Exam:
Stochastic Gradient Techniques in Optimization and Learning period 4.2

December 2019

Problem 1 (15 Credits)

Let $J(\theta) = 3\sin(2\theta + c)$, for some $c > 0$, and $\theta \in \mathbb{R}$. For $\alpha \in (0, 1)$, we want to find θ^* such that

$$J(\theta^*) = \alpha. \tag{1}$$

Let

$$G(\theta) = -(J(\theta) - \alpha)).$$

(a). [5 Credits] Show that the stationary points of $J(\theta)$ fail to be stable points of $G(\theta)$.

(b). [5 Credits] Under what conditions on α does θ^* become asymptotically stable?

(c). [5 Credits] Rewrite $J(\theta^*) = \alpha$ as optimization problem

$$\min_\theta \frac{1}{2}(J(\theta) - \alpha)^2.$$

Argue that $G(\theta)$ is not coercive for this optimization problem, i.e., for tracking the solutions of $J(\theta^*) = \alpha$.

Answer Problem 1:

(a) At stationary points the value of $J(\theta)$ is either -3 or 3. The vector field $G(\theta)$ moves towards $3 > \alpha > -3$ and therefore the stationary points are not stable points of $G(\theta)$.

(b) θ^* is asymptotically stable if $G(\theta)$ moves towards θ^* when it is in the neighborhood of θ^*. This only holds for the chosen G if $J(\theta)$ is monotone increasing in the neighborhood of θ^*.

(c) $J(\theta) = \alpha$ has infinitely many solutions. Moreover, as $J(\theta)$ is not monotone $G(\theta)$ moves in the areas where $J(\theta)$ is monotone decreasing towards a maximum.

Problem 2 (total 15 Credits)

The gradient-field of a function $J(\theta)$ is shown in Figure 1. Apply a steepest descent algorithm for

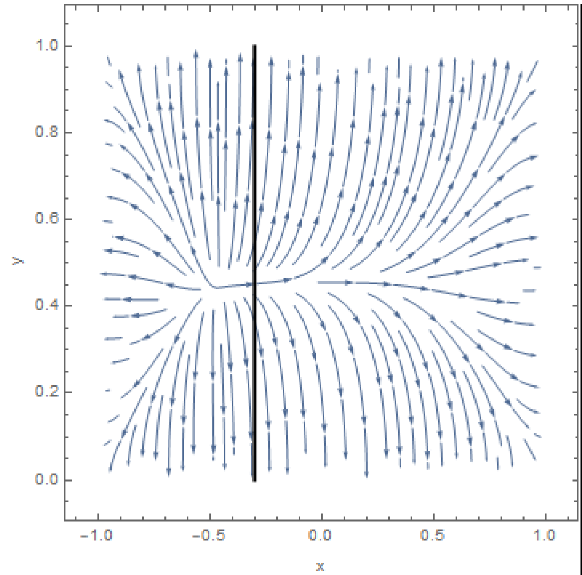

Figure 1: Gradient-field of $J(\theta)$

finding the minimum of $J(\theta)$.

(a). [5 Credits] Discuss with Figure 1 (where you ignore the black line for this part of the problem) for the ODE

$$\frac{d}{dt}x(t) = -\nabla J(x(t))$$

the nature of point $(0, 0.5)$ (stable, asymptotically stable, or unstable).

(b). [5 Credits] Judging from the figure (where you ignore the black line for this part of the problem), is this problem well-posed and is the vector-field coercive?

(c). [5 Credits] Now suppose we use a descent algorithm for finding the minimum of $J(\theta)$ on the constraint set given by the black line, i.e., only the points on the black line are admissible. Argue that $(-0.3, 0.425)$ is an asymptotically stable point for the ODE living on the constraint set.

Answer Problem 2:

(a) $(0, 0.5)$ is not stable (and therefor not asymptotically stable) as some arrows points towards this point while other points away from it.

(b) Yes, moving in opposite direction of the arrows always leads to approximately $(-0.5, 0.45)$

(c) The ODE will move on the black line following the opposite direction of the arrows as much as possible, and this ODE will therefore move to $(-0.3, 0.425)$.

Problem 3 (total 20 Credits)

Consider the algorithm

$$\theta_{n+1} = \theta_n + \epsilon_n Y_n,$$

for finding some optimal solution θ^*, for $\epsilon_n = 1/(n+1)$, for $n \in \mathbb{N}$. Suppose that evaluating Y_n requires one sample from an underlying process. Suppose your computational budget is sufficient to sample N samples from the underlying process, and you split your simulation budget to produce k independent runs of the algorithm yielding $\theta_n(\omega_i)$, $1 \leq i \leq k$, for each of the runs, where $kn = N$. Running your experiment with $k = 1000$, $n = 100$, and $N = 10^5$ yields the histogram in Figure 2. The black line is the density of the normal distribution fitted to the data.

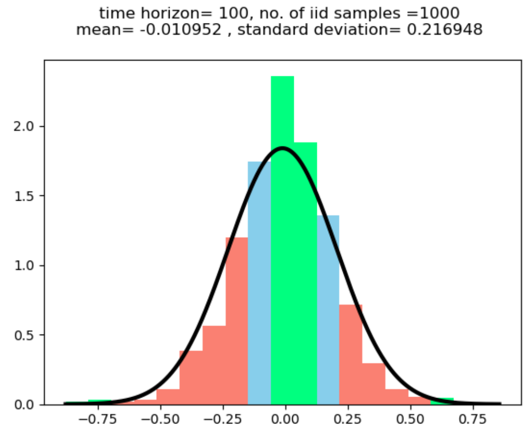

Figure 2: Histogram of θ_{100}

(a). [5 Credits] Given the available information, can the claim "$\theta^* = 0.0$" be rejected at confidence level 0.05? (You may use that $1/\sqrt{1000} \approx 0.031$, and $1.96 \times 0.031 \times 0.2169 \approx 0.0134$.)

(b). [5 Credits] Suppose you would run the algorithm with fixed ϵ rather than decreasing ϵ_n. What would you except to find for the resulting output? Will the mean value be effected, will the variance be be effected? Explain!

(c). [10 Credits] Keeping the budget fixed, provide a new choice for n and k that may improve the statistical properties of θ_n.

Answer Problem 3:

(a) 0.0 lies within a 95 % confidence interval and the claim $\theta^* = 0.0$ cannot be rejected.

(b) Fixed ϵ yields weak convergence, so θ_n is a random variable, and for decreasing ϵ, θ_n converges a.s. So, fixed ϵ will result in θ_n with larger variance (decreasing will have eventually zero variance).

(c) The histogram shows that the normal distribution fit is not very good. This suggests that n is too small, and it is advisable to increase n by a factor of, say, 2, and use $\hat{n} = 2n = 200$, $\hat{k} = 500$.

Problem 4 (total 20 Credits)

Let $X(\theta)$ be a continuous random variable with a probability cumulative distribution function

$$F_\theta(x) = \begin{cases} 0, & x \le 0; \\ 1 - e^{-\theta\sqrt{x}}, & x > 0, \end{cases} \tag{2}$$

for $\theta > 0$.

(a). [5 Credits] Show that the score function is

$$S(\theta, x) = \frac{1}{\theta} - \sqrt{x}.$$

(b). [10 Credits] Let the parameter space be $\Theta = [1, 4]$. Show that

$$\frac{d}{d\theta}\mathbb{E}\big[h\big(X(\theta)\big)\big] = \mathbb{E}\big[h\big(X(\theta)\big)\, S\big(\theta, X(\theta)\big)\big].$$

for $\theta \in \Theta$ and all polynomially bounded cost functions h, i.e., $h(x) = O(x^k)$ for some positive integer k.

(c). [5 Credits] Note that (2) is the Weibull distribution with scale parameter θ, and with shape parameter $\alpha = \frac{1}{2}$. Set $\theta = 1$. How would you generate samples from this distribution? Work out a few details.

Answer Problem 4:

(a) First, determine the PDF,

$$f_\theta(x) \doteq \frac{\partial}{\partial x} F_\theta(x) = \frac{\theta}{2\sqrt{x}}\, e^{-\theta\sqrt{x}}, \quad x > 0.$$

The score function is

$$S(\theta, x) \doteq \frac{\frac{\partial}{\partial\theta} f_\theta(x)}{f_\theta(x)} = \frac{\partial}{\partial\theta} \log f_\theta(x)$$
$$= \frac{\partial}{\partial\theta}\Big(\log\theta - \log(2\sqrt{x}) - \theta\sqrt{x}\Big) = \frac{1}{\theta} - \sqrt{x}, \quad x > 0.$$

(b)

$$\frac{d}{d\theta}\mathbb{E}\big[h\big(X(\theta)\big)\big] = \frac{d}{d\theta}\int_0^\infty h(x) f_\theta(x)\, dx.$$

A sufficient condition for interchanging $\frac{d}{d\theta}$ and $\int$ on $\Theta = [1, 4]$ is bounded convergence, which holds if

$$\int_0^\infty |h(x)| \sup_{\theta\in\Theta} \Big|\frac{\partial}{\partial\theta} f_\theta(x)\Big|\, dx < \infty.$$

Do the calculus,

$$\Big|\frac{\partial}{\partial\theta} f_\theta(x)\Big| = \Big|\frac{1}{2\sqrt{x}} - \frac{\theta}{2}\Big|\, e^{-\theta\sqrt{x}} \le \Big(\frac{1}{\sqrt{x}} + 2\Big)\, e^{-\sqrt{x}},$$

for $\theta \in \Theta = [1, 4]$. Clearly, for any $k \geq 0$,

$$\int_0^\infty x^k \big(\frac{1}{\sqrt{x}} + 2\big)\, e^{-\sqrt{x}}\, dx < \infty.$$

Hence, interchange is allowed for $h(x) = O(x^k)$, and results in

$$\begin{aligned}\frac{d}{d\theta}\mathbb{E}\big[h\big(X(\theta)\big)\big] &= \frac{d}{d\theta}\int_0^\infty h(x) f_\theta(x)\, dx = \int_0^\infty h(x) \frac{d}{d\theta} f_\theta(x)\, dx \\ &= \int_0^\infty h(x) \frac{\frac{d}{d\theta} f_\theta(x)}{f_\theta(x)}\, f_\theta(x)\, dx = \mathbb{E}\big[h\big(X(\theta)\big)\, S\big(\theta, X(\theta)\big)\big].\end{aligned}$$

(c) Let $X = X(1)$ be the random variable for $\theta = 1$. It has CDF

$$F(x) = 1 - e^{-\sqrt{x}}, \quad x > 0.$$

It holds that $F(X) = U$, the uniform random variable on $(0, 1)$, thus $X = F^{-1}(U)$. In other words, to get a sample x of X, it suffices to solve $F(x) = u$ for any $u \in (0, 1)$.

$$\begin{aligned} &F(x) = u \;\Leftrightarrow\; 1 - e^{-\sqrt{x}} = u \;\Leftrightarrow\; e^{-\sqrt{x}} = 1 - u \\ &\Leftrightarrow\; -\sqrt{x} = \ln(1 - u) \;\Leftrightarrow\; x = \big(-\ln(1 - u)\big)^2. \end{aligned}$$

Note that because $1 - U \stackrel{\mathcal{D}}{=} U$, and $(-1)^2 = 1$, you might do

$$x = \big(\ln(u)\big)^2, \quad u \in (0, 1).$$

Problem 5 (total 30 Credits)

(a). [6 Credits] Let $\Theta = (a, b) \subset \mathbb{R}$ be a finite interval, and $f : \Theta \to \mathbb{R}$ a (real-valued) function on it.

(i). Give the definition of Lipschitz continuity of f on (a, b).

(ii). Suppose that f is differentiable on (a, b); then give a sufficient condition for Lipschitz continuity that is more easy to check than the definition in (i).

(b). [9 Credits] Are the following (deterministic) functions Lipschitz continuous? If yes, show by applying (a)-(i) or (a)-(ii); if no, argue that (a)-(i) does not hold.

(i). $f(\theta) = \theta\, e^{\theta}$ on $(1, 2)$.

(ii). $f(\theta) = |\theta|$ on $(-1, 1)$.

(iii). $f(\theta) = \log\theta$ on $(0, 1)$.

(c). [15 Credits] Is $Y(\theta)$ almost surely Lipschitz continuous in the following cases? And if so, is the Lipschtiz modulus integrable? Just an answer is not sufficient. Provide an analysis where you proof your clams.

(i). $Y(\theta) = X/\theta$ where $X \overset{\mathcal{D}}{\sim} \mathsf{Ex}(1)$ (the exponential distribution with parameter 1), and $\theta \in \Theta = (1, 2)$.

(ii). $Y(\theta) = 1/(\theta U)$ where $U \overset{\mathcal{D}}{\sim} U(0, 1)$ (uniform distribution), and $\theta \in \Theta = (1, 2)$.

(iii). $Y(\theta) = \sqrt{|X - \theta|}$ where $X \overset{\mathcal{D}}{\sim} \mathsf{Ex}(1)$ (the exponential distribution with parameter 1), and $\theta \in \Theta = (1, 2)$.

Answer Problem 5:

(a) (i). There exists $0 < K < \infty$ such that for any $\theta_1, \theta_2 \in \Theta$,

$$|f(\theta_1) - f(\theta_2)| \le K|\theta_1 - \theta_2|.$$

(ii). You may take

$$K = \sup_{\theta\in\Theta} |f'(\theta)|$$

as Lipschitz constant.

(b) (i). Yes. Apply (a)(ii):

$$\sup_{\theta\in(1,2)} |f'(\theta)| = \sup_{\theta\in(1,2)} |1 + \theta|e^{\theta} = 3e^2 < \infty.$$

(ii). Yes. Apply (a)(i): w.l.o.g., assume $-1 < \theta_2 < \theta_1 < 1$.

$$|f(\theta_1) - f(\theta_2)| = \big||\theta_1| - |\theta_2|\big| = \begin{cases} |\theta_1 - \theta_2|, & -1 < \theta_2 < \theta_1 \le 0; \\ |\theta_1 - \theta_2|, & 0 \le \theta_2 < \theta_1 < 1; \\ |\theta_1 - (-\theta_2)| \le |\theta_1 - \theta_2|, & -1 < \theta_2 < 0 < \theta_1 < 1. \end{cases}$$

Thus Lipschitz constant $K = 1$.

(iii). No. Let $\theta_1 = 0.5$ and $\theta_2 = 1/n$, where $n = 1, 2, \ldots$. Then

$$|\log\theta_1 - \log\theta_2| = \log(n/2),$$

which you cannot bound for all n. Equivalently,

$$|f'(\theta)| = \frac{1}{\theta} \overset{\theta\downarrow 0}{\to} \infty.$$

(c) (i). Let $x > 0$ be a random element of X. Then, the function $Y(\theta) = x/\theta$ is Lipschitz continuous on $\Theta = (1, 2)$. For instance, apply (a)(ii):

$$K(x) = \sup_{\theta\in\Theta} |Y'(\theta)| = \sup_{\theta\in(1,2)} x/\theta^2 = x < \infty.$$

This holds for any $x > 0$ drawn from the exponential(1) distribution, thus it holds with probability one. The Lipschitz constant becomes the random variable $K = X$, for which $\mathbb{E}[X] = 1 < \infty$.

(ii). Let $u \in (0, 1)$ be a random element of U. Then, the function $Y(\theta) = 1/(u\theta)$ is Lipschitz continuous on $\Theta = (1, 2)$. The Lipschitz constant can be taken to be

$$K(u) = \sup_{\theta\in\Theta} |Y'(\theta)| = \sup_{\theta\in(1,2)} 1/(u\theta^2) = 1/u.$$

This holds for any $u \in (0, 1)$ drawn from the uniform distribution, thus it holds with probability one. The Lipschitz constant becomes the random variable $K = 1/U$, for which

$$\mathbb{E}[1/U] = \int_0^1 \frac{1}{u}\,du = \infty.$$

(iii). Let $x > 0$ be a random element of X, and suppose that $x \in (1, 2)$. Then

$$Y(\theta) = \sqrt{|x-\theta|} = \begin{cases} \sqrt{x-\theta}, & 1 < \theta \le x; \\ \sqrt{\theta - x}, & x \le \theta < 1. \end{cases}$$

This function is not Lipschitz. Taking $\theta = x+\epsilon$, and letting $\epsilon \downarrow 0$ results in an unbounded derivative

$$Y'(\theta) = Y'(x+\epsilon) = \frac{1}{2\sqrt{\epsilon}}.$$

This holds for all $x \in (1, 2)$, and thus with positive probability.

Bonus Problem (total 20 Credits)

Consider the one-dimensional fitting problem

$$\min_{\theta} \mathbb{E}[(\theta X - X^2)^2]$$

for finding the best "scaling" of X that produces X^2.

(a). [10 Credits] Find an SA for solving this problem.

(b). [10 Credits] Show that, in general, $\theta = \mathbb{E}[X]$ is not the correct answer.

Answer Bonus Problem:

(a) Apply IPA, which is straightforward in this case:

$$\frac{d}{d\theta}\mathbb{E}[(\theta X - X^2)^2] = \mathbb{E}[2X(\theta X - X^2)].$$

The negative gradient is coercive for this problem as we deal with minimizing a distance. The SA looks like

$$\theta_{n+1} = \theta_n - \epsilon_n 2X_n(\theta X_n - X_n^2)$$

for X_n the n-th observation. To ensure convergence we have to control the variance. As the variance of Y_n scales in θ^2, we need to use a truncation argument.

(b)

$$\mathbb{E}[(\theta X - X^2)^2] = \theta^2\mathbb{E}[X^2] - 2\theta\mathbb{E}[X^3] + \mathbb{E}[X^4]$$

Taking derivatives gives,

$$\mathbb{E}'[(\theta X - X^2)^2] = 2\theta\mathbb{E}[X^2] - 2\mathbb{E}[X^3],$$

yielding as stationary point

$$\theta = \frac{\mathbb{E}[X^3]}{\mathbb{E}[X^2]}.$$

As the the second order derivative of $\mathbb{E}[(\theta X - X^2)^2]$ is positive, this point is a minimum. The fact that we only find one stationary point, show that we have found the global minimum.