Optimization and Learning via Stochastic Gradient Search

Including Answers

(Period 2, 2022/2023)
Exam 20 December 2022

The exam offers **10 bonous points**. The total number of points is 110 while the grade is capped at 100 points, i.e., your grade will be computed by $\min(\text{credits}, 100)/10$.

Problem 1 (45 points)

Let $X(\theta)$ be a parameterized random variable with cumulative distribution function (CDF) $F_\theta(x)$, given by

$$F_\theta(x) = \begin{cases} 0 & x \le 0; \\ e^{-(\theta/x)^8} & x > 0, \end{cases}$$

where parameter $\theta > 0$. Furthermore, a cost function $h(x, \theta)$ is given by

$$h(x,\theta) = x^2 + \frac{1}{\theta^2},$$

and an objective function

$$J(\theta) = \mathbb{E}\big[h(X(\theta), \theta)\big] = \mathbb{E}\Big[X^2(\theta) + \frac{1}{\theta^2}\Big]$$

(a). (10 points) Apply the inverse transform method (ITM) for generating samples of $X(\theta)$. Give the resulting one-line formula for $X(\theta)$.

(b). (10 points) Another method for sampling from distributions is the accept-reject method (ARM). Discuss some pros and cons of the ITM and ARM.

(c). (10 points) Derive the infinite perturbation analysis (IPA) estimator of $J'(\theta)$.

(d). (10 points) Derive the score function method (SFM) estimator of $J'(\theta)$.

(e). (5 points) Discuss some pros and cons of the IPA and SFM estimators.

Solution:

(a). Solve $F_\theta(x) = u$ for $x > 0$ for any $u \in (0, 1)$:

$$e^{-(\theta/x)^8} = u \quad \Leftrightarrow \quad x = \theta\,(-\log u)^{-1/8}.$$

Thus, $X(\theta) = \theta\,(-\log U)^{-1/8}$ for the uniform $U = U(0, 1)$.

(b). ITM: (+) one-line formula; (+) can be vectorized (+) fixed number of calls to RNG; (-) not often applicable; (-) might have complicated mathematical functions; (-) sensitive for numerical errors.
ARM: (+) generally applicable; (-) random number of calls to RNG; (-) acceptance probability could be small; (-) need an efficient proposal distribution.

(c). Since θ is scale parameter (see (a)), $X'(\theta) = X(\theta)/\theta$. Thus,

$$\begin{aligned} D^{\text{IPA}}(\theta) &= \frac{\partial}{\partial x} h(x,\theta)|_{x=X(\theta)} \times X'(\theta) + \frac{\partial}{\partial \theta} h(x,\theta)|_{x=X(\theta)} \\ &= \frac{2X^2(\theta)}{\theta} - \frac{2}{\theta^3}. \end{aligned}$$

(d). First the PDF:

$$\begin{aligned} f_\theta(x) &= \frac{\partial}{\partial x} F_\theta(x) = \frac{\partial}{\partial x} e^{-(\theta/x)^8} = \frac{\partial}{\partial x} e^{-(x/\theta)^{-8}} \\ &= \frac{8}{\theta}\big(\frac{x}{\theta}\big)^{-9} e^{-(x/\theta)^{-8}} = \frac{8}{\theta}\big(\frac{\theta}{x}\big)^{9} e^{-(\theta/x)^{8}}. \end{aligned}$$

Then the score function

$$\begin{aligned} S(\theta,x) &= \frac{\partial}{\partial \theta} \log f_\theta(x) = \frac{\partial}{\partial \theta}\Big(-\log\theta + 9\log\theta - \big(\frac{\theta}{x}\big)^8\Big) \\ &= \frac{8}{\theta} - \frac{8\theta^7}{x^8} = \frac{8}{\theta}\Big(1 - \big(\frac{\theta}{x}\big)^8\Big). \end{aligned}$$

The objective function is rewritten to $J(\theta) = \mathbb{E}[X^2(\theta)] + 1/\theta^2$, which yields the score function estimator,

$$D^{\text{SFM}}(\theta) = X^2(\theta) S(\theta, X(\theta)) - \frac{2}{\theta^3} = \frac{8X^2(\theta)}{\theta}\Big(1 - \big(\frac{\theta}{X(\theta)}\big)^8\Big) - \frac{2}{\theta^3}.$$

(e). IPA: (+) unbiased estimator; (+) CRN usage; (+) low variance; (-) structural parameter; (-) interchange conditions.
SFM: (+) unbiased estimator; (+) easily implementable; (-) distributional parameter; (-) interchange conditions; (-) high variance.

Problem 2 (20 points)

Consider a container terminal within a large harbour. Ships arrive (randomly, for instance according to a Poisson process) at the terminal, they carry a load of (exactly) 20 containers, and after arrival they are immediately served, or they wait for service. Service is the unloading of the carried containers. When all the containers of the ship are unloaded, the ship leaves, and service of the next ship (if present) starts immediately. There is a single unloading dock that serves ships on a first-come-first-served basis. The containers of the first ship are unloaded one-by-one, for each it takes a random time (for instance, exponentially distributed). An unloaded container is transferred to a truck, that after getting the container drives immediately away to bring the container to its destination. If there is no truck available, the unloading process is stopped until a truck arrival. Trucks arrive at the terminal (randomly, for instance according to a Poisson process).

(a). (10 points) Provide a DES modelling of the container terminal operations. Specify states, state space, events, event lists, and transition functions.

(b). (10 points) Check whether the commuting condition holds.

Solution:

(a). DES model.

- States: $s = (n_1, n_2, n_3)$, where

  $n_1 =$ number of ships present $\in \{0, 1, \ldots\}$;
  $n_2 =$ number of trucks present $\in \{0, 1, \ldots\}$;
  $n_3 =$ number of containers (at the first ship) to unload at the dock $\in \{0, 1, \ldots, 20\}$.

- State space: $\mathcal{S} = \{0, 1, \ldots\} \times \{0, 1, \ldots\} \times \{0, 1, \ldots, 20\} \setminus I$, where I are the infeasible states $(0, n_2, n_3 \geq 1)$ (no ship, thus there can be no containers), and $(n_1 \geq 1, n_2, 0)$ (ships present, thus there are containers to unload).
- Events $\mathcal{E} = \{\alpha_1, \alpha_2, \beta\}$, where

$$\begin{aligned} \alpha_1 &= \text{ship arrival;} \\ \alpha_2 &= \text{truck arrival;} \\ \beta &= \text{container unloaded.} \end{aligned}$$

- Event lists $L(s) \subset \mathcal{E}$ for $s \in \mathcal{S}$:

$$\begin{aligned} L(0, n_2, 0) &= \{\alpha_1, \alpha_2\}; \\ L(n_1 \geq 1, 0, n_3 \geq 1) &= \{\alpha_1, \alpha_2\}; \\ L(n_1 \geq 1, n_2 \geq 1, n_3 \geq 1) &= \{\alpha_1, \alpha_2, \beta\}. \end{aligned}$$

- State transitions $\phi(s, e) \in \mathcal{S}$ for $s \in \mathcal{S}$ and $e \in L(s)$:

$$\begin{aligned} \phi\big((0, n_2, 0), \alpha_1\big) &= (1, n_2, 20); && (\textit{ship arrival}) \\ \phi\big((n_1 \geq 1, n_2, n_3 \geq 1), \alpha_1\big) &= (n_1 + 1, n_2, n_3); && (\textit{ship arrival}) \\ \phi\big((n_1, n_2, n_3), \alpha_2\big) &= (n_1, n_2 + 1, n_3); && (\textit{truck arrival}) \\ \phi\big((n_1 \geq 1, n_2 \geq 1, n_3 \geq 2), \beta\big) &= (n_1, n_2 - 1, n_3 - 1); && (\textit{container unloaded}) \\ \phi\big((n_1 \geq 2, n_2 \geq 1, 1), \beta\big) &= (n_1 - 1, n_2 - 1, 20); && (\textit{container unloaded}) \\ \phi\big((1, n_2 \geq 1, 1), \beta\big) &= (0, n_2 - 1, 0). && (\textit{container unloaded}) \end{aligned}$$

(b). The commuting condition says $\phi(\phi(s, e), e') = \phi(\phi(s, e'), s)$ for any state s and events $e, e' \in L(s)$. This holds clearly for the two arrival events, i.e., $e = \alpha_1, e' = \alpha_2$. For instance

$$\begin{aligned} (0, 0, 0) &\overset{\alpha_1}{\to} (1, 0, 20) \overset{\alpha_2}{\to} (1, 1, 20); \\ (0, 0, 0) &\overset{\alpha_2}{\to} (0, 1, 0) \overset{\alpha_1}{\to} (1, 1, 20). \end{aligned}$$

Now consider the states with the unloading event and an arrival event. We show the CC in case of $\{\beta, \alpha_1\}$; with α_2 goes similarly.

$$(n_1 \geq 1, n_2 \geq 1, n_3 \geq 2) \overset{\beta}{\rightarrow} (n_1, n_2 - 1, n_3 - 1) \overset{\alpha_1}{\rightarrow} (n_1 + 1, n_2 - 1, n_3 - 1);$$
$$(n_1 \geq 1, n_2 \geq 1, n_3 \geq 2) \overset{\alpha_1}{\rightarrow} (n_1 + 1, n_2, n_3) \overset{\beta}{\rightarrow} (n_1 + 1, n_2 - 1, n_3 - 1);$$

$- -$

$$(n_1 \geq 2, n_2 \geq 1, 1) \overset{\beta}{\rightarrow} (n_1 - 1, n_2 - 1, 20) \overset{\alpha_1}{\rightarrow} (n_1, n_2 - 1, 20);$$
$$(n_1 \geq 2, n_2 \geq 1, 1) \overset{\alpha_1}{\rightarrow} (n_1 + 1, n_2, 1) \overset{\beta}{\rightarrow} (n_1, n_2 - 1, 20);$$

$- -$

$$(1, n_2 \geq 1, 1) \overset{\beta}{\rightarrow} (0, n_2 - 1, 0) \overset{\alpha_1}{\rightarrow} (1, n_2 - 1, 20);$$
$$(1, n_2 \geq 1, 1) \overset{\alpha_1}{\rightarrow} (2, n_2, 1) \overset{\beta}{\rightarrow} (1, n_2 - 1, 20);$$

Problem 3 (25 points)

Consider the vector field $G(\theta)$, for $\theta \in \mathbb{R}^2$.

(i). (10 points) Judging from Figure 1, what is the nature of point θ^*? (stationary, asymptotically stable, globally asymptotically stable, none)

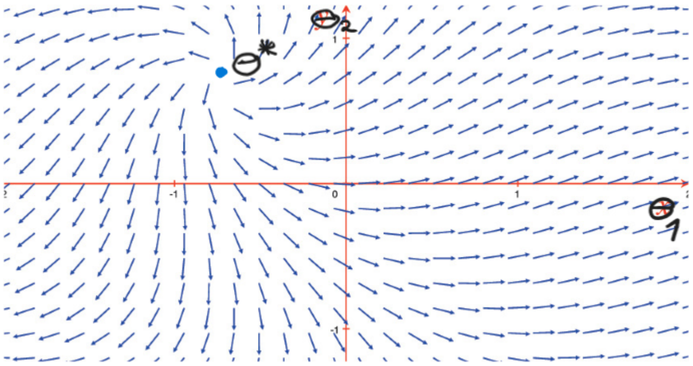

Figure 1: Vector Field $G(\theta)$

(ii). (5 points) Given is the problem of finding the minimizer of $J(\theta)$, for $J(\theta) \in \mathcal{C}^2$, i.e., you are looking for $\theta^* = \arg\min J(\theta)$. Suppose that G in (i) is coercive for this problem. What can be said about the relation between $G(\theta)$ and the gradient of $J(\theta)$?

(iii). (10 points) Assume that you use the algorithm

$$\theta_{n+1} = \theta_n - \epsilon_n G(\theta_n),$$

for $\epsilon_n = 3/(n + 10)^{2/3}$. Does this choice satisfy the convergence conditions for decreasing ϵ?

Solution:

(i). From the arrows you see that the arrows point away from θ^*. This indicates that θ^* is a stable point.

Alternatively: You may have misinterpreted the graph as showing the gradient field. Then θ^* is a globally asymptomatically stable point for $G(\theta) = -\nabla J(\theta)$, which is obtained from turning the arrows around in the figure.

(ii). We can conclude the $G(\theta)$ is a descent direction, i.e., $G(\theta)\nabla J(\theta) < 0$ for all non-stationary points θ

(iii). We note that $\sum \frac{1}{n^p}$ is finite if and only if $p < 1$. Applying this to $p = 2/3$ and $p = 4/3$ shows the result.

Problem 4 (20 points)

For a stochastic approximation of the form

$$\theta_{n+1} = \theta_n + \epsilon Y_n$$

the output shown in Figure 2 was obtained.

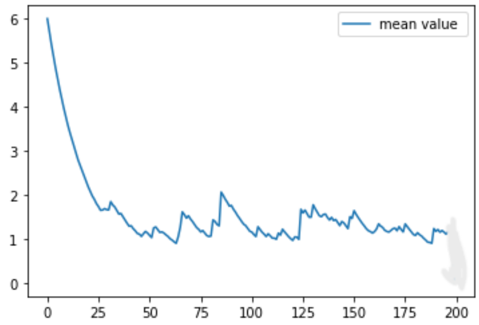

Figure 2: Vector Field $G(\theta)$

(i). (10 points) Argue that, judging from the output,

$$\theta^* \approx \frac{1}{N}\sum_{i=k+1}^{k+N} \theta_i,$$

for $k = 50$.

(ii). (10 points) How would you design an experiment so that you can build a confidence interval for θ^* rather than just producing a point estimator as above?

Solution:

(i). We use a fixed gain size algorithm. Provided that the appropriate conditions hold, $\{\theta_n : n \geq L\}$, for L sufficiently large, approximates a mean reverting stationary process (this is actually an Orstein-Uhlenbeck process). Judging from the figure, $L = 50$ is sufficient. As the θ_n process is (approximately) stationary after $n = 50$, averaging will lead an estimator for the mean value (which is a proxy for the solution).

(ii). To begin, determine n such that θ_n is approximately normally distributed (which is will become for n large enough due to our theory). Then sample as many realization of θ_n as the budget allows, and produce a confidence interval with it.