# FoSML: Lecture #24

- Let $\pi^*$ be an optimal deterministic policy, which may be thought of as just a function $\pi^*: S \to A$ from states to actions.
- Let $V^*(s) = Q^*(s, \pi^*(s))$ be the associated optimal value function. The Bellman condition is:

$$\text{(★)} \quad \forall s \in S, \; V^*(s) = \max_{a \in A} \left\{ \mathbb{E}(r(s,a)) + \gamma \sum_{s' \in S} \mathbb{P}(s' \mid s, a) \cdot V^*(s') \right\}.$$

This is a non-linear system of equations, since the max function is non-linear.

- We can alternatively consider a linear programming approach. That is, we can try to phrase solving for the optimal policy/value in our MDP as a optimization problem with a linear objective function and linear constraints.
- Note that we can characterize (★) as attempting to minimize $\{V(s) \mid s \in S\}$ subject to the constraint that

$$V(s) \geq \mathbb{E}(r(s,a)) + \gamma \sum_{s' \in S} \mathbb{P}(s' \mid s, a) \cdot V(s'),$$

since minimizing over $\{V(s)\}_{s \in S}$ while maintaining the above leads to equality with the RHS as big as possible.

- So, let $\{\alpha(s)\}_{s \in S}$ be a set of fixed positive weights $\alpha(s) \in \mathbb{R}_+$. We recast (★) as the following linear programming (LP) problem:

$$\min_V \sum_{s\in S} \alpha(s) V(s) \quad \text{subject to } \forall s\in S,\ \forall a\in A,\ V(s) \geq \mathbb{E}(r(s,a)) + \gamma \sum_{s'\in S} \mathbb{P}(s'|s,a)\cdot V(s'),$$

where $\alpha > 0$ is the vector of $\{\alpha(s)\}_{s\in S}$.

- We could also add the normalizing constraint that $\sum_{s\in S} \alpha(s) = 1$.
- The constraint set here is quite large.
- We may also give this a dual formulation, which may be easier for optimization:

$$\max_X \sum_{\substack{s\in S\\ a\in A}} \mathbb{E}(r(s,a)) x(s,a) \quad \text{subject to } \forall s\in S,\ \sum_{a\in A} x(s,a) = \alpha(s) + \gamma \sum_{\substack{s'\in S\\ a\in A}} \mathbb{P}(s'|s,a)\, x(s',a)$$

- $\forall s\in S,\ \forall a\in A,\ x(s,a) \geq 0.$

- In general, one can try a simplex method approach or interior point algorithm to solve a linear program. Problem-specific tools may also be considered.

=

- Suppose now that we cannot access $P(s'|s,a)$ and $r(s,a)$, and instead need to estimate these as we go along?
- We can imagine the agent starting with no knowledge, then following the hidden reward-transition environment to simultaneously (1.) Collect rewards
  (2.) Estimate $\mathbb{P}(s'|s,a)$, $r(s,a)$.
- We consider two approaches:

(a.) Model-free approach: just try to learn a good $\pi^*$

(b.) Model-based approach: try to learn $\mathbb{P}(s'|s,a)$, $r(s,a)$ as we go, then use them to estimate $\pi^*$.

- To help introduce these ideas, we consider the more general process of stochastic approximation (SA)

- SA helps to estimate solutions to optimization problems where objective function is the expectation of a random variable, and thus cannot be easily evaluated directly.

- Methods from statistical inference are highly relevant, so we state a few.

Thm (Mean Estimation): Let $X$ be a r.v. with values in $[0,1]$. Let $\{x_i\}_{i=0}^{m} \overset{iid}{\sim} X$. Let $\{\mu_m\}_{m\in\mathbb{N}}$ be defined inductively by

$$\mu_{m+1} = \mu_m \cdot (1-\alpha_m) + \alpha_m x_m,$$

where $\mu_0 = x_0$, $\alpha_m \in [0,1]$, and $\sum_{m=0}^{\infty} \alpha_m = \infty$, $\sum_{m=0}^{\infty} \alpha_m^2 < \infty$. Then

$$\mu_m \xrightarrow{a.s.} \mathbb{E}(X), \text{ i.e. } \mathbb{P}(\mu_m \to \mathbb{E}(X)) = 1.$$

Proof: We will show convergence in $L^2$; the result follows from slightly more work (and in general convergence in $L^2 \Rightarrow$ a.s. convergence).

Since the sample $\{x_i\}_{i=0}^{m}$ are independent,

$$\mathrm{Var}(\mu_{m+1}) = \mathrm{Var}(\mu_m \cdot (1-\alpha_m) + \alpha_m x_m)$$

$$= \mathrm{Var}(\mu_m \cdot (1-\alpha_m)) + \mathrm{Var}(\alpha_m \cdot x_m)$$

$= (1-\alpha_m)^2 \operatorname{Var}(\mu_m) + \alpha_m^2 \operatorname{Var}(X_m)$

$\leq (1-\alpha_m)^2 \operatorname{Var}(\mu_m) + \alpha_m^2$, since $X_m \in [0,1]$.

We will show $\operatorname{Var}(\mu_m) \to 0$, from which the desired $L^2$-convergence follows. Let us proceed by contradiction: suppose $\exists \varepsilon > 0$ s.t. for all $m \geq N$, $\operatorname{Var}(\mu_m) \geq \varepsilon$. Then, for all $m \geq N$,

$$\begin{aligned} &\operatorname{Var}(\mu_{m+1}) \\ &\leq (1-\alpha_m)\operatorname{Var}(\mu_m) + \alpha_m^2 \\ &\leq \operatorname{Var}(\mu_m) - \alpha_m \varepsilon + \alpha_m^2 \quad (\star) \end{aligned}$$

Iterating this argument yields $\operatorname{Var}(\mu_{m+N}) \leq \operatorname{Var}(\mu_m) - \varepsilon \sum_{n=N}^{m+N} \alpha_n + \sum_{n=N}^{m+N} \alpha_n^2$

But, by assumption on $\{\alpha_m\}_{m=0}^{\infty}$, $\left(-\varepsilon \sum_{n=N}^{m+N} \alpha_n + \sum_{n=N}^{m+N} \alpha_n^2\right) \to -\infty$ as $m \to \infty$, which violates the non-negativity of $\operatorname{Var}(\mu_{m+N})$. So, $\forall \varepsilon > 0$, $\exists m_0 \geq N$ s.t. $\operatorname{Var}(\mu_{m_0}) \leq \varepsilon$.

Now, let $N$ be sufficiently large so that for all $m \geq N$, $\alpha_m \leq \varepsilon$ (this holds because $\alpha_m \to 0$). We claim $\forall m \geq m_0$, $\operatorname{Var}(\mu_m) \leq \varepsilon$, from whence the desired convergence in $L^2$ follows. Indeed, suppose $\operatorname{Var}(\mu_m) \leq \varepsilon$ for some $m \geq m_0$. Then by $(\star)$, and the fact that $\alpha_m \leq \varepsilon$,

$$\begin{aligned} \operatorname{Var}(\mu_{m+1}) &\leq (1-\alpha_m)\operatorname{Var}(\mu_m) + \alpha_m^2 \\ &\leq (1-\alpha_m)\varepsilon + \varepsilon \cdot \alpha_m \end{aligned}$$

$= \varepsilon$, as desired. ∎

Remark: When $\alpha_m = \frac{1}{m}$, we recover the classical strong law of large numbers.

- A prototypical problem in stochastic approximation looks like $x = H(x)$, i.e. finding a fixed point of a function $H$, where $H$ is not easily accessed. Instead, we get noisy samples of the form $\{H(x_i) + w_i\}_{i=1}^m$, where $w_i$ is some noise.

- We can approximately solve for $x^* = H(x^*)$ by using our above theorem. Indeed, let us define a sequence of iterates $\{X_t\}_{t=0}^\infty$ as

$$X_{t+1} = X_t + \alpha_t \left( H(X_t) + w_t - X_t \right),$$

where $\{\alpha_t\}_{t=0}^\infty$ are as in the theorem ($\sum_{t=0}^\infty \alpha_t = \infty$, $\sum_{t=0}^\infty \alpha_t^2 < \infty$, $\alpha_t > 0$.)

- A more general formulation replaces $\left[ H(X_t) + w_t - X_t \right]$ by a general function $D(X_t, w_t)$, where $D : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^N$, i.e. we consider updates of the form

$$X_{t+1} = X_t + \alpha_t D(X_t, w_t).$$

Q: Under what conditions on $D$ does this converge?

- We will use results on martingales to develop guarantees of convergence.

**Theorem (supermartingale convergence)**: Let $\{X_t\}_{t\in\mathbb{N}}, \{Y_t\}_{t\in\mathbb{N}}, \{Z_t\}_{t\in\mathbb{N}}$ be sequences of non-negative random variables. Suppose the following hypotheses hold:

(a.) $\sum_{t=0}^{\infty} Y_t < \infty$

(b.) For $\mathcal{F}_t = \left\{\{X_{t'}\}_{t'\le t}, \{Y_{t'}\}_{t'\le t}, \{Z_{t'}\}_{t'\le t}\right\}$,

$\mathbb{E}(X_{t+1} \mid \mathcal{F}_t) \le X_t + Y_t - Z_t.$

Then:

(c.) $\{X_t\}_{t=0}^{\infty}$ has an a.s. limit.

(d.) $\sum_{t=0}^{\infty} Z_t < \infty$.

**Remark**: A sequence of r.v. $\{X_t\}_{t\in\mathbb{N}}$ is a martingale if

(a.) $\mathbb{E}(|X_t|) < \infty \quad \forall t$

(b.) $\mathbb{E}(X_{t+1} \mid X_t, X_{t-1}, \dots, X_0) = X_t$

Since we have "$\le$" in (b), we call this result a "supermartingale" result.

- We can use our supermartingale convergence theorem to prove the following.

**Theorem**: Let $H: \mathbb{R}^N \to \mathbb{R}^N$, $\{w_t\}_{t\in\mathbb{N}}$ a sequence of r.v. in $\mathbb{R}^N$, $\{\alpha_t\}_{t\in\mathbb{N}}$ a sequence of reals, and let the sequence $\{X_t\}_{t\in\mathbb{N}}$ be

$\forall s \in \{1,\dots,N\}, \quad X_{t+1}(s) = X_t(s) + \alpha_t(s)\left[H(X_t)(s) - X_t(s) + w_t(s)\right].$

Let $\mathcal{F}_t$ denote the history up to time $t$, i.e. $\mathcal{F}_t = \{\{x_{t'}\}_{t'\le t}, \{w_{t'}\}_{t'\le t}, \{\alpha_{t'}\}_{t'\le t}\}$

Let $\Psi(x) = \frac{1}{2}\|x - x^*\|_2^2$ for some $x^* \in \mathbb{R}^N$. Suppose $D$, $\{\alpha_t\}_{t\in\mathbb{N}}$ satisfy:

(a.) $\exists K_1, K_2 \in \mathbb{R}$ s.t. $\mathbb{E}\left(\|D(x_t, w_t)\|_2^2 \mid \mathcal{F}_t\right) \le K_1 + K_2 \Psi(x_t)$.

(b.) $\exists c \ge 0$ s.t. $\nabla \Psi(x_t)^T \mathbb{E}\left(D(x_t, w_t) \mid \mathcal{F}_t\right) \le -c\,\Psi(x_t)$.

(c.) $\alpha_t \ge 0 \quad \forall t$

(d.) $\sum_{t=0}^{\infty} \alpha_t = +\infty, \quad \sum_{t=0}^{\infty} \alpha_t^2 < \infty$.

Then $x_t \xrightarrow{a.s.} x^*$ as $t \to \infty$.

Proof: Notice that $\Psi$ may be expanded in Taylor expansion as

$$\Psi(x_{t+1}) = \Psi(x_t) + \nabla\Psi(x_t)^T \cdot (x_{t+1} - x_t) + \frac{1}{2}(x_{t+1} - x_t) \cdot \nabla^2 \Psi(x_t)(x_{t+1} - x_t).$$

We get equality, since $\Psi$ is quadratic. So,

$$\begin{aligned}
\mathbb{E}\left(\Psi(x_{t+1}) \mid \mathcal{F}_t\right) &= \Psi(x_t) + \alpha_t \nabla\Psi(x_t)^T \mathbb{E}\left(D(x_t, w_t \mid \mathcal{F}_t)\right) \\
&\quad + \frac{\alpha_t^2}{2} \mathbb{E}\left(\|D(x_t, w_t)\|^2 \mid \mathcal{F}_t\right) \\
&\le \Psi(x_t) - \alpha_t c\,\Psi(x_t) + \frac{\alpha_t^2}{2}\left(K_1 + K_2 \Psi(x_t)\right) \\
&= \Psi(x_t) + \frac{\alpha_t^2 K_1}{2} - \left(\alpha_t c - \frac{\alpha_t^2 K_2}{2}\right)\Psi(x_t)
\end{aligned}$$

Since $\alpha_t \to 0$ as $t \to \infty$, for $t$ large enough, $\operatorname{sgn}\left(\alpha_t c - \frac{\alpha_t^2 K_2}{2}\right)$

$$= \operatorname{sgn}(\alpha_t c).$$

Since $\Psi \ge 0$, $\alpha_t \ge 0$, $c > 0$, we have that $\left(\alpha_t c - \frac{\alpha_t^2 K_2}{2}\right)\Psi(x_t)$ is

eventually non-negative. By our supermartingale convergence theorem, $\Psi(x_t)$ converges and $\sum_{t=0}^{\infty} \left(\alpha_t C - \frac{\alpha_t^2 K_2}{2}\right) \Psi(x_t) < \infty$. (*)

$\sum_{t=0}^{\infty}$

Now, we know $\sum_{t=0}^{\infty} \frac{\alpha_t^2 K_2}{2} \Psi(x_t) < \infty$ since $\Psi(x_t)$ converges and $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$. So, the only way (*) holds is if $\sum_{t=0}^{\infty} \alpha_t \cdot C \cdot \Psi(x_t) < \infty$

$\Leftrightarrow \sum_{t=0}^{\infty} \alpha_t \cdot \Psi(x_t) < \infty$.

Since $\sum_{t=0}^{\infty} \alpha_t = \infty$, it follows that $\Psi(x_t) \to 0$, as desired. ∎