Exam including Answers: Stochastic Gradient Techniques in Optimization and Learning period 4.2

December 2020

Problem 1 (10 Credits + 5 Bonus)

Let $J(\theta) = (\theta_1 - 5)^2 + (\theta_2 - 8)^2$, $\theta \in \mathbb{R}^2$. Mapping $J(\theta)$ has the obvious global minimum $\theta_1^* = 5$ and $\theta_2^* = 8$. Let $B = \{\theta \in \mathbb{R}^2 : ||\theta|| \leq 2\}$ and consider the optimization problem

$$\min_{\theta \in B} J(\theta). \tag{1}$$

(a). [10 Credits] Write the problem in (1) in standard form, i.e., specify g and h such that (1) reads

$$\min_{\theta \in \Theta} J(\theta), \tag{2}$$
$$\Theta = \{\theta \in \mathbb{R}^d \colon g(\theta) \leq 0, h(\theta) = 0\}$$

Is the problem well-posed?

(b). [5 Credits]**[Bouns Question]** Argue that the problem

$$\min_{\theta \in \hat{B}} J(\theta),$$

with $J(\theta)$ as in (a), and $\hat{B} = \{\theta \in \mathbb{R}^2 : ||\theta|| < 12\}$ is not well-posed. How can we cast the above problem into a well-posed problem?

Answer Problem 1

(a) **Version 1 (Qualitative Argument)** J is continuous and B is compact. By the theorem of Weierstrass J attains its minimum on B at some point θ^* in B. Since B is given with a smooth boundary $g(\theta) = ||\theta|| - 2$ and J is smooth as well, the only candidates for a solution satisfy the KKT conditions. As the solution set is compact, it is not possible to have a sequence of feasible points θ_n such that $||\theta_n||$ tends to infinity.

Version 2 (Analytical Argument) Let $g(\theta) = ||\theta|| - 2$, then (1) reads in standard form

$$\min_{\theta \in \in \mathbb{R}^d} J(\theta), \tag{3}$$
$$\text{s.t.} g(\theta) \leq 0$$

The solution set is a bounded and closed set (as the ball is compact). Moreover, as $J(\theta), g(\theta) \in \mathcal{C}^1$, the KKT conditions are necessary conditions for a solution of (1). As the solution set is compact, it is not possible to have a sequence of feasible points θ_n such that $||\theta_n||$ tends to infinity. Hence, the problem is well-posed.

Version 3 (Lagrange Argument) The solution of the unconstrained problem is $(5, 8)$. As $(5, 8) \notin B$ the solution to the constrained problem is the point on the surface of B closest to $(5, 8)$. Therefore the problem is equivalent

$$\min_{\theta \in \Theta} J(\theta), \tag{4}$$
$$\Theta = \{\theta \in \mathbb{R}^d \colon g(\theta) = ||\theta|| - 2 = 0\}.$$

Then, we to solve

$$\mathcal{L}(\theta, \lambda) = J(\theta) + \lambda(g(\theta) - 2).$$

The KKT conditions come down to $\nabla \mathcal{L} = 0$ and we can compute the solution θ^* (and λ^*). As the solution set is compact, it is not possible to have a sequence of feasible points θ_n such that $||\theta_n||$ tends to infinity.

(b) The problem is ill-posed as the constraint reads $g(\theta) < 0$ for $g(\theta) = ||\theta|| - 12$. However, as the global minimum does ly inside the ball B, we can - without changing the outcome of the optimization - replace the problem by

$$\min_{\theta \in \in \mathbb{R}^d} J(\theta), \tag{5}$$
$$\text{s.t.} \hat{g}(\theta) \leq 0,$$

for $\hat{g}(\theta) \leq 0$. Then according to (a), the problem is the well-posed. Moreover, the constraint is not active at the solution, so that we can solve the problem from $J(\theta) = 0$ with obvious solution $(5, 8)$.

Problem 2 (total 20 Credits)

Recall the dynamic fitting model from the lecture. We denote by $Z(X)$ the system response to input X, where X is a random variable distributed in the experimentation range: $X \in S$. We model the system response by the mapping

$$h(\theta, x) = \theta_1 + \theta_2 x.$$

For $\theta = (\theta_1, \theta_2)^\top \in \mathbb{R}^2$, let

$$J(\theta) = \frac{1}{2}\mathbb{E}[(Z(X) - (\theta_1 + \theta_2 X))^2]$$

and the least squares optimization problem becomes

$$\min_{\theta \in \mathbb{R}^2} J(\theta).$$

Let θ^* denote the solution of the problem. Let $(x_n z_n)$ be samples of $(X, Z(X))$. For solving this problem, we apply the gradient descent from the lecture $Y_n = (z_n - \theta_{n,1} - \theta_{n,2}\, x_n)\,(1, x_n)^\top$ and the SA algorithm becomes:

$$\theta_{n+1,1} = \theta_{n,1} + \epsilon_n \, (z_n - \theta_{n,1} - \theta_{n,2} \, x_n),$$
$$\theta_{n+1,2} = \theta_{n,2} + \epsilon_n \, x_n (z_n - \theta_{n,1} - \theta_{n,2} \, x_n).$$

From the theory of SA we know that as n tends to infinity, $\theta_{n,i}$ tends in distribution to a normal distribution with mean θ_i^* and variance σ_i^2, for $i = 1, 2$. Due to time constraints you can only collect N observation pairs $(X, Z(X))$. We want to test the hypothesis that $\theta_1^* \neq 0$.

(a). [10 Credits] Design a SA algorithm for obtaining estimates of θ^* that lend themselves for carrying out a statistical analysis. How would you "optimally" allocate your computational budget?

(b). [10 Credits] Describe the actual test of the hypothesis $\theta_1^* \neq 0$ for confidence level of $\alpha = 0.05$.

Answer Problem 2 (a) Let N be the computational budget given in terms of samples of $(X, Z(X))$ available. We split N according to $N = nk$ where n is denoting the number of updates for SA and k denotes the number if iid replications for the SA algorithm. This produces as output k approximate solutions $\theta_n(\omega_i)$, $1 \leq i \leq k$. The connection between the size n and k is as follows: k should be as large as possible to produce the maximal number of samples for building confidence intervals; n should a small as possible provided that θ_n is approximate normal distributed. Moreover, in case of decreasing ϵ, we should for n observe that $\theta_n(\omega_i)$ becomes stable, and for fixed ϵ we should see that $\theta_{n+j}(\omega_i)$, $j \geq 1$, is moving around a fixed mean (= becomes a mean reverting process).
(b) Determine n, k as in (a). Then produce two confidence intervals for the components of thesolution θ_i^* for confidence level α. If $(0, 0)^\top$ is in both confidence intervals, then we cannot reject the hypothesis $H = \theta^* = 0$. More specifically, let $\bar{\theta}_{n,i}$ be the sample average over $(\theta_{n,i}(\omega_j) : 1 \leq j \leq k)$, and $st(\bar{\theta}_{n,i})$ the standard deviation. Then the confidence intervals are given by

$$(\bar{\theta}_{n,i} - z_{1-\alpha/2} st(\bar{\theta}_{n,i})/k, \bar{\theta}_{n,i} + z_{1-\alpha/2} st(\bar{\theta}_{n,i})/k)$$

for $i = 1, 2$, where z_β denotes the β quantile of the normal distribution.

Problem 3 (total 20 Credits)

We consider the problem of finding the minimum of the mapping

$$J(\theta) = 3\theta^2 + 6\theta \sin(\theta), \quad \theta \in \mathbb{R},$$

which has a unique global minimum at $\theta^* = 0$. Moreover,

$$J'(\theta) = 6\theta + 6 \sin(\theta) + 6\theta \cos(\theta), \quad \theta \in \mathbb{R}.$$

We consider an SA algorithm of the form (deterministic gradient descent with measurement noise)

$$\theta_{n+1} = \theta_n - \epsilon_n \big(J'(\theta_n) + Z_n \big) = \theta_n - \epsilon_n Y_n,$$

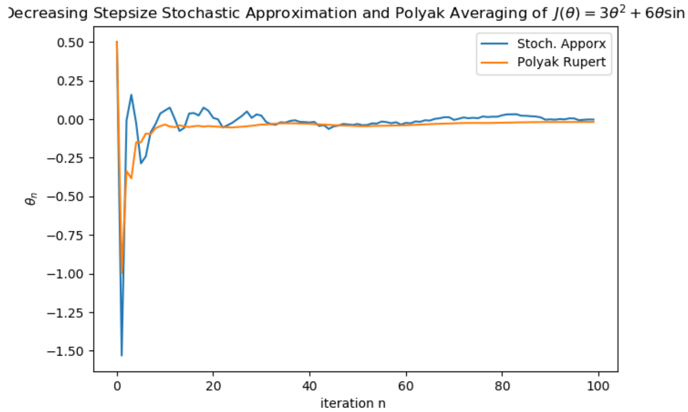

Figure 1: Decreasing Stepsize

where Z_n is the noise in the estimation of $J'(\theta_n)$, and we assume that $\{Z_n\}$ are mutually independent (not necessarily identical distributed). We denote the variance of the nth update by

$$V_n = \mathbb{E}[(Y_n - J'(\theta_n))^2 | \mathcal{F}_{n-1}] = \mathrm{Var(Z_n)}.$$

As output of the SA we consider either θ_n or the so-called *Polyak-Rupert average*

$$\bar{\theta}_n = \frac{1}{n+1} \sum_{i=0}^{n} \theta_i.$$

(a). [10 Credits] Figure 1 shows an example for $\epsilon_n = 0.25/(n+1)$ and $\theta_0 = 0.5$. The standard deviation of Z_n is set to 2.5. Explain the output: Why is the Polyak-Rupert estimate more stable than the standard algorithm? Argue that the figure illustrates the fact that the noise in θ_n asymptotically disappears, and explain why this is to be expected.

(b). [10 Credits] Figure 2 shows an example for $\epsilon_n = 0.25/(n+1)$ and $\theta_0 = 0.5$, where the standard deviation of Z_n is set to $n/2$. Can any conclusions be drawn from this output? Explain the behaviour of the algorithm.

Answer Problem 3 (a) The problem is well-posed and the negative gradient is coercive and unbiased. The noise is iid with bounded variance. Hence, θ_n converges a.s. to the true solution of the problem. In the picture θ_n becomes a stable horizontal line, which shows that the noise disappears, as expected. The PR version is more stable as the effect a deviation from the mean is damped by averaging.

(b) The variance of the gradient estimator (=vector field) is $n^2/4$. Hence, it follows from

$$\sum_n \epsilon_n^2 \frac{n^2}{4} = \infty$$

that the variance condition is violated. Indeed, for n increasing the variance gets bigger and eventually the entire outpurt of the algorithm is blurred by the noise. No answers can be drawn from this figure.

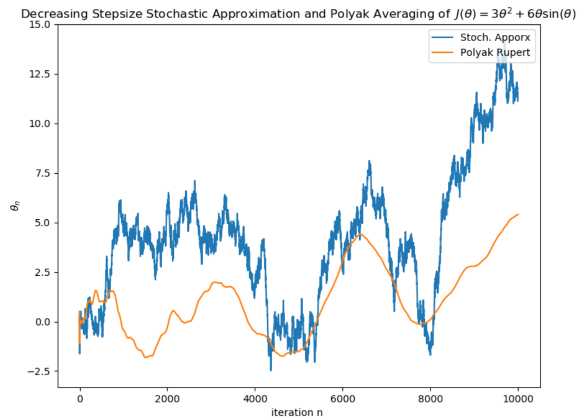

Figure 2: Decreasing Stepsize

(The only remedy is to use the variance control scheme, which is this case, however, would mean to use n^2 iid samples for update θ_n.)

Problem 4 (total 30 Credits)

Let $Z_i(\theta)$, $i = 1, 2, 3, 4$ be independent, identically distributed random variables on $\mathbb{R}$ with the normal $(\theta, 1)$ distribution for $\theta \in \mathbb{R}$. Define random variables $X_i(\theta) = e^{Z_i(\theta)}$, $i = 1, 2, 3, 4$, and define the output function

$$L\big(x_1, x_2, x_3, x_4\big) = \min\{x_1 x_2, x_3 x_4\}, \quad x_1, \ldots, x_4 \in (0, \infty).$$

The objective is to minimize the cost function

$$J(\theta) = \mathbb{E}\big[L\big(X_1(\theta), X_2(\theta), X_3(\theta), X_4(\theta)\big)\big] + \frac{1}{2}\theta^2, \tag{6}$$

with respect to θ. Figure 3 shows that there is a local minimum on $[-3, 1]$.

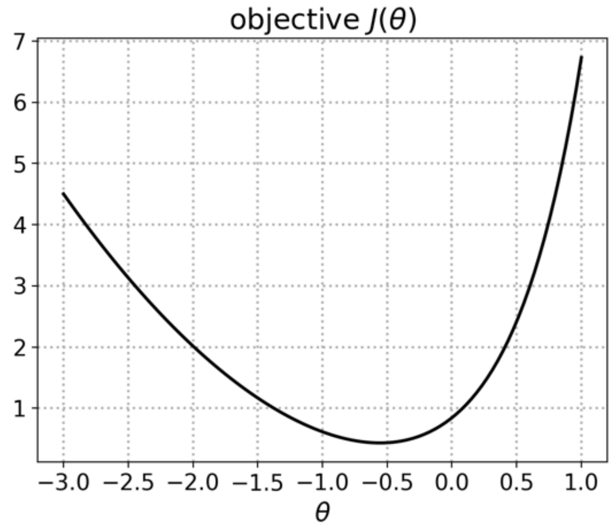

Figure 3: The objective function of Problem 4.

(a). [5 Credits] Argue that the problem is well-posed.

(b). [5 Credits] Set $G(\theta) = -dJ(\theta)/d\theta$ and argue that $G(\theta)$ is coercive for the optimization problem.

(c). [10 Credits] Compute the infinitesimal perturbation analysis (IPA) estimator for $dJ(\theta)/d\theta$.

(d). [10 Credits] Give the complete simulation algorithm (in pseudocode) that would implement the stochastic approximation iteration starting at some $\theta_0 \in (-3, 1)$ using a decreasing stepsize (starting at some $\epsilon_0 > 0$) and using the unbiased derivative estimator based on the IPA method. Include how to draw samples of the $X_i(\theta)$ samples.

Answers:

(a). The graph of the function $J(\theta)$ shows that $J(\theta)$ is continuous and at least twice differentiable, with $J(\theta) \to \infty$ when $|\theta| \to \infty$. Thus, we can find $0 < K < \infty$ such that (i) the global minimization $\min_{\theta \in \mathbb{R}} J(\theta)$ is equivalent to $\min_{-K \le \theta \le K} J(\theta)$, meaning that $[-K, K]$ contains the global minimum; and (ii) the boundaries are not active, meaning that the KKT points are the stationary points in $[-K, K]$.

Analytical approach: Note, the problem can be reduced to a simple optimization. Let $W_i, i = 1, 2, 3, 4$ be i.i.d. standard normal random variables and define

$$\alpha = \mathbb{E}\big[\min\big\{e^{W_1+W_2},\, e^{W_3+W_4}\big\}\big],$$

then $\alpha > 0$, and

$$\mathbb{E}\big[L\big(X_1(\theta), X_2(\theta), X_3(\theta), X_4(\theta)\big)\big] = \mathbb{E}\big[\min\big\{e^{Z_1(\theta)+Z_2(\theta)},\, e^{Z_3(\theta)+Z_4(\theta)}\big\}\big]$$
$$= \mathbb{E}\big[\min\big\{e^{2\theta+W_1+W_2},\, e^{2\theta+W_3+W_4}\big\}\big] = \alpha e^{2\theta}.$$

Thus $J(\theta) = \alpha e^{2\theta} + \theta^2/2$, which shows that (i) $J(\theta)$ is smooth on $\mathbb{R}$; (ii) $J(\theta)$ is convex; (iii) $\lim_{\theta\to-\infty} J(\theta) = \infty$; (iv) $\lim_{\theta\to\infty} J(\theta) = \infty$. Thus $J(\theta)$ has a unique minimum that satisfies $2\alpha e^{2\theta} + \theta = 0$.

(b). The function $J(\theta)$ is convex on $[-K, K]$, thus with $G = -J'$, any trajectory $\{\theta_n, n = 0, 1, \ldots\}$ defined by $\theta_{n+1} = \theta_n + \epsilon_n G(\theta_n)$ stays within the compact set $[-K, K]$ when the gradients remain bounded and $\epsilon_n > 0$ sufficiently small. Clearly this is the case, $|G(\theta)| \leq C < \infty$ for all $\theta \in [-K, K]$. Hence, the trajectory moves into the direction of the stable point of the ODE $dx(t)/dt = G(x(t))$ which is the stationary point.

(c). We may assume interchange of differentiation and expectation.

$$\begin{aligned}\frac{d}{d\theta}J(\theta) &= \frac{d}{d\theta}\Big(\mathbb{E}\big[L\big(X_1(\theta), X_2(\theta), X_3(\theta), X_4(\theta)\big)\big] + \frac{1}{2}\theta^2\Big)\\ &= \mathbb{E}\Big[\frac{d}{d\theta}L\big(X_1(\theta), X_2(\theta), X_3(\theta), X_4(\theta)\big)\Big] + \theta\\ &= \mathbb{E}\Big[\sum_{i=1}^{4}\frac{\partial}{\partial X_i}L\big(X_1(\theta), X_2(\theta), X_3(\theta), X_4(\theta)\big)\frac{d}{d\theta}X_i(\theta) + \theta\Big],\end{aligned}$$

with

$$\begin{aligned}\frac{\partial}{\partial x_1}L(x_1, x_2, x_3, x_4) &= x_2 I\{x_1x_2 < x_3x_4\}\\ \frac{\partial}{\partial x_2}L(x_1, x_2, x_3, x_4) &= x_1 I\{x_1x_2 < x_3x_4\}\\ \frac{\partial}{\partial x_3}L(x_1, x_2, x_3, x_4) &= x_4 I\{x_3x_4 < x_1x_2\}\\ \frac{\partial}{\partial x_4}L(x_1, x_2, x_3, x_4) &= x_3 I\{x_3x_4 < x_1x_2\}\\ \frac{d}{d\theta}X_i(\theta) = \frac{d}{d\theta}e^{Z_i(\theta)} &= e^{Z_i(\theta)}\frac{d}{d\theta}Z_i(\theta) = X_i(\theta),\end{aligned}$$

because $Z_i(\theta) = \theta + N(0, 1)$. Hence, the IPA estimator is

$$\begin{aligned}D &= \begin{cases}2X_1(\theta)X_2(\theta) + \theta & \text{if } X_1(\theta)X_2(\theta) < X_3(\theta)X_4(\theta)\\ 2X_3(\theta)X_4(\theta) + \theta & \text{if } X_3(\theta)X_4(\theta) < X_1(\theta)X_2(\theta)\end{cases}\\ &= 2\min\big\{X_1(\theta)X_2(\theta), X_3(\theta)X_4(\theta)\big\} + \theta.\end{aligned}$$

(d). This is the algorithm with a single sample of the IPA estimator D per iteration.

Algorithm 1 Stochastic Approximation Iteration

Require: θ_0 {*initial point*}
Require: ϵ {*initial stepsize*}
Require: M {*number of iterations*}
for $n = 0$ **to** $M - 1$ **do**
  for $i = 1$ **to** 4 **do**
    Generate $W_i \sim N(0,1)$ {*standard normal*}
    $Z_i \leftarrow \theta_n + W_i$
    $X_i \leftarrow \exp(Z_i)$
  end for
  Compute $L = \min\{X_1 X_2, X_3 X_4\}$ {*output function*}
  $D \leftarrow 2L + \theta_n$ {*ipa estimator*}
  $\theta_{n+1} = \theta_n - \epsilon/(n+1) \times D$
end for
return θ_M

Problem 5

Let $Z(\theta)$ be a random variable on $\mathbb{R}$, parameterized by $\theta \in \Theta$. Denote its probability density function (PDF) by $f_Z(z;\theta)$. Assume that f_Z is differentiable with respect to θ for all $z \in \mathbb{R}$.

Define a random variable $X(\theta) = h\big(Z(\theta)\big)$ for a differentiable monotone function $h : \mathbb{R} \to \mathbb{R}$. Recall from Probability Theory that $X(\theta)$ has PDF

$$f_X(x;\theta) = f_Z\big(h^{-1}(x);\theta\big)\,\frac{d}{dx}h^{-1}(x).$$

(a). Let $Z(\theta)$ have the normal $(\theta, 1)$ distribution, and $X(\theta) = e^{Z(\theta)}$, where $\theta \in \Theta = \mathbb{R}$. Recall

$$f_Z(z;\theta) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}(z-\theta)^2}, \quad z \in \mathbb{R}.$$

Compute the score functions $S_Z\big(\theta; Z(\theta)\big)$ and $S_X\big(\theta; X(\theta)\big)$.

(b). Refer to Problem 4 for the objective function $J(\theta)$ in display (6). Compute the score function estimator for $dJ(\theta)/d\theta$.

(c). [**Bonus**]. Let $Z(\theta)$ have some general distribution with a given score function $S_Z\big(\theta; Z(\theta)\big)$. Now find the expression for the score function of $X(\theta) = h\big(Z(\theta)\big)$.

Answers:

(a). The score function of $Z(\theta)$:

$$S_Z(\theta; z) = \frac{d}{d\theta}\log f_Z(z;\theta) = \frac{d}{d\theta}\Big(-\log\sqrt{2\pi} - \frac{1}{2}(z-\theta)^2\Big) = z - \theta.$$

Thus $S_Z\big(\theta; Z(\theta)\big) = Z(\theta) - \theta$.

The score function of $X(\theta) = e^{Z(\theta)}$. For $h(z) = e^z$, the inverse is $h^{-1}(x) = \log x$, thus the PDF of $X(\theta)$ is

$$f_X(x;\theta) = f_Z\big(h^{-1}(x);\theta\big)\,\frac{d}{dx}h^{-1}(x) = \frac{1}{x\sqrt{2\pi}}\,e^{-\frac{1}{2}(\log x-\theta)^2},\ x>0.$$

From this,

$$S_X(\theta;x) = \frac{d}{d\theta}\log f_X(x;\theta) = \frac{d}{d\theta}\Big(-\log(x\sqrt{2\pi}) - \frac{1}{2}(\log x-\theta)^2\Big) = \log x - \theta.$$

Thus $S_X\big(\theta; X(\theta)\big) = \log X(\theta) - \theta$.

(b). First, the $X_i(\theta)$'s are i.i.d., thus their joint PDF is the product

$$f(x_1,x_2,x_3,x_4;\theta) = \prod_{i=1}^{4} f_X(x_i;\theta).$$

Second, the score function of this PDF is

$$\frac{d}{d\theta}\log f(x_1,x_2,x_3,x_4;\theta) = \frac{d}{d\theta}\sum_{i=1}^{4}\log f_X(x_i;\theta) = \sum_{i=1}^{4} S_X(\theta;x_i)$$
$$= \sum_{i=1}^{4}(\log x_i - \theta) = \sum_{i=1}^{4}\log x_i - 4\theta = \log\prod_{i=1}^{4} x_i - 4\theta.$$

Hence, for the score function estimator of $dJ(\theta)/\theta$ we assume interchange of differentiation and integration to get.

$$\frac{d}{d\theta}J(\theta) = \frac{d}{d\theta}\Big(\mathbb{E}\big[L\big(X_1(\theta),X_2(\theta),X_3(\theta),X_4(\theta)\big)\big] + \frac{1}{2}\theta^2\Big)$$
$$= \frac{d}{d\theta}\int L(x_1,x_2,x_3,x_4)\,f(x_1,x_2,x_3,x_4;\theta)\,dx_1dx_2dx_3dx_4 + \theta$$
$$= \int L(x_1,x_2,x_3,x_4)\,\frac{d}{d\theta}f(x_1,x_2,x_3,x_4;\theta)\,dx_1dx_2dx_3dx_4 + \theta$$
$$= \int L(x_1,x_2,x_3,x_4)\,\Big(\frac{d}{d\theta}\log f(x_1,x_2,x_3,x_4;\theta)\Big)f(x_1,x_2,x_3,x_4;\theta)\,dx_1dx_2dx_3dx_4 + \theta$$
$$= \int L(x_1,x_2,x_3,x_4)\,\big(\log\prod_{i=1}^{4}x_i - 4\theta\big)f(x_1,x_2,x_3,x_4;\theta)\,dx_1dx_2dx_3dx_4 + \theta$$
$$= \mathbb{E}\big[L\big(X_1(\theta),X_2(\theta),X_3(\theta),X_4(\theta)\big)\,\big(\log\prod_{i=1}^{4}X_i(\theta) - 4\theta\big)\big] + \theta.$$

The estimator is

$$D = L\big(X_1(\theta),X_2(\theta),X_3(\theta),X_4(\theta)\big)\big(\log\prod_{i=1}^{4}X_i(\theta) - 4\theta\big) + \theta$$
$$= \min\big\{X_1(\theta)X_2(\theta)\,X_3(\theta)X_4(\theta)\big\}\,\big(\log\prod_{i=1}^{4}X_i(\theta) - 4\theta\big) + \theta.$$

(c). Using the definition of score function and the transformation formula,

$$
\begin{aligned}
S_X(\theta; x) &= \frac{d}{d\theta} \log f_X(x;\theta) = \frac{d}{d\theta} \log\Big(f_Z\big(h^{-1}(x);\theta\big) \frac{d}{dx} h^{-1}(x)\Big) \\
&= \frac{d}{d\theta}\Big(\log f_Z\big(h^{-1}(x);\theta\big) + \log \frac{d}{dx} h^{-1}(x)\Big) \\
&= \frac{d}{d\theta} \log f_Z\big(h^{-1}(x);\theta\big) = S_Z\big(\theta; h^{-1}(x)\big).
\end{aligned}
$$

Thus,

$$
S_X\big(\theta; X(\theta)\big) = S_Z\big(\theta; h^{-1}\big(X(\theta)\big)\big).
$$