$

$$V \succeq \lim_{l \to \infty} (B^*)^l (V) = V^*$$

$$\Rightarrow \sum_{s \in S} V(s) \geq \sum_{s \in S} V^*(s)$$

Goal: Maximise $\left(-\sum_{s\in S} V(s)\right)$

Subject to $V(s) \geq \sum_{s'\in S} T(s,a,s')\{R(s,a,s') + \gamma V(s')\} \quad \forall s\in S,\ a\in A$

LP $\rightarrow$ n variables, nk constraint

POLICY IMPROVEMENT:

Given π, Pick one or more improvable states, and in them, switch to an arbitrary improvement action. Resulting state $\rightarrow \pi'$

① For $\pi \in \Pi$, $s \in S$

IMPROVING ACTIONS $\Leftarrow IA(\pi, s) \stackrel{def}{=} \{a \in A : Q^{\pi}(s,a) > V^{\pi}(s)\}$

② For $\pi \in \Pi$

IMPROVABLE STATES $\Leftarrow IS(\pi) \stackrel{def}{=} \{s \in S : |IA(\pi,s)| \geq 1\}$

$\pi' \in \Pi$ is obtained by Policy Improvement $\rightarrow$
- $\forall s \in S : \pi'(s) = \pi(s)$ or $\pi'(s) \in IA(\pi, s)$
- $\exists s \in S : \pi'(s) \in IA(\pi, s)$

POLICY IMPROVEMENT THEOREM

(1) If $IS(\pi) = \phi$, then π is optimal

(2) If π' is obtained by policy improvement on π, then $\pi' \succ \pi$*

* If $IS(\pi) \neq \phi$ then $\exists\ \pi' \in \Pi$ such that $\pi' \succ \pi$
* But Π has finite policies (k^n)
* $\Rightarrow \exists\ \pi^*$ such that $IS(\pi^*) = \phi$
* $IS(\pi^*) = \phi \Leftrightarrow B^*(V^{\pi^*}) = V^{\pi^*}$

Convention $\rightarrow V^*$ for V^{π^*}

POLICY ITERATION ALGORITHM

$\pi \leftarrow$ Arbitrary Policy

While π has improvable states

$\quad \pi' \leftarrow$ Policy improvement (π)

$\quad \pi \leftarrow \pi'$

Return π

*SWITCHING STRATEGY

Path taken and iterations depend on which improved state is chosen.

① Random policy iteration

② Simple Policy iteration

★ HOWARD'S POLICY ITERATION:

$\Rightarrow$ Greedy $\rightarrow$ Switch all improvable states

★ RANDOM POLICY ITERATION:

$\Rightarrow$ Switch a non-empty subset of improvable states chosen uniformly at random. $\rightarrow 2^n - 1$ subsets

PROOF OF POLICY IMPROVEMENT THEOREM

BELLMAN OPERATOR $B^{\pi}: (S \to R) \to (S \to R)$ * It is a contraction mapping.

$$(B^{\pi}(X))(s) \stackrel{def}{=} \sum_{s' \in S} T(s,\pi(s),s')\left(R(s,\pi(s),s') + \gamma X(s')\right)$$

For $X: S \to R$: $\lim_{l \to \infty} (B^{\pi})^{l}(X) = V^{\pi}$

For $X: S \to R, Y: S \to R$: $X \succeq Y \Rightarrow B^{\pi}(X) \succeq B^{\pi}(Y)$

Observe that : $B^{\pi'}(V^{\pi})(s) = Q^{\pi}(s, \pi'(s))$

PROOF : $IS(\pi) = \phi$

$\Rightarrow \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi})$

$\Rightarrow \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^{2}(V^{\pi}) \dots \succeq \lim_{l \to \infty} (B^{\pi'})^{l}(V^{\pi})$ $V^{\pi'}$

$IS(\pi) \neq \phi$

$\Rightarrow B^{\pi'}(V^{\pi}) \succ V^{\pi}$

$\Rightarrow (B^{\pi'})^{2}(V^{\pi}) \succeq B^{\pi'}(V^{\pi}) \succ V^{\pi}$

$\Rightarrow V^{\pi'} \succ V^{\pi}$

COMPUTATIONAL COMPLEXITY :

RUNNING TIME BOUNDS
- Value Iteration $\Rightarrow$ Upper bound : $poly\left(n, k, B, \frac{1}{1-\gamma}\right)$
- Linear Programming $\Rightarrow$ $poly(n, k, B)$
  - $poly(n,k) \cdot \exp\left(O\left(\sqrt{n \log(n)}\right)\right)$ Expected

POLICY ITERATION

	$k=2$	General k
Deterministic	$O\left(\frac{2^n}{n}\right)$	$O\left(\frac{k^n}{n}\right)$
Randomised	1.7172^n	$O\left(\frac{k}{2}\right)^n$

$\rightarrow$ UPPER BOUND

REINFORCEMENT LEARNING PROBLEM:

Agent does not know Transition Probability and sometimes reward function

* Can the agent eventually take optimal actions?

$h^t = (s^0, a^0, r^0, s^1 \dots s^t) \Rightarrow$ History.

* Learning Algorithm L is a mapping from set of all histories to set of probability distribution over arms.

LEARNING PROBLEM

* CONTROL PROBLEM

$$\lim_{T\to\infty} \frac{1}{T}\left(\sum_{t=0}^{T-1} \mathbb{P}\{a^t \sim L(h^t) \text{ is an optimal action for } s^t\}\right) = 1$$

$\Downarrow$

We want to be sublinear.

PREDICTION PROBLEM

* We are give π that agent follows. AIM: To estimate V^π.

PROBLEM: Can we construct L such that $\lim_{t\to\infty} \hat{V}^t = V^\pi$?

ASSUMPTION 1: IRREDUCIBILITY

If there is a directed path from s to s' for every $s, s' \in S$,

↳ Comes from non-zero-probability transition under π.

then M is irreducible under π. M is irreducible if it is irreducible under ALL π.

ASSUMPTION 2: APERIODIC

$X(s,t) \rightarrow$ set of all states possible after time t following policy π.

$Y(s) \rightarrow$ set of all t such that $s \in X(s,t)$

$p(s) \rightarrow$ GCD $(Y(s))$

M is aperiodic under π if for all $s \in S$: $p(s) = 1$

If M is aperiodic under all $\pi \in \Pi$, then M is aperiodic.

ERGODICITY: Mdp which is APERIODIC and IRREDUCIBLE.

* Every policy induces a unique steady state distribution $\mu^\pi : S \rightarrow (0,1)$, subject to $\sum_{s\in S} \mu^\pi(s) = 1$, which is independent of start. → For Ergodic MDP

$$\mu^\pi(s) = \lim_{t\to\infty} p(s,t)$$

MODEL BASED APPROACH :

Estimate of MDP $\hat{T} \longrightarrow T$
$\hat{R} \longrightarrow R$

* At convergence, acting optimally for MDP $(S, A, \hat{T}, \hat{R}, \gamma)$ must be optimal for the original MDP (S, A, T, R, γ)

* We must visit every state-action pair infinitely often

Because reward is stochastic.

* Algorithm discussed in slides. ⇒ Takes sub-linear number of sub-optimal steps.

→ Needs Irreducibility, not Aperiodicity.

MODEL BASED

→ uses $\Theta(|S|^2|A|)$ memory.

PREDICTION

MONTE CARLO methods estimate based on sample averages.

NOTATION :
$s \rightarrow$ state
$i \rightarrow$ Episode Number
$j \rightarrow$ occurrence in an episode.

$1(s,i,j) \longrightarrow$ If s is visited at least j times on episode i

$G(s,i,j) \longrightarrow$ Discounted long term reward starting from j^{th} visit of s on episode i.

FIRST VISIT MONTE CARLO

$$\hat{V}^{T}_{\text{First-visit}}(s) = \frac{\sum_{i=1}^{T} G(s,i,1)}{\sum_{i=1}^{T} 1(s,i,1)}$$

EVERY VISIT MONTE CARLO

$$\hat{V}^{T}_{\text{Every-visit}}(s) = \frac{\sum_{i=1}^{T} \sum_{j=1}^{\infty} G(s,i,j)}{\sum_{i=1}^{T} \sum_{j=1}^{\infty} 1(s,i,j)}$$

$$\lim_{T \to \infty} \hat{V}^{T}_{\text{First-visit}} = V^{\pi}$$

$$\lim_{T \to \infty} \hat{V}^{T}_{\text{Every-visit}} = V^{\pi}$$

⇒ More careful analysis

$$\lim_{T \to \infty} \hat{V}^{T}_{\text{second-visit}} = V^{\pi}$$

$$\lim_{T \to \infty} \hat{V}^{T}_{\text{Last-visit}} \neq V^{\pi}$$

First Visit Monte Carlo

$$\hat{V}^t(s) = \frac{1}{t}\sum_{i=1}^{t} G(s,i,1)$$

Online Implementation

$$\hat{V}^t(s) = \frac{1}{t}\sum_{i=1}^{t} G(s,t,i)$$

$$= \frac{1}{t}\left(\sum_{i=1}^{t-1} G(s,t,i) + G(s,t,1)\right)$$

$$= \frac{1}{t}\left((t-1)\hat{V}^{t-1}(s) + G(s,t,1)\right)$$

$$= (1-\alpha_t)\hat{V}^{t-1}(s) + \alpha_t G(s,t,1) \qquad \alpha_t = \frac{1}{t}$$

Stochastic Approximation

Let the sequence $(\alpha_t)_{t\geq 1}$ satisfy:

① $\sum_{t=1}^{\infty} \alpha_t = \infty$

② $\sum_{t=1}^{\infty} \alpha_t^2 \leq \infty$

For $t \geq 1$:

$$\hat{V}^t(s) \leftarrow (1-\alpha_t)\hat{V}^{t-1}(s) + \alpha_t G(s,t,1)$$

$\alpha_t \rightarrow$ learning rate.

Then $\lim_{t\to\infty} \hat{V}^t(s) = V^\pi(s)$

Bootstrapping

Replacing M by older estimates of the value function of state encountered in the episode.

M ↓ Monte Carlo Estimate

Temporal Difference Learning

$$\hat{V}^{t+1}(s^t) \leftarrow \hat{V}^t(s^t) + \alpha_{t+1}\{r^t + \gamma\hat{V}^t(s^{t+1}) - \hat{V}^t(s^t)\}$$

$r^t + \gamma\hat{V}^t(s^{t+1}) - \hat{V}^t(s^t) \Rightarrow$ TEMPORAL DIFFERENCE PREDICTION ERROR.

$\hat{V}^t(s_T) = 0$ For episodic Tasks

$\uparrow$ Terminal States.

Error

First Visit Monte Carlo

$$Error_{First}(V,s) \overset{def}{=} \sum_{i=1}^{T} 1(s,i,1)(V(s) - G(s,i,1))^2$$

Every Visit Monte Carlo

$$Error_{Every}(V,s) \overset{def}{=} \sum_{i=1}^{T}\sum_{j=1}^{\infty} 1(s,i,j)\left(V(s) - G(s,i,j)\right)^2$$

* For TD estimate for a finite set of episodes, V^t depends on $\hat{V}_0$. Therefore to remove that dependence, convert finite to infinite episodes by repeating the episodes.

Batch TD(0) Estimate:

$$\hat{V}^T_{Batch-TD(0)} = V^\pi \text{ on MLE.}$$

3 Control Algorithms

From state s^t, $a^t \sim \pi^t(s^t)$

Update $\hat{Q}_t$ after observing s^t, a^t, r^t, s^{t+1}

$$\hat{Q}^{t+1}(s^t,a^t) \leftarrow \hat{Q}^t(s^t,a^t) + \alpha_{t+1}\{Target - \hat{Q}^t(s_t,a^t)\}$$

Q-LEARNING $\Rightarrow r^t + \gamma\max_{a\in A}\hat{Q}^t(s^{t+1},a)$ * $\longrightarrow$ OFF POLICY

SARSA $\Rightarrow r^t + \gamma\hat{Q}^t(s^{t+1},a^{t+1})$

EXPECTED SARSA $\Rightarrow r^t + \gamma\sum_{a\in A}\pi^t(s^{t+1},a)\hat{Q}^t(s^{t+1},a)$ $\longrightarrow$ ON POLICY

$\lim_{t\to\infty}\hat{Q}^t = Q^* \longleftarrow \pi^t$ is ϵ_t-greedy w.r.t. $\hat{Q}^t$.

For $\pi^t = \pi \Rightarrow$ Time invariant

Q-learning $\longrightarrow Q^*$

SARSA, EXPECTED SARSA $\longrightarrow Q^\pi$

MULTI-STEP RETURNS

$$s_1 \longrightarrow s_2 \longrightarrow s_3$$

$$V^{new}(s_1) \leftarrow V^{old}(s_1) + \alpha\{\underbrace{2+\gamma+\gamma^2 V^{old}(s_3)}_{\text{2-step return}} - V^{old}(s_1)\}$$

n-STEP RETURN

$$G_{t:t+n} \stackrel{def}{=} r^t + \gamma r^{t+1} + \gamma^2 r^{t+2} \dots \gamma^{n-1} r^{t+n-1} + \gamma^n V^{t+n-1}(s^{t+n})$$

For episodic tasks $\longrightarrow$ $G_{t:t+n} = G_{t:t+n'}$

$n' \longrightarrow$ END.

n-STEP TD

$$V^{t+n}(s^t) \leftarrow V^{t+n-1}(s^t) + \alpha\{G_{t:t+n} - V^{t+n-1}(s^t)\}$$

For $n \geq 1$

$$\lim_{t\to\infty} V^t = V^\pi$$

COMBINING RETURNS :

$\rightarrow$ Convex combinations of $G_{t:t+k}$ where $k \geq 1$

$\downarrow$

weighted average where $w_i \geq 0$.

λ-RETURN

$$G_t^\lambda \stackrel{def}{=} (1-\lambda) \sum_{n=1}^{T-t-1} \lambda^{n-1} G_{t:t+n} + \lambda^{T-t-1} G_{t:T}$$

Notice that :

$G_t^0 = G_{t:t+1} \longrightarrow$ Full Bootstrapping.

$G_t^1 = G_{t:\infty} \longrightarrow$ Monte Carlo Estimate

LINEAR ARCHITECTURE

$$\hat{V}(w,s) = w \cdot x(s) \qquad X : S \longrightarrow R^d$$

$$w^* = \underset{w \in R^d}{\text{argmin}}\ MSVE(w) \qquad MSVE(w) \stackrel{def}{=} \frac{1}{2} \sum_{s\in S} \mu^\pi(s) \{V^\pi(s) - \hat{V}(w,s)\}^2$$

$$MSVE(w_\lambda^\infty) \leq \left(\frac{1-\gamma\lambda}{1-\gamma}\right) MSVE(w^*)$$

TILE CODING

* A tiling partitions x into equal width regions called tiles
  Multiple tiles are created with an offset from previous one.

REPRESENTING $\hat{Q}$

$$\hat{Q}(s,a) = \sum_{j=1}^{d} T_{aj}(x_j(s))$$

★ CONJUNCTION OF FEATURES

To represent more complex functions