Reinforcement Learning Algorithms

A. LAZARIC (*SequeL Team @INRIA-Lille*)
ENS Cachan - Master 2 MVA

SequeL – INRIA Lille

How to *solve incrementally* an RL problem

Reinforcement Learning Algorithms

Reinforcement Learning Algorithms

Tools

Policy Evaluation

Policy Learning

Notice

From now on we often work on the ***episodic discounted*** setting.

Most results smoothly extend to other settings.

Notice

From now on we often work on the ***episodic discounted*** setting.

Most results smoothly extend to other settings.

The value functions can be represented ***exactly*** (no approximation error).

In This Lecture

- Dynamic programming algorithms require an *explicit* definition of
  - transition probabilities $p(\cdot|x, a)$
  - reward function $r(x, a)$

In This Lecture

- Dynamic programming algorithms require an *explicit* definition of
  - transition probabilities $p(\cdot|x, a)$
  - reward function $r(x, a)$
- This knowledge is often *unavailable* (i.e., wind intensity, human-computer-interaction).

In This Lecture

- Dynamic programming algorithms require an *explicit* definition of
  - transition probabilities $p(\cdot|x, a)$
  - reward function $r(x, a)$
- This knowledge is often *unavailable* (i.e., wind intensity, human-computer-interaction).
- *Can we relax this assumption?*

In This Lecture

- *Learning with generative model.* A *black-box simulator f* of the environment is available. Given (x, a),

$$f(x,a) = \{y, r\} \text{ with } y \sim p(\cdot|x,a), r = r(x,a).$$

In This Lecture

- *Learning with generative model.* A *black-box simulator f* of the environment is available. Given (x, a),

$$f(x, a) = \{y, r\} \text{ with } y \sim p(\cdot|x, a), r = r(x, a).$$

- *Episodic learning.* Multiple *trajectories* can be repeatedly generated from the same state x and terminating when a *reset* condition is achieved:

$$(x_0^i = x, x_1^i, \ldots, x_{T_i}^i)_{i=1}^n.$$

In This Lecture

- *Learning with generative model.* A *black-box simulator f* of the environment is available. Given (x, a),

$$f(x, a) = \{y, r\} \text{ with } y \sim p(\cdot|x, a), r = r(x, a).$$

- *Episodic learning.* Multiple *trajectories* can be repeatedly generated from the same state x and terminating when a *reset* condition is achieved:

$$(x_0^i = x, x_1^i, \ldots, x_{T_i}^i)_{i=1}^n.$$

- *Online learning.* At each time t the agent is at state x_t, it takes action a_t, it observes a transition to state x_{t+1}, and it receives a reward r_t. We *assume* that $x_{t+1} \sim p(\cdot|x_t, a_t)$ and $r_t = r(x_t, a_t)$ (i.e., MDP assumption).

How to *solve incrementally* an RL problem

Reinforcement Learning Algorithms

How to *solve incrementally* an RL problem

Reinforcement Learning Algorithms

Tools

Policy Evaluation

Policy Learning

Concentration Inequalities

Let X be a random variable and $\{X_n\}_{n\in\mathbb{N}}$ a sequence of r.v.

Concentration Inequalities

Let X be a random variable and $\{X_n\}_{n\in\mathbb{N}}$ a sequence of r.v.

- $\{X_n\}$ converges to X *almost surely*, $X_n \xrightarrow{a.s.} X$, if

$$\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1,$$

Concentration Inequalities

Let X be a random variable and $\{X_n\}_{n\in\mathbb{N}}$ a sequence of r.v.

- $\{X_n\}$ converges to X *almost surely*, $X_n \xrightarrow{a.s.} X$, if

$$\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1,$$

- $\{X_n\}$ converges to X *in probability*, $X_n \xrightarrow{P} X$, if for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}[|X_n - X| > \epsilon] = 0,$$

Concentration Inequalities

Let X be a random variable and $\{X_n\}_{n\in\mathbb{N}}$ a sequence of r.v.

- $\{X_n\}$ converges to X *almost surely*, $X_n \xrightarrow{a.s.} X$, if

$$\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1,$$

- $\{X_n\}$ converges to X *in probability*, $X_n \xrightarrow{P} X$, if for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}[|X_n - X| > \epsilon] = 0,$$

- $\{X_n\}$ converges to X *in law* (or in distribution), $X_n \xrightarrow{D} X$, if for any bounded continuous function f

$$\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)].$$

Concentration Inequalities

Let X be a random variable and $\{X_n\}_{n\in\mathbb{N}}$ a sequence of r.v.

- $\{X_n\}$ converges to X *almost surely*, $X_n \xrightarrow{a.s.} X$, if

$$\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1,$$

- $\{X_n\}$ converges to X *in probability*, $X_n \xrightarrow{P} X$, if for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}[|X_n - X| > \epsilon] = 0,$$

- $\{X_n\}$ converges to X *in law* (or in distribution), $X_n \xrightarrow{D} X$, if for any bounded continuous function f

$$\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)].$$

Remark: $X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{P} X \implies X_n \xrightarrow{D} X.$

Concentration Inequalities

Proposition (Markov Inequality)

Let X be a *positive* random variable. Then for any $a > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}X}{a}.$$

Concentration Inequalities

Proposition (Markov Inequality)

Let X be a *positive* random variable. Then for any $a > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}X}{a}.$$

Proof.

$$\mathbb{P}(X \geq a) = \mathbb{E}[\mathbb{I}\{X \geq a\}] = \mathbb{E}[\mathbb{I}\{X/a \geq 1\}] \leq \mathbb{E}[X/a]$$

■

Concentration Inequalities

Proposition (Hoeffding Inequality)

Let X be a *centered* random variable bounded in $[a, b]$. Then for any $s \in \mathbb{R}$,

$$\mathbb{E}[e^{sX}] \leq e^{s^2(b-a)^2/8}.$$

Concentration Inequalities

Proof.

From *convexity* of the exponential function, for any $a \leq x \leq b$,

$$e^{sx} \leq \frac{x-a}{b-a} e^{sb} + \frac{b-x}{b-a} e^{sa}.$$

Let $p = -a/(b-a)$ then (recall that $\mathbb{E}[X] = 0$)

$$\begin{aligned} \mathbb{E}[e^{sx}] &\leq \frac{b}{b-a} e^{sa} - \frac{a}{b-a} e^{sb} \\ &= (1 - p + pe^{s(b-a)}) e^{-ps(b-a)} = e^{\phi(u)} \end{aligned}$$

with $u = s(b-a)$ and $\phi(u) = -pu + \log(1 - p + pe^{u})$ whose derivative is

$$\phi'(u) = -p + \frac{p}{p + (1-p)e^{-u}},$$

and $\phi(0) = \phi'(0) = 0$ and $\phi''(u) = \frac{p(1-p)e^{-u}}{(p+(1-p)e^{-u})^2} \leq 1/4$.

Thus from *Taylor's theorem*, the exists a $\theta \in [0, u]$ such that

$$\phi(\theta) = \phi(0) + \theta\phi'(0) + \frac{u^2}{2}\phi''(\theta) \leq \frac{u^2}{8} = \frac{s^2(b-a)^2}{8}.$$

Concentration Inequalities

Proposition (Chernoff-Hoeffding Inequality)

Let $X_i \in [a_i, b_i]$ be n *independent* r.v. with mean $\mu_i = \mathbb{E}X_i$. Then

$$\mathbb{P}\Big[\Big|\sum_{i=1}^{n}(X_i - \mu_i)\Big| \geq \epsilon\Big] \leq 2\exp\Big(-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\Big).$$

Concentration Inequalities

Proof.

$$
\begin{aligned}
\mathbb{P}\Big(\sum_{i=1}^n X_i - \mu_i \geq \epsilon\Big) &= \mathbb{P}(e^{s\sum_{i=1}^n X_i - \mu_i} \geq e^{s\epsilon}) \\
&\leq e^{-s\epsilon}\mathbb{E}[e^{s\sum_{i=1}^n X_i - \mu_i}], \quad \text{Markov inequality} \\
&= e^{-s\epsilon}\prod_{i=1}^n \mathbb{E}[e^{s(X_i-\mu_i)}], \quad \text{independent random variables} \\
&\leq e^{-s\epsilon}\prod_{i=1}^n e^{s^2(b_i-a_i)^2/8}, \quad \text{Hoeffding inequality} \\
&= e^{-s\epsilon+s^2\sum_{i=1}^n (b_i-a_i)^2/8}
\end{aligned}
$$

If we choose $s = 4\epsilon/\sum_{i=1}^n (b_i - a_i)^2$, the result follows.
Similar arguments hold for $\mathbb{P}\big(\sum_{i=1}^n X_i - \mu_i \leq -\epsilon\big)$.

Monte-Carlo Approximation of a Mean

Definition

Let X be a random variable with mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \mathbb{V}[X]$ and $x_n \sim X$ be n i.i.d. realizations of X. The Monte-Carlo approximation of the mean (i.e., the empirical mean) built on n i.i.d. realizations is defined as

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Monte-Carlo Approximation of a Mean

- *Unbiased estimator*: Then $\mathbb{E}[\mu_n] = \mu$ (and $\mathbb{V}[\mu_n] = \frac{\mathbb{V}[X]}{n}$)

Monte-Carlo Approximation of a Mean

- *Unbiased estimator*: Then $\mathbb{E}[\mu_n] = \mu$ (and $\mathbb{V}[\mu_n] = \frac{\mathbb{V}[X]}{n}$)
- *Weak law of large numbers:* $\mu_n \xrightarrow{P} \mu$.

Monte-Carlo Approximation of a Mean

- *Unbiased estimator*: Then $\mathbb{E}[\mu_n] = \mu$ (and $\mathbb{V}[\mu_n] = \frac{\mathbb{V}[X]}{n}$)
- *Weak law of large numbers:* $\mu_n \xrightarrow{P} \mu$.
- *Strong law of large numbers*: $\mu_n \xrightarrow{a.s.} \mu$.

Monte-Carlo Approximation of a Mean

- *Unbiased estimator*: Then $\mathbb{E}[\mu_n] = \mu$ (and $\mathbb{V}[\mu_n] = \frac{\mathbb{V}[X]}{n}$)
- *Weak law of large numbers:* $\mu_n \xrightarrow{P} \mu$.
- *Strong law of large numbers*: $\mu_n \xrightarrow{a.s.} \mu$.
- *Central limit theorem (CLT)*: $\sqrt{n}(\mu_n - \mu) \xrightarrow{D} \mathcal{N}(0, \mathbb{V}[X])$.

Monte-Carlo Approximation of a Mean

- *Unbiased estimator*: Then $\mathbb{E}[\mu_n] = \mu$ (and $\mathbb{V}[\mu_n] = \frac{\mathbb{V}[X]}{n}$)
- *Weak law of large numbers:* $\mu_n \xrightarrow{P} \mu$.
- *Strong law of large numbers*: $\mu_n \xrightarrow{a.s.} \mu$.
- *Central limit theorem (CLT)*: $\sqrt{n}(\mu_n - \mu) \xrightarrow{D} \mathcal{N}(0, \mathbb{V}[X])$.
- *Finite sample guarantee*:

$$\mathbb{P}\Bigg[\underbrace{\Big| \frac{1}{n} \sum_{t=1}^{n} X_t - \mathbb{E}[X_1] \Big|}_{deviation} > \underbrace{\epsilon}_{accuracy} \Bigg] \leq \underbrace{2 \exp\Big(- \frac{2n\epsilon^2}{(b-a)^2} \Big)}_{confidence}$$

Monte-Carlo Approximation of a Mean

- *Unbiased estimator*: Then $\mathbb{E}[\mu_n] = \mu$ (and $\mathbb{V}[\mu_n] = \frac{\mathbb{V}[X]}{n}$)
- *Weak law of large numbers:* $\mu_n \xrightarrow{P} \mu$.
- *Strong law of large numbers*: $\mu_n \xrightarrow{a.s.} \mu$.
- *Central limit theorem (CLT)*: $\sqrt{n}(\mu_n - \mu) \xrightarrow{D} \mathcal{N}(0, \mathbb{V}[X])$.
- *Finite sample guarantee*:

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{t=1}^{n} X_t - \mathbb{E}[X_1]\right| > (b-a)\sqrt{\frac{\log 2/\delta}{2n}}\right] \leq \delta$$

Monte-Carlo Approximation of a Mean

- *Unbiased estimator*: Then $\mathbb{E}[\mu_n] = \mu$ (and $\mathbb{V}[\mu_n] = \frac{\mathbb{V}[X]}{n}$)
- *Weak law of large numbers:* $\mu_n \xrightarrow{P} \mu$.
- *Strong law of large numbers*: $\mu_n \xrightarrow{a.s.} \mu$.
- *Central limit theorem (CLT)*: $\sqrt{n}(\mu_n - \mu) \xrightarrow{D} \mathcal{N}(0, \mathbb{V}[X])$.
- *Finite sample guarantee*:

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{t=1}^{n} X_t - \mathbb{E}[X_1]\right| > \epsilon\right] \leq \delta$$

if $n \geq \frac{(b-a)^2 \log 2/\delta}{2\epsilon^2}$.

Exercise

Simulate n Bernoulli of probability p and verify the correctness and the accuracy of the C-H bounds.

Stochastic Approximation of a Mean

Definition

Let X a random variable bounded in $[0,1]$ with mean $\mu = \mathbb{E}[X]$ and $x_n \sim X$ be n i.i.d. realizations of X. The stochastic approximation of the mean is,

$$\mu_n = (1 - \eta_n)\mu_{n-1} + \eta_n x_n$$

with $\mu_1 = x_1$ and where (η_n) is a sequence of learning steps.

Stochastic Approximation of a Mean

Definition

Let X a random variable bounded in $[0,1]$ with mean $\mu = \mathbb{E}[X]$ and $x_n \sim X$ be n i.i.d. realizations of X. The stochastic approximation of the mean is,

$$\mu_n = (1 - \eta_n)\mu_{n-1} + \eta_n x_n$$

with $\mu_1 = x_1$ and where (η_n) is a sequence of learning steps.

Remark: When $\eta_n = \frac{1}{n}$ this is the *recursive* definition of empirical mean.

Stochastic Approximation of a Mean

Proposition (Borel-Cantelli)

Let $(E_n)_{n\geq 1}$ be a *sequence* of events such that $\sum_{n\geq 1}\mathbb{P}(E_n) < \infty$, then the probability of the *intersection of an infinite subset* is 0. More formally,

$$\mathbb{P}\Big(\limsup_{n\to\infty} E_n\Big) = \mathbb{P}\Big(\bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} E_k\Big) = 0.$$

Stochastic Approximation of a Mean

Proposition

If for any n, $\eta_n \geq 0$ and are such that

$$\sum_{n\geq 0} \eta_n = \infty; \qquad \sum_{n\geq 0} \eta_n^2 < \infty,$$

then

$$\mu_n \xrightarrow{a.s.} \mu,$$

and we say that μ_n is a *consistent* estimator.

Stochastic Approximation of a Mean

Proof. We focus on the case $\eta_n = n^{-\alpha}$.
In order to satisfy the two conditions we need $1/2 < \alpha \leq 1$. In fact, for instance

$$\alpha = 2 \Rightarrow \sum_{n \geq 0} \frac{1}{n^2} = \frac{\pi^2}{6} < \infty \quad \text{(see the Basel problem)}$$

$$\alpha = 1/2 \Rightarrow \sum_{n \geq 0} \left(\frac{1}{\sqrt{n}}\right)^2 = \sum_{n \geq 0} \frac{1}{n} = \infty \quad \text{(harmonic series).}$$

Stochastic Approximation of a Mean

Proof (cont'd).

Case $\alpha = 1$

Let $(\epsilon_k)_k$ a sequence such that $\epsilon_k \to 0$, *almost sure* convergence corresponds to

$$\mathbb{P}\Big(\lim_{n\to\infty} \mu_n = \mu\Big) = \mathbb{P}(\forall k, \exists n_k, \forall n \geq n_k, \big|\mu_n - \mu\big| \leq \epsilon_k) = 1.$$

From Chernoff-Hoeffding inequality for any ***fixed*** n

$$\mathbb{P}\big(\big|\mu_n - \mu\big| \geq \epsilon\big) \leq 2e^{-2n\epsilon^2}. \tag{1}$$

Let $\{E_n\}$ be a sequence of events $E_n = \{\big|\mu_n - \mu\big| \geq \epsilon\}$. From C-H

$$\sum_{n\geq 1} \mathbb{P}(E_n) < \infty,$$

and from Borel-Cantelli lemma we obtain that with probability 1 there exist only a *finite* number of n values such that $\big|\mu_n - \mu\big| \geq \epsilon$.

Stochastic Approximation of a Mean

Proof (cont'd).

Case $\alpha = 1$

Then for any ϵ_k there exist only a finite number of instants were $\big|\mu_n - \mu\big| \geq \epsilon_k$, which corresponds to have $\exists n_k$ such that

$$\mathbb{P}(\forall n \geq n_k, \big|\mu_n - \mu\big| \leq \epsilon_k) = 1$$

Repeating for all ϵ_k in the sequence leads to the statement.

Stochastic Approximation of a Mean

Proof (cont'd).
Case $\alpha = 1$
Then for any ϵ_k there exist only a finite number of instants were $\big|\mu_n - \mu\big| \geq \epsilon_k$, which corresponds to have $\exists n_k$ such that

$$\mathbb{P}(\forall n \geq n_k, \big|\mu_n - \mu\big| \leq \epsilon_k) = 1$$

Repeating for all ϵ_k in the sequence leads to the statement.

Remark: when $\alpha = 1$, μ_n is the Monte-Carlo estimate and this corresponds to the strong law of large numbers. A more precise and accurate proof is here:
http://terrytao.wordpress.com/2008/06/18/the-strong-law-of-large-numbers/

Stochastic Approximation of a Mean

Proof (cont'd).
Case $1/2 < \alpha < 1$. The stochastic approximation μ_n is

$$\mu_1 = x_1$$
$$\mu_2 = (1-\eta_2)\mu_1 + \eta_2 x_2 = (1-\eta_2)x_1 + \eta_2 x_2$$
$$\mu_3 = (1-\eta_3)\mu_2 + \eta_3 x_3 = (1-\eta_2)(1-\eta_3)x_1 + \eta_2(1-\eta_3)x_2 + \eta_3 x_3$$
$$\ldots$$
$$\mu_n = \sum_{i=1}^{n} \lambda_i x_i,$$

with $\lambda_i = \eta_i \prod_{j=i+1}^{n}(1-\eta_j)$ such that $\sum_{i=1}^{n} \lambda_i = 1$.
By C-H inequality

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{n} \lambda_i x_i - \sum_{i=1}^{n} \lambda_i \mathbb{E}[x_i]\Big| \geq \epsilon\Big) = \mathbb{P}\big(\big|\mu_n - \mu\big| \geq \epsilon\big) \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n}\lambda_i^2}}.$$

Stochastic Approximation of a Mean

Proof (cont'd).
Case $1/2 < \alpha < 1$**.**
From the definition of λ_i

$$\log \lambda_i = \log \eta_i + \sum_{j=i+1}^{n} \log(1-\eta_j) \leq \log \eta_i - \sum_{j=i+1}^{n} \eta_j$$

since $\log(1-x) < -x$. Thus $\lambda_i \leq \eta_i e^{-\sum_{j=i+1}^{n} \eta_j}$ and for any $1 \leq m \leq n$,

$$\begin{aligned}
\sum_{i=1}^{n} \lambda_i^2 &\leq \sum_{i=1}^{n} \eta_i^2 e^{-2\sum_{j=i+1}^{n} \eta_j} \\
&\overset{(a)}{\leq} \sum_{i=1}^{m} e^{-2\sum_{j=i+1}^{n} \eta_j} + \sum_{i=m+1}^{n} \eta_i^2 \\
&\overset{(b)}{\leq} m e^{-2(n-m)\eta_n} + (n-m)\eta_m^2 \\
&\overset{(c)}{=} m e^{-2(n-m)n^{-\alpha}} + (n-m)m^{-2\alpha}.
\end{aligned}$$

Stochastic Approximation of a Mean

Proof (cont'd).
Case $1/2 < \alpha < 1$.
Let $m = n^{\beta}$ with $\beta = (1 + \alpha/2)/2$ (i.e. $1 - 2\alpha\beta = 1/2 - \alpha$):

$$\sum_{i=1}^{n} \lambda_i^2 \leq n e^{-2(1-n^{-1/4})n^{1-\alpha}} + n^{1/2-\alpha} \leq 2n^{1/2-\alpha}$$

for *n big enough*, which leads to

$$\mathbb{P}\Big(\big|\mu_n - \mu\big| \geq \epsilon\Big) \leq e^{-\frac{\epsilon^2}{n^{1/2-\alpha}}}.$$

From this point we follow the same steps as for $\alpha = 1$ (application of the Borel-Cantelli lemma) and obtain the convergence result for μ_n.

Stochastic Approximation of a Fixed Point

Definition

Let $\mathcal{T} : \mathbb{R}^N \rightarrow \mathbb{R}^N$ be a contraction in some norm $||\cdot||$ with fixed point V. For any function W and state x, a noisy observation $\widehat{\mathcal{T}}W(x) = \mathcal{T}W(x) + b(x)$ is available.
For any $x \in X = \{1, \dots, N\}$, we defined the stochastic approximation

$$\begin{aligned} V_{n+1}(x) &= (1 - \eta_n(x))V_n(x) + \eta_n(x)(\hat{\mathcal{T}}V_n(x)) \\ &= (1 - \eta_n(x))V_n(x) + \eta_n(x)(\mathcal{T}V_n(x) + b_n), \end{aligned}$$

where η_n is a sequence of learning steps.

Stochastic Approximation of a Fixed Point

Proposition

Let $\mathcal{F}_n = \{V_0, \ldots, V_n, b_0, \ldots, b_{n-1}, \eta_0, \ldots, \eta_n\}$ the filtration of the algorithm and assume that

$$\mathbb{E}[b_n(x)|\mathcal{F}_n] = 0 \quad \text{and} \quad \mathbb{E}[b_n^2(x)|\mathcal{F}_n] \leq c(1 + ||V_n||^2)$$

for a constant c.
If the learning rates $\eta_n(x)$ are positive and satisfy the stochastic approximation conditions

$$\sum_{n\geq 0} \eta_n = \infty, \qquad \sum_{n\geq 0} \eta_n^2 < \infty,$$

then for any $x \in X$

$$V_n(x) \xrightarrow{a.s.} V(x).$$

Stochastic Approximation of a Zero

Robbins-Monro (1951) algorithm. Given a noisy function f, find x^* such that $f(x^*) = 0$.
In each x_n, observe $y_n = f(x_n) + b_n$ (with b_n a zero-mean independent noise) and compute

$$x_{n+1} = x_n - \eta_n y_n.$$

Stochastic Approximation of a Zero

Robbins-Monro (1951) algorithm. Given a noisy function f, find x^* such that $f(x^*) = 0$.
In each x_n, observe $y_n = f(x_n) + b_n$ (with b_n a zero-mean independent noise) and compute

$$x_{n+1} = x_n - \eta_n y_n.$$

If f is an *increasing* function, then under the same assumptions on the learning step

$$x_n \xrightarrow{a.s.} x^*$$

Stochastic Approximation of a Minimum

Kiefer-Wolfowitz (1952) algorithm. Given a function f and noisy observations of its gradient, find $x^* = \arg\min f(x)$.
In each x_n, observe $g_n = \nabla f(x_n) + b_n$ (with b_n a zero-mean independent noise) and compute

$$x_{n+1} = x_n - \eta_n g_n.$$

Stochastic Approximation of a Minimum

Kiefer-Wolfowitz (1952) algorithm. Given a function f and noisy observations of its gradient, find $x^* = \arg\min f(x)$.
In each x_n, observe $g_n = \nabla f(x_n) + b_n$ (with b_n a zero-mean independent noise) and compute

$$x_{n+1} = x_n - \eta_n g_n.$$

If the Hessian $\nabla^2 f$ is *positive*, then under the same assumptions on the learning step

$$x_n \xrightarrow{a.s.} x^*$$

Remark: this is often referred to as the **stochastic gradient** algorithm.

How to *solve incrementally* an RL problem

Reinforcement Learning Algorithms

Tools

Policy Evaluation

Policy Learning

The RL Interaction Protocol

For $i = 1, \dots, n$

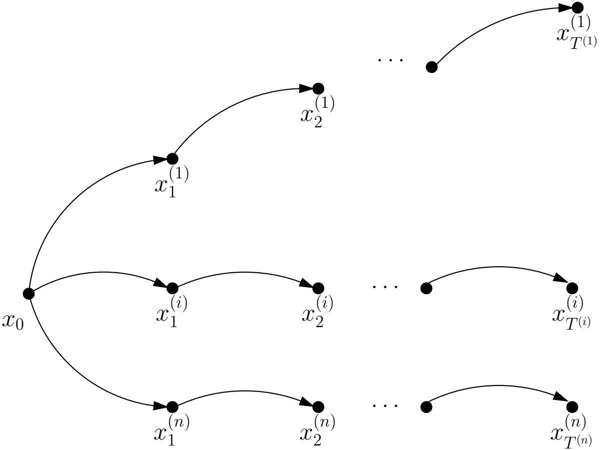
1. Set $t = 0$
2. Set initial state x_0 *[possibly random]*

   [execute one trajectory]
3. **While** (x_t not terminal)
   - 3.1 Take action a_t
   - 3.2 Observe next state x_{t+1} and reward r_t
   - 3.3 Set $t = t + 1$

   EndWhile

EndFor

The RL Interaction Protocol



Policy Evaluation

Objective: given a policy π evaluate its quality at the (fixed) initial state x_0

Policy Evaluation

Objective: given a policy π evaluate its quality at the (fixed) initial state x_0

For $i = 1, \ldots, n$

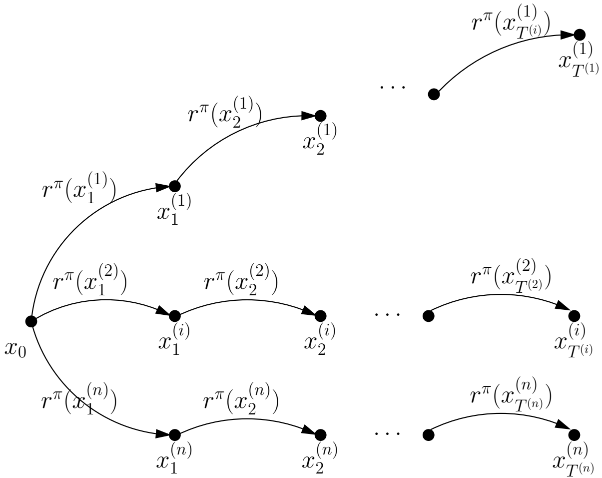
1. Set $t = 0$
2. Set initial state x_0 ~~***[possibly random]***~~

   [execute one trajectory]
3. **While** (x_t not terminal)
   - 3.1 Take action $a_t = \pi(x_t)$
   - 3.2 Observe next state x_{t+1} and ***reward*** $r_t = r^\pi(x_t)$
   - 3.3 Set $t = t + 1$

   EndWhile

EndFor

The RL Interaction Protocol



State Value Function

- *Infinite time horizon with terminal state*: the problem never terminates but the agent will eventually reach a *termination state*.

$$V^\pi(x) = \mathbb{E}\Big[\sum_{t=0}^{T} \gamma^t r(x_t, \pi(x_t)) | x_0 = x; \pi\Big],$$

where T is the first (*random*) time when the *termination state* is achieved.

Monte-Carlo Approximation

Idea: we can approximate an *expectation* by an *average*!

- Return of trajectory i

$$\widehat{R}_i(x_0) = \sum_{t=0}^{T^{(i)}} \gamma^t r^\pi(x_t^{(i)})$$

- Estimated value function

$$\widehat{V}_n^\pi(x_0) = \frac{1}{n}\sum_{i=1}^{n} \widehat{R}_i(x_0)$$

Monte-Carlo Approximation

For $i = 1, \ldots, n$

1. Set $t = 0$
2. Set initial state x_0 ~~***[possibly random]***~~

   [execute one trajectory]
3. **While** (x_t not terminal)
   - 3.1 Take action $a_t = \pi(x_t)$
   - 3.2 Observe next state x_{t+1} and ***reward*** $r_t = r^\pi(x_t)$
   - 3.3 Set $t = t + 1$

   EndWhile

EndFor

Collect trajectories and compute $\widehat{V}_n^\pi(x_0)$ *using MC approximation*

Monte-Carlo Approximation: Properties

- All returns are unbiased estimators of $V^\pi(x)$

$$\mathbb{E}[\widehat{R}^{(i)}(x_0)] = \mathbb{E}\Big[r^\pi(x_0^{(i)}) + \gamma r^\pi(x_1^{(i)}) + \cdots + \gamma^{T^{(i)}} r^\pi(x_{T^{(i)}}^{(i)})\Big] = V^\pi(x)$$

- Thus

$$\widehat{V}_n^\pi(x_0) \xrightarrow{a.s.} V^\pi(x_0).$$

- Finite-sample guarantees are also possible

Monte-Carlo Approximation: Extensions

Non-episodic problems

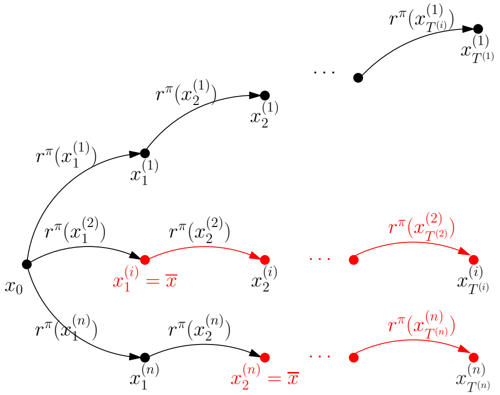
- Interrupt trajectories after H steps

$$\widehat{R}_i(x_0) = \sum_{t=0}^{H} \gamma^t r^\pi(x_t^{(i)})$$

- Loss in accuracy limited to $\gamma^H \frac{r_{\max}}{1-\gamma}$

Monte-Carlo Approximation: Extensions

Multiple subtrajectories



All *subtrajectories* starting with $\overline{x}$ can be used to estimate $V^\pi(\overline{x})$

First-visit and Every-Visit Monte-Carlo

Remark: any trajectory $(x_0, x_1, x_2, \ldots, x_T)$ contains also the sub-trajectory $(x_t, x_{t+1}, \ldots, x_T)$ whose return $\widehat{R}(x_t) = r^\pi(x_t) + \cdots + r^\pi(x_{T-1})$ could be used to build an estimator of $V^\pi(x_t)$.

First-visit and Every-Visit Monte-Carlo

Remark: any trajectory $(x_0, x_1, x_2, \ldots, x_T)$ contains also the sub-trajectory $(x_t, x_{t+1}, \ldots, x_T)$ whose return $\widehat{R}(x_t) = r^\pi(x_t) + \cdots + r^\pi(x_{T-1})$ could be used to build an estimator of $V^\pi(x_t)$.

- *First-visit MC.* For each state x we only consider the sub-trajectory when x is first achieved. *Unbiased estimator*, *only one sample per trajectory*.

First-visit and Every-Visit Monte-Carlo

Remark: any trajectory $(x_0, x_1, x_2, \dots, x_T)$ contains also the sub-trajectory $(x_t, x_{t+1}, \dots, x_T)$ whose return $\widehat{R}(x_t) = r^\pi(x_t) + \dots + r^\pi(x_{T-1})$ could be used to build an estimator of $V^\pi(x_t)$.

- *First-visit MC.* For each state x we only consider the sub-trajectory when x is first achieved. *Unbiased estimator*, *only one sample per trajectory*.
- *Every-visit MC.* Given a trajectory $(x_0 = x, x_1, x_2, \dots, x_T)$, we list all the m sub-trajectories starting from x up to x_T and we average them all to obtain an estimate. *More than one sample per trajectory*, *biased estimator*.

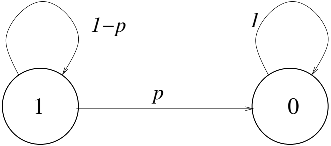
Question

More samples or no bias?

$\Rightarrow$ *Sometimes a biased estimator is preferable if consistent!*

First-visit vs Every-Visit Monte-Carlo

Example: 2-state Markov Chain



The reward is 1 while in state 1 (while is 0 in the terminal state). All trajectories are $(x_0 = 1, x_1 = 1, \ldots, x_T = 0)$. By Bellman equations

$$V(1) = 1 + (1-p)V(1) + 0 \cdot p = \frac{1}{p},$$

since $V(0) = 0$.

First-visit vs Every-Visit Monte-Carlo

We measure the mean squared error (MSE) of $\widehat{V}$ w.r.t. V

$$\mathbb{E}\big[(\widehat{V} - V)^2\big] = \underbrace{\big(\mathbb{E}[\widehat{V}] - V\big)^2}_{Bias^2} + \underbrace{\mathbb{E}\big[\big(\widehat{V} - \mathbb{E}[\widehat{V}]\big)^2\big]}_{Variance}$$

First-visit vs Every-Visit Monte-Carlo

First-visit Monte-Carlo. All the trajectories start from state 1, then the return over one single trajectory is exactly T, i.e., $\widehat{V} = T$.
The time-to-end T is a *geometric* r.v. with expectation

$$\mathbb{E}[\widehat{V}] = \mathbb{E}[T] = \frac{1}{p} = V^{\pi}(1) \Rightarrow \textit{unbiased estimator.}$$

Thus the MSE of $\widehat{V}$ coincides with the variance of T, which is

$$\mathbb{E}\Big[\Big(T - \frac{1}{p}\Big)^2\Big] = \frac{1}{p^2} - \frac{1}{p}.$$

First-visit vs Every-Visit Monte-Carlo

Every-visit Monte-Carlo. Given one trajectory, we can construct $T-1$ sub-trajectories (number of times state 1 is visited), where the t-th trajectory has a return $T-t$.

$$\widehat{V} = \frac{1}{T}\sum_{t=0}^{T-1}(T-t) = \frac{1}{T}\sum_{t'=1}^{T} t' = \frac{T+1}{2}.$$

The corresponding expectation is

$$\mathbb{E}\Big[\frac{T+1}{2}\Big] = \frac{1+p}{2p} \neq V^{\pi}(1) \Rightarrow \textit{ biased estimator}.$$

First-visit vs Every-Visit Monte-Carlo

Let's consider *n independent trajectories*, each of length T_i.
Total number of samples $\sum_{i=1}^n T_i$ and the estimator $\widehat{V}_n$ is

$$\widehat{V}_n = \frac{\sum_{i=1}^n \sum_{t=0}^{T_i-1}(T_i - t)}{\sum_{i=1}^n T_i} = \frac{\sum_{i=1}^n T_i(T_i+1)}{2\sum_{i=1}^n T_i}$$

$$= \frac{1/n \sum_{i=1}^n T_i(T_i+1)}{2/n \sum_{i=1}^n T_i}$$

$$\xrightarrow{a.s.} \frac{\mathbb{E}[T^2] + \mathbb{E}[T]}{2\mathbb{E}[T]} = \frac{1}{p} = V^\pi(1) \Rightarrow \textit{ consistent estimator}.$$

The MSE of the estimator

$$\mathbb{E}\Big[\big(\frac{T+1}{2} - \frac{1}{p}\big)^2\Big] = \frac{1}{2p^2} - \frac{3}{4p} + \frac{1}{4} \leq \frac{1}{p^2} - \frac{1}{p}.$$

First-visit vs Every-Visit Monte-Carlo

In general

- *Every-visit MC*: *biased* but *consistent* estimator.
- *First-visit MC*: *unbiased* estimator with potentially *bigger MSE*.

First-visit vs Every-Visit Monte-Carlo

In general

- *Every-visit MC*: *biased* but *consistent* estimator.
- *First-visit MC*: *unbiased* estimator with potentially *bigger MSE*.

Remark: when the state space is large the probability of visiting multiple times the same state is low, then the performance of the two methods tends to be the same.

Monte-Carlo Approximation: Extensions

Full estimate of V^π over any $x \in X$

- Use subtrajectories
- Restart from random states over X

Monte-Carlo Approximation: Limitations

- The estimate $\widehat{V}^{\pi}(x_0)$ is computed when ***all*** trajectories are terminated

Temporal Difference $TD(1)$

Idea: we can approximate an *expectation* by an ***incremental*** average!

- Return of trajectory i

$$\widehat{R}_i(x_0) = \sum_{t=0}^{T^{(i)}} \gamma^t r^\pi(x_t^{(i)})$$

- Estimated value function *after trajectory i*

$$\widehat{V}_i^\pi(x_0) = (1 - \alpha_i)\widehat{V}_{i-1}^\pi(x_0) + \alpha_i \widehat{R}_i(x_0)$$

Temporal Difference $TD(1)$

For $i = 1, \ldots, n$

1. Set $t = 0$
2. Set initial state x_0 ~~***[possibly random]***~~

   [execute one trajectory]
3. **While** (x_t not terminal)
   3.1 Take action $a_t = \pi(x_t)$
   3.2 Observe next state x_{t+1} and ***reward*** $r_t = r^\pi(x_t)$
   3.3 Set $t = t + 1$

   EndWhile
4. *Update* $\widehat{V}_i^\pi(x_0)$ using $TD(1)$ approximation

EndFor

~~*Collect trajectories and compute* $\widehat{V}_n^\pi(x_0)$ *using MC approximation*~~

Temporal Difference $TD(1)$: Properties

- If $\alpha_i = 1/i$, then $TD(1)$ is just the incremental version of the empirical mean

$$\widehat{V}_i^{\pi}(x_0) = \frac{n-1}{n}\widehat{V}_{i-1}^{\pi}(x_0) + \frac{1}{n}\widehat{R}_i(x_0)$$

- Using a generic step-size (learning rate) α_i gives *flexibility* to the algorithm

Temporal Difference $TD(1)$: Properties

Proposition

If the learning rate satisfies the Robbins-Monro conditions

$$\sum_{i=0}^{\infty} \alpha_i = \infty, \qquad \sum_{i=0}^{\infty} \alpha_i^2 < \infty,$$

then

$$\widehat{V}_n^{\pi}(x_0) \xrightarrow{a.s.} V^{\pi}(x_0)$$

Temporal Difference $TD(1)$: Extensions

- *Non-episodic problems*: Truncated trajectories
- *Multiple sub-trajectories*
  - Updates of all the states using sub-trajectories
  - *state-dependent learning rate* $\alpha_i(x)$
  - i is the index of the number of updates in that specific state

Temporal Difference $TD(1)$: Limitations

- The estimate $\widehat{V}^{\pi}(x_0)$ is updated when the trajectory is ***completely terminated***

The Bellman Equation

Proposition

For any stationary policy $\pi = (\pi, \pi, \dots)$, the state value function at a state $x \in X$ satisfies the *Bellman equation*:

$$V^\pi(x) = r(x, \pi(x)) + \gamma \sum_y p(y|x, \pi(x)) V^\pi(y).$$

Temporal Difference $TD(0)$

Idea: we can approximate V^π by estimating the Bellman error

Temporal Difference $TD(0)$

Idea: we can approximate V^π by estimating the Bellman error

- *Bellman error* of a function V in a state x

$$\mathcal{B}^\pi(V;x) = r^\pi(x) + \gamma \sum_y p(y|x,\pi(x))V(y) - V(x).$$

Temporal Difference $TD(0)$

Idea: we can approximate V^π by estimating the Bellman error

- *Bellman error* of a function V in a state x

$$\mathcal{B}^\pi(V;x) = r^\pi(x) + \gamma \sum_y p(y|x,\pi(x))V(y) - V(x).$$

- *Temporal difference* of a function $\widehat{V}^\pi$ for a transition $\langle x_t, r_t, x_{t+1}\rangle$

$$\delta_t = r_t + \gamma \widehat{V}^\pi(x_{t+1}) - \widehat{V}^\pi(x_t)$$

Temporal Difference $TD(0)$

Idea: we can approximate V^π by estimating the Bellman error

- *Bellman error* of a function V in a state x

$$\mathcal{B}^\pi(V;x) = r^\pi(x) + \gamma \sum_y p(y|x,\pi(x))V(y) - V(x).$$

- *Temporal difference* of a function $\widehat{V}^\pi$ for a transition $\langle x_t, r_t, x_{t+1} \rangle$

$$\delta_t = r_t + \gamma \widehat{V}^\pi(x_{t+1}) - \widehat{V}^\pi(x_t)$$

- Estimated value function *after transition* $\langle x_t, r_t, x_{t+1} \rangle$

$$\begin{aligned}\widehat{V}^\pi(x_t) &= \big(1 - \alpha_i(x_t)\big)\widehat{V}^\pi(x_t) + \alpha_i(x_t)\big(r_t + \gamma \widehat{V}^\pi(x_{t+1})\big) \\ &= \widehat{V}^\pi(x_t) + \alpha_i(x_t)\delta_t\end{aligned}$$

Temporal Difference $TD(0)$

For $i = 1, \dots, n$

1. Set $t = 0$
2. Set initial state x_0 ~~***[possibly random]***~~

   [execute one trajectory]
3. **While** (x_t not terminal)
   - 3.1 Take action $a_t = \pi(x_t)$
   - 3.2 Observe next state x_{t+1} and ***reward*** $r_t = r^\pi(x_t)$
   - 3.3 Set $t = t + 1$
   - 3.4 *Update* $\widehat{V}^\pi(x_t)$ using $TD(0)$ approximation

   EndWhile
4. ~~*Update* $\widehat{V}_i^\pi(x_0)$ using $TD(1)$ approximation~~

EndFor

~~*Collect trajectories and compute* $\widehat{V}_n^\pi(x_0)$ *using MC approximation*~~

Temporal Difference $TD(0)$: Properties

- The update rule

$$\widehat{V}^\pi(x_t) = \big(1 - \alpha_i(x_t)\big)\widehat{V}^\pi(x_t) + \alpha_i(x_t)\big(r_t + \gamma \widehat{V}^\pi(x_{t+1})\big)$$

is *bootstrapping* the current estimate of $\widehat{V}^\pi$ in other state.

- The temporal difference is an unbiased sample of the Bellman error

$$\mathbb{E}[\delta_t] = \mathbb{E}[r_t + \gamma \widehat{V}^\pi(x_{t+1}) - \widehat{V}^\pi(x_t)] = \mathcal{T}^\pi \widehat{V}^\pi(x_t) - \widehat{V}^\pi(x_t)$$

Temporal Difference $TD(0)$: Properties

Proposition

If the learning rate satisfies the Robbins-Monro conditions in all states $x \in X$

$$\sum_{i=0}^{\infty} \alpha_i(x) = \infty, \qquad \sum_{i=0}^{\infty} \alpha_i^2(x) < \infty,$$

and all states are visited *infinitely often*, then for all $x \in X$

$$\widehat{V}^{\pi}(x) \xrightarrow{a.s.} V^{\pi}(x)$$

Temporal Difference $TD(0)$

For $i = 1, \dots, n$

1. Set $t = 0$
2. Set $\widehat{V}^\pi(x) = 0, \quad \forall x \in X$
3. Set initial state x_0
4. **While** (x_t not terminal)
   4.1 Take action $a_t = \pi(x_t)$
   4.2 Observe next state x_{t+1} and ***reward*** $r_t = r^\pi(x_t)$
   4.3 Set $t = t + 1$
   4.4 Compute the TD $\delta_t = r_t + \gamma \widehat{V}^\pi(x_{t+1}) - \widehat{V}^\pi(x_t)$
   4.5 Update the value function estimate in x_t as
   $$\widehat{V}^\pi(x_t) = \widehat{V}^\pi(x_t) + \alpha_i(x_t)\delta_t$$
   4.6 Update the learning rate, e.g.,
   $$\alpha(x_t) = \frac{1}{\#\ \text{visits}(x_t)}$$

   EndWhile

Comparison between TD(1) and TD(0)

TD(1)

- Update rule

$$\widehat{V}^{\pi}(x_t) = \widehat{V}^{\pi}(x_t) + \alpha(x_t)[\delta_t + \gamma\delta_{t+1} + \cdots + \gamma^{T-1}\delta_T].$$

- No bias, large variance

TD(0)

- Update rule

$$\widehat{V}^{\pi}(x_t) = \widehat{V}^{\pi}(x_t) + \alpha(x_t)\delta_t.$$

- Potential bias, small variance

Comparison between TD(1) and TD(0)

TD(1)

- Update rule

$$\widehat{V}^\pi(x_t) \quad = \quad \widehat{V}^\pi(x_t) + \alpha(x_t)[\delta_t + \gamma\delta_{t+1} + \cdots + \gamma^{T-1}\delta_T].$$

- No bias, large variance

TD(0)

- Update rule

$$\widehat{V}^\pi(x_t) = \widehat{V}^\pi(x_t) + \alpha(x_t)\delta_t.$$

- Potential bias, small variance

$\Rightarrow$ $TD(\lambda)$ perform intermediate updates!

The $\mathcal{T}_\lambda^\pi$ Bellman operator

Definition

Given $\lambda < 1$, *then the Bellman operator* $\mathcal{T}_\lambda^\pi$ *is*

$$\mathcal{T}_\lambda^\pi = (1-\lambda) \sum_{m \geq 0} \lambda^m (\mathcal{T}^\pi)^{m+1}.$$

The $\mathcal{T}^\pi_\lambda$ Bellman operator

Definition

Given $\lambda < 1$, then the Bellman operator $\mathcal{T}^\pi_\lambda$ is

$$\mathcal{T}^\pi_\lambda = (1-\lambda)\sum_{m\geq 0} \lambda^m (\mathcal{T}^\pi)^{m+1}.$$

Remark: convex combination of the m-step Bellman operators $(\mathcal{T}^\pi)^m$ weighted by a sequences of coefficients defined as a function of a λ.

Temporal Difference $TD(\lambda)$

Idea: use the whole series of temporal differences to update $\widehat{V}^\pi$

- *Temporal difference* of a function $\widehat{V}^\pi$ for a transition $\langle x_t, r_t, x_{t+1}\rangle$

$$\delta_t = r_t + \gamma \widehat{V}^\pi(x_{t+1}) - \widehat{V}^\pi(x_t)$$

- Estimated value function

$$\widehat{V}^\pi(x_t) = \widehat{V}^\pi(x_t) + \alpha_i(x_t) \sum_{s=t}^{T} (\gamma\lambda)^{s-t} \delta_s$$

Temporal Difference $TD(\lambda)$

Idea: use the whole series of temporal differences to update $\widehat{V}^\pi$

- *Temporal difference* of a function $\widehat{V}^\pi$ for a transition $\langle x_t, r_t, x_{t+1}\rangle$

$$\delta_t = r_t + \gamma \widehat{V}^\pi(x_{t+1}) - \widehat{V}^\pi(x_t)$$

- Estimated value function

$$\widehat{V}^\pi(x_t) = \widehat{V}^\pi(x_t) + \alpha_i(x_t) \sum_{s=t}^{T} (\gamma\lambda)^{s-t} \delta_s$$

$\Rightarrow$ Still requires the whole trajectory before updating...

Temporal Difference $TD(\lambda)$: Eligibility Traces

- *Eligibility* traces $z \in \mathbb{R}^N$
- For every transition $x_t \to x_{t+1}$

Temporal Difference $TD(\lambda)$: Eligibility Traces

- *Eligibility* traces $z \in \mathbb{R}^N$
- For every transition $x_t \rightarrow x_{t+1}$
    1. Compute the temporal difference

$$d_t = r^\pi(x_t) + \gamma \widehat{V}^\pi(x_{t+1}) - \widehat{V}^\pi(x_t)$$

Temporal Difference $TD(\lambda)$: Eligibility Traces

- *Eligibility* traces $z \in \mathbb{R}^N$
- For every transition $x_t \to x_{t+1}$
  1. Compute the temporal difference

$$d_t = r^\pi(x_t) + \gamma \widehat{V}^\pi(x_{t+1}) - \widehat{V}^\pi(x_t)$$

  2. Update the eligibility traces

$$z(x) = \begin{cases} \lambda z(x) & \text{if } x \neq x_t \\ 1 + \lambda z(x) & \text{if } x = x_t \\ 0 & \text{if } x_t = x_0 \text{ (reset the traces)} \end{cases}$$

Temporal Difference $TD(\lambda)$: Eligibility Traces

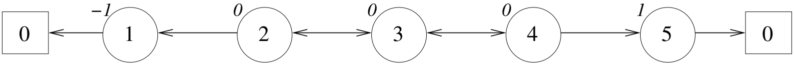
- *Eligibility* traces $z \in \mathbb{R}^N$
- For every transition $x_t \to x_{t+1}$
  1. Compute the temporal difference
  $$d_t = r^\pi(x_t) + \gamma \widehat{V}^\pi(x_{t+1}) - \widehat{V}^\pi(x_t)$$
  2. Update the eligibility traces
  $$z(x) = \begin{cases} \lambda z(x) & \text{if } x \neq x_t \\ 1 + \lambda z(x) & \text{if } x = x_t \\ 0 & \text{if } x_t = x_0 \text{ (reset the traces)} \end{cases}$$
  3. For all state $x \in X$
  $$\widehat{V}^\pi(x) \leftarrow \widehat{V}^\pi(x) + \alpha(x) z(x) \delta_t.$$

Sensitivity to λ

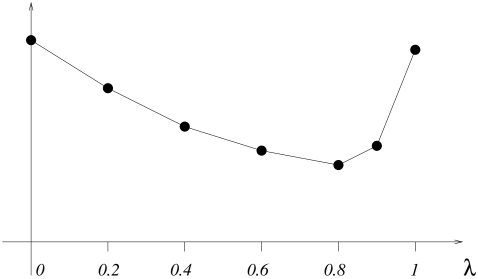
- $\lambda < 1$: *smaller variance* w.r.t. $\lambda = 1$ (MC/TD(1)).
- $\lambda > 0$: *faster propagation* of rewards w.r.t. $\lambda = 0$.

Example: Sensitivity to λ

Linear chain example



The MSE of V_n w.r.t. V^π after $n = 100$ trajectories:



How to *solve incrementally* an RL problem

Reinforcement Learning Algorithms

Tools

Policy Evaluation

Policy Learning

Question

How do we compute the optimal policy online?

$\Rightarrow$ *Q-learning!*

Learning the Optimal Policy

Objective: learn the optimal policy π^* with direct interaction with the environment

Learning the Optimal Policy

Objective: learn the optimal policy π^* with direct interaction with the environment

For $i = 1, \ldots, n$

1. Set $t = 0$
2. Set initial state x_0
3. **While** (x_t not terminal)
    3.1 Take action a_t
    3.2 Observe next state x_{t+1} and reward r_t
    3.3 Set $t = t + 1$

   EndWhile

EndFor

Policy Iteration

1. Let π_0 be *any* stationary policy
2. At each iteration $k = 1, 2, \dots, K$
   - *Policy evaluation* given π_k, compute Q^{π_k}.
   - *Policy improvement*: compute the *greedy* policy

$$\pi_{k+1}(x) \in \arg\max_{a \in A} Q_k^{\pi}(x)$$

3. Return the last policy π_K

SARSA

Idea: alternate policy evaluation and policy improvement

SARSA

Idea: alternate policy evaluation and policy improvement

- Define a *greedy exploratory* policy with temperature τ

$$\pi_Q(a|x) = \frac{\exp(Q(x,a)/\tau)}{\sum_{a'} \exp(Q(x,a')/\tau)}$$

The higher $Q(x,a)$, the more probability to take action a in state x

SARSA

Idea: alternate policy evaluation and policy improvement

- Define a *greedy exploratory* policy with temperature τ

$$\pi_Q(a|x) = \frac{\exp(Q(x,a)/\tau)}{\sum_{a'} \exp(Q(x,a')/\tau)}$$

  The higher $Q(x,a)$, the more probability to take action a in state x

- Compute the temporal difference on the trajectory $\langle x_t, a_t, r_t, x_{t+1}, a_{t+1} \rangle$ (with actions chosen according to $\pi_Q(a|x)$)

$$\delta_t = r_t + \gamma \widehat{Q}(x_{t+1}, a_{t+1}) - \widehat{Q}(x_t, a_t)$$

SARSA

Idea: alternate policy evaluation and policy improvement

- Define a *greedy exploratory* policy with temperature τ

$$\pi_Q(a|x) = \frac{\exp(Q(x,a)/\tau)}{\sum_{a'} \exp(Q(x,a')/\tau)}$$

The higher $Q(x,a)$, the more probability to take action a in state x

- Compute the temporal difference on the trajectory $\langle x_t, a_t, r_t, x_{t+1}, a_{t+1} \rangle$ (with actions chosen according to $\pi_Q(a|x)$)

$$\delta_t = r_t + \gamma \widehat{Q}(x_{t+1}, a_{t+1}) - \widehat{Q}(x_t, a_t)$$

- Update the estimate of Q as

$$\widehat{Q}(x_t, a_t) = \widehat{Q}(x_t, a_t) + \alpha(x_t, a_t)\delta_t$$

SARSA: Properties

- The *TD* updates make $\widehat{Q}$ converge to Q^{π}
- The update of π_Q allows to improve the policy
- A decreasing temperature allows to become more and more greedy

$\Rightarrow$ If $\tau \to 0$ with a proper rate, then $\widehat{Q} \to Q^*$ and $\pi_Q \to \pi^*$

SARSA: Limitations

The actions a_t need to be selected according to the current Q

$\Rightarrow$ ***On-policy learning***

The Optimal Bellman Equation

Proposition

The optimal value function V^* (i.e., $V^* = \max_\pi V^\pi$) is the solution to the *optimal Bellman equation*:

$$V^*(x) = \max_{a \in A} \Big[r(x,a) + \gamma \sum_y p(y|x,a) V^*(y) \Big].$$

Q-Learning

Idea: use *TD* for the optimal Bellman operator

- Compute the (optimal) temporal difference on the trajectory $\langle x_t, a_t, r_t, x_{t+1} \rangle$ (with actions chosen ***arbitrarily***!)

$$\delta_t = r_t + \gamma \max_{a'} \widehat{Q}(x_{t+1}, a') - \widehat{Q}(x_t, a_t)$$

- Update the estimate of Q as

$$\widehat{Q}(x_t, a_t) = \widehat{Q}(x_t, a_t) + \alpha(x_t, a_t)\delta_t$$

Q-Learning: Properties

Proposition

If the learning rate satisfies the Robbins-Monro conditions in all states $x \in X$

$$\sum_{i=0}^{\infty} \alpha_i(x) = \infty, \qquad \sum_{i=0}^{\infty} \alpha_i^2(x) < \infty,$$

and all states are visited *infinitely often*, then for all $x \in X$

$$\widehat{Q}(x) \xrightarrow{a.s.} Q^*(x)$$

Remark: "infinitely often" requires a steady exploration policy

Learning the Optimal Policy

For $i = 1, \ldots, n$

1. Set $t = 0$
2. Set initial state x_0
3. **While** (x_t not terminal)

   3.1 Take action a_t ***according to a suitable exploration policy***

   3.2 Observe next state x_{t+1} and reward r_t

   3.3 Compute the temporal difference

   $$\delta_t = r_t + \gamma \widehat{Q}(x_{t+1}, a_{t+1}) - \widehat{Q}(x_t, a_t) \quad \textit{(SARSA)}$$

   $$\delta_t = r_t + \gamma \max_{a'} \widehat{Q}(x_{t+1}, a') - \widehat{Q}(x_t, a_t) \quad \textit{(Q-learning)}$$

   3.4 Update the Q-function

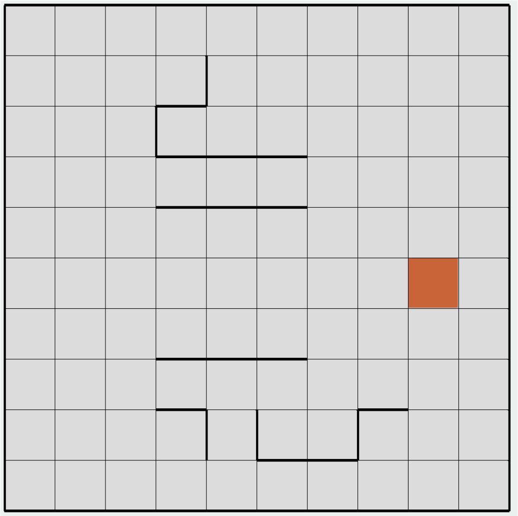
   $$\widehat{Q}(x_t, a_t) = \widehat{Q}(x_t, a_t) + \alpha(x_t, a_t)\delta_t$$

   3.5 Set $t = t + 1$

   EndWhile

EndFor

The Grid-World Problem



Bibliography I

Reinforcement Learning

Alessandro Lazaric
alessandro.lazaric@inria.fr

sequel.lille.inria.fr