Faculté des Sciences et Techniques **Monday, December 4, 2023**
Université de Limoges
M2 ACSYON

Reinforcement learning and stochastic optimization

Exam (3 hours)
No documents are allowed.

Problem 1 A gambler plays bets at n periods. At each period he bets any nonnegative amount up to his present fortune. The gambler wins the bet amount with probability $p \in]0,1[$ and loses it otherwise. The gambler's aim is to maximize the expectation of the logarithm of his final fortune. For $x > 0$, denote by $V_k(x)$ the maximal expected return if the gambler has a present fortune of x and is allowed k ($k = 0, \ldots, n$) more bets.

(i) **[2pts]** Write the dynamic programming equation for V_k.

(ii) **[3.5pts]** For any $x > 0$ compute $V_n(x)$ and the optimal betting strategy.

Problem 2 Let $H \colon \mathbb{R}^d \to \mathbb{R}^d$ and $x^* \in \mathbb{R}^d$ be such that $H(x^*) = x^*$. Assume that there exists $\beta \in]0,1[$ such that

$$\|H(x) - x^*\| \leq \beta \|x - x^*\| \quad \text{for all } x \in \mathbb{R}^d,$$

where $\| \cdot \|$ denotes the euclidian norm. In order to approximate x^*, we consider the following stochastic algorithm

$$x_{k+1} = (1 - \gamma_k)x_k + \gamma_k (H(x_k) + w_k),$$

where $(\gamma_k)_{k\in\mathbb{N}} \subset]0,1[$ is a deterministic sequence and $(w_k)_{k\in\mathbb{N}}$ is a sequence of random variables taking values in $\mathbb{R}^d$. Let $\mathcal{F}_0 = \{x_0\}$ and, for $k \geq 1$, let us set $\mathcal{F}_k = \{x_0, \ldots, x_k, \ldots, w_0, \ldots, w_{k-1}\}$. We assume that

- $\diamond$ $\mathbb{E}\left(w_k | \mathcal{F}_k\right) = 0 \quad \text{for all } k \in \mathbb{N}.$
- $\diamond$ $\mathbb{E}\left(\|w_k\|^2 \,|\, \mathcal{F}_k\right) \leq A + B\|x_k - x^*\|^2,$

for some positive constants A and B.

(i) **[2pts]** Show that

$$\langle H(x) - x, x - x^* \rangle \leq -(1 - \beta)\|x - x^*\|^2 \quad \text{for all } x \in \mathbb{R}^d,$$

where, for $y,\, z \in \mathbb{R}^d$, $\langle y, z \rangle$ denotes the usual scalar product in $\mathbb{R}^d$.

(ii) **[3.5pts]** Show that if

$$\left(\sum_{k=0}^{\infty} \gamma_k = \infty \right) \wedge \left(\sum_{k=0}^{\infty} \gamma_k^2 < \infty \right).$$

then, almost surely, $x_k \underset{k\to\infty}{\longrightarrow} x^*$

Problem 3 Let $n \in \mathbb{N}^*$, $\mathcal{X} = \{1, \ldots, n\}$, and A be a finite set. Let $\{p(j|i,a) \,|\, i,\, j \in \mathcal{X},\, a \in A\}$ be a *controlled Markov kernel in* $\mathcal{X}$, i.e.

$$p(j|i,a) \geq 0 \quad \text{for all } i,\, j \in \mathcal{X},\, a \in A, \quad \sum_{j\in\mathcal{X}} p(j|i,a) = 1 \quad \text{for all } i \in \mathcal{X},\, a \in A.$$

Let $\gamma \in]0,1[$, and let $r\colon \mathcal{X}\times A \to \mathbb{R}$ be bounded. Consider the discounted infinite horizon Markov decision problem

$$V_i^* = \sup_{\pi\colon \mathcal{X}\to A} \mathbb{E}\left(\sum_{k=0}^{\infty} \gamma^k r\left(X_k^{i,\pi}, \pi\left(X_k^{i,\pi}\right)\right)\right) \quad \text{for all } i \in \mathcal{X},$$

where $X_k^{i,\pi}$ ($k \in \mathbb{N}$) is the Markov chain, taking values in $\mathcal{X}$, starting at i and with probability transitions given by

$$\mathbb{P}(X_{k+1}^{i,\pi} = i_{k+1} \,|\, X_k^{i,\pi} = i_k) = p(i_{k+1}|i_k, \pi(i_k)) \quad \text{for all } k \in \mathbb{N},\ i_k,\ i_{k+1} \in \mathcal{X}.$$

Now, define $\widetilde{\mathcal{X}} = \{0, 1, \dots, n\}$ and, for all $i,\ j \in \widetilde{\mathcal{X}}$ and $a \in A$, set

$$\widetilde{p}(j|i,a) = \begin{cases} \gamma p(j|i,a) & \text{if } i,\ j \in \mathcal{X}, \\ 1-\gamma & \text{if } i \in \mathcal{X},\ j = 0, \\ 1 & \text{if } i = j = 0, \\ 0 & \text{otherwise.} \end{cases}$$

1. **[2pts]** Show that $\{\widetilde{p}(j|i,a) \,|\, i,\ j \in \widetilde{\mathcal{X}},\ a \in A\}$ is a controlled Markov kernel in $\widetilde{\mathcal{X}}$.

Given a policy $\widetilde{\pi}\colon \widetilde{\mathcal{X}} \to A$ in $\widetilde{\mathcal{X}}$ and $i \in \widetilde{\mathcal{X}}$, define $\widetilde{X}_k^{i,\widetilde{\pi}}$ ($k \in \mathbb{N}$) as the Markov chain, with values in $\widetilde{\mathcal{X}}$, starting at i and with probability transitions given by

$$\mathbb{P}(\widetilde{X}_{k+1}^{i,\widetilde{\pi}} = i_{k+1} \,|\, \widetilde{X}_k^{i,\widetilde{\pi}} = i_k) = \widetilde{p}(i_{k+1}|i_k, \widetilde{\pi}(i_k)) \quad \text{for all } k \in \mathbb{N},\ i_k,\ i_{k+1} \in \widetilde{\mathcal{X}}$$

and set

$$\tau^{i,\widetilde{\pi}} = \inf\{k \in \mathbb{N} \,|\, \widetilde{X}_k^{i,\widetilde{\pi}} = 0\}$$

2. **[2pts]** Show that, for every $i \in \widetilde{\mathcal{X}}$ and $\widetilde{\pi}\colon \widetilde{\mathcal{X}} \to A$, we have $\tau^{i,\widetilde{\pi}} < +\infty$ almost surely.

Consider the following undiscounted infinite horizon Markov decision problem

$$\begin{aligned} \widetilde{V}_i^* &= \max_{\widetilde{\pi}\colon \widetilde{\mathcal{X}}\to A} \mathbb{E}\left(\sum_{k=0}^{\tau^{i,\widetilde{\pi}}-1} r\left(\widetilde{X}_k^{i,\widetilde{\pi}}, \widetilde{\pi}\left(\widetilde{X}_k^{i,\widetilde{\pi}}\right)\right)\right) \quad \text{for all } i \in \mathcal{X}, \\ \widetilde{V}_0^* &= 0. \end{aligned} \tag{1}$$

3. **[3pts]** Show that, for every $i \in \mathcal{X}$, we have $V_i^* = \widetilde{V}_i^*$. Does there exist an optimal policy for both problems?

The previous result shows that this modification of a discounted Markov decision problem to an undiscounted one with terminal state preserves the value function.

4. **[2pts]** Let $\lambda \in [0,1]$ and $i \in \mathcal{X}$. Use the previous idea to provide a $\mathrm{TD}(\lambda)$ method to approximate

$$V_i^\pi = \mathbb{E}\left(\sum_{k=0}^{\infty} \gamma^k r\left(X_k^{i,\pi}, \pi\left(X_k^{i,\pi}\right)\right)\right)$$

for a given policy π.[1]

[1] A different $\mathrm{TD}(\lambda)$ method can be constructed by working directly with the discounted problem. It can be shown that the latter has smaller variance at the price of having to observe infinite trajectories.