Computational Statistics : 1st project.

Cardoso Ruben

1 Exercise 1:

1.1

For i.i.d $X_1, .., X_N \overset{law}{=} \mathcal{U}(0, \theta)$, it's density is $p_\theta(x) = \frac{1}{\theta}\mathbf{1}_{[0,\theta]}(x)$. Hence

$$\mathbb{E}_\theta[X_1] = \int_{\mathbb{R}} x\, p_\theta(x)\, dx = \int_0^\theta x\, \frac{1}{\theta}\, dx = \frac{1}{\theta}\left[\tfrac{x^2}{2}\right]_0^\theta = \frac{\theta}{2}.$$

Moreover, since $(X_k)_{k\leq N}$ are i.i.d. and square–integrable, the Strong Law of Large Numbers gives

$$\bar{X}_N \;=\; \frac{1}{N}\sum_{k=1}^{N} X_k \underset{N\to\infty}{\overset{a.s}{\to}} \mathbb{E}_\theta[X_1] = \frac{\theta}{2}$$

so by continuity,

$$\hat{\theta}_1 = 2\bar{X}_n \xrightarrow{\text{a.s.}} \theta,$$

i.e., $\hat{\theta}_1$ is a consistent estimator of θ.

1.2

We have $\mathbb{E}_\theta[\hat{\theta}_1] = \theta$ so $\hat{\theta}_1$ is unbiased. Under squared loss, the quadratic risk is thus

$$R(\theta, \hat{\theta}_1) = \mathbb{E}_\theta\left[(\hat{\theta}_1 - \theta)^2\right] = \mathrm{Var}_\theta(\hat{\theta}_1).$$

Computing the variance:

$$\mathrm{Var}_\theta(\hat{\theta}_1) = \mathrm{Var}_\theta\left(\frac{2}{N}\sum_{k=1}^{N} X_k\right) = \frac{4}{N^2}\sum_{k=1}^{N}\mathrm{Var}_\theta(X_k) \overset{\text{i.i.d.}}{=} \frac{4}{N}\,\mathrm{Var}_\theta(X_1).$$

Now :

$$\mathbb{E}_\theta[X_1^2] = \int_0^\theta x^2\, \frac{1}{\theta}\, dx = \frac{1}{\theta}\left[\tfrac{x^3}{3}\right]_0^\theta = \frac{\theta^2}{3}$$

so

$$\mathrm{Var}_\theta(X_1) = \mathbb{E}_\theta[X_1^2] - \big(\mathbb{E}_\theta[X_1]\big)^2 = \frac{\theta^2}{3} - \frac{\theta^2}{4} = \frac{\theta^2}{12}.$$

Therefore,

$$R(\theta, \hat{\theta}_1) = \mathrm{Var}_\theta(\hat{\theta}_1) = \frac{4}{N}\cdot\frac{\theta^2}{12} = \frac{\theta^2}{3N}.$$

1.3

We consider the likelihood function based on independent samples $X_1, \dots, X_N$ from $\mathcal{U}(0, \theta)$:

$$L(X_1, \dots, X_N; \theta) = \prod_{i=1}^{N} p_\theta(X_i) = \prod_{i=1}^{N}\frac{1}{\theta}\,\mathbf{1}_{[0,\theta]}(X_i) = \frac{1}{\theta^N}\,\mathbf{1}_{\{\theta\geq \max_i X_i\}}.$$

Hence, the likelihood is positive if and only if $\theta \geq X_{(N)}$, where $X_{(n)} = \max\{X_1, \dots, X_N\}$. Since L is decreasing in θ for $\theta \geq X_{(N)}$, the first θ that maximizes L is

$$\hat{\theta}_2 = X_{(N)}$$

1.4

For any $t \in [0, \theta]$,

$$\mathbb{P}_\theta(X_{(N)} \leq t) = \mathbb{P}_\theta(X_1 \leq t, \dots, X_N \leq t) = \big(\mathbb{P}_\theta(X_1 \leq t)\big)^N = \left(\frac{t}{\theta}\right)^N.$$

Differentiating with respect to t, we get its density:

$$f_{X_{(N)}}(t) = \frac{d}{dt}\left(\frac{t}{\theta}\right)^N = \frac{N\, t^{N-1}}{\theta^N}, \quad t \in [0, \theta].$$

So:

$$\mathbb{E}_\theta[X_{(N)}] = \int_0^\theta t\, f_{X_{(N)}}(t)\, dt = \int_0^\theta t\, \frac{Nt^{N-1}}{\theta^N}\, dt = \frac{N}{\theta^N}\, \frac{\theta^{N+1}}{N+1} = \frac{N}{N+1}\, \theta.$$

Similarly,

$$\mathbb{E}_\theta[X_{(N)}^2] = \int_0^\theta t^2\, f_{X_{(N)}}(t)\, dt = \frac{N}{\theta^N}\, \frac{\theta^{N+2}}{N+2} = \frac{N}{N+2}\, \theta^2.$$

Then the variance is

$$\mathrm{Var}_\theta(X_{(N)}) = \mathbb{E}_\theta[X_{(N)}^2] - \big(\mathbb{E}_\theta[X_{(N)}]\big)^2 = \theta^2\left(\frac{N}{N+2} - \frac{N^2}{(N+1)^2}\right) = \frac{N\,\theta^2}{(N+1)^2(N+2)}.$$

Since $\hat{\theta}_2$ is biased, the risk is

$$R(\theta, \hat{\theta}_2) = \mathrm{Var}_\theta(\hat{\theta}_2) + \big(\mathrm{Bias}_\theta(\hat{\theta}_2)\big)^2.$$

We have $\mathrm{Bias}_\theta(\hat{\theta}_2) = \mathbb{E}_\theta[X_{(N)}] - \theta = -\frac{\theta}{N+1}$, hence

$$R(\theta, \hat{\theta}_2) = \frac{N\,\theta^2}{(N+1)^2(N+2)} + \left(\frac{\theta}{N+1}\right)^2 = \frac{\theta^2}{(N+1)^2}\left(\frac{N}{N+2} + 1\right) = \frac{2\,\theta^2}{(N+1)(N+2)}.$$

1.5

We compare the quadratic risks of $\hat{\theta}_1$ and $\hat{\theta}_2$:

$$R_1(\theta) = \frac{\theta^2}{3N}, \qquad R_2(\theta) = \frac{2\theta^2}{(N+1)(N+2)}.$$

Since $\theta^2 > 0$, the sign of $R_1(\theta) - R_2(\theta)$ depends only on the fractions:

$$R_1 - R_2 \overset{\text{sign}}{=} \frac{1}{3N} - \frac{2}{(N+1)(N+2)}.$$

The sign thus depends on the numerator polynomial

$$P(N) = N^2 - 3N + 2 = (N-1)(N-2).$$

Conclusion. As long as sample size $N \geq 3$, the maximum–likelihood estimator $\hat{\theta}_2 = X_{(N)}$ has a smaller quadratic risk than the unbiased estimator $\hat{\theta}_1 = 2\bar{X}_N$, and is therefore more efficient. Otherwise, $\hat{\theta}_1$ has a smaller risk.

2 Exercise 2

2.1 2.1

Define the mapping

$$\Psi: \ (r, \theta) \mapsto (x, y) = (r\cos\theta,\ r\sin\theta).$$

Its Jacobian matrix and determinant are

$$D\Psi(r, \theta) = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}, \qquad \big|\det D\Psi(r, \theta)\big| = r.$$

The inverse transform is $r = \sqrt{x^2 + y^2}$ and $\theta = \mathrm{atan2}(y, x) \in [0, 2\pi[$.[1]

[1]The function $\mathrm{atan2}(y, x)$ returns the angle $\Theta \in (-\pi, \pi]$ such that $\tan(\Theta) = \frac{y}{x}$ and $\mathrm{sign}(\cos\Theta) = \mathrm{sign}(\sin\Theta)$.

Since $R \perp\!\!\!\perp \Theta$, the joint density of (R, Θ) is

$$f_{R,\Theta}(r,\theta) = f_R(r) f_\Theta(\theta) = \frac{1}{2\pi}\, r\, e^{-r^2/2}\, \mathbf{1}_{\{r>0\}} \mathbf{1}_{[0,2\pi]}(\theta).$$

By the change-of-variables formula,

$$f_{X,Y}(x,y) = f_{R,\Theta}(\Psi(R,\Theta))\, \frac{1}{|\det D\Psi|} = \frac{1}{2\pi} e^{-(x^2+y^2)/2}, \quad (x,y) \in \mathbb{R}^2.$$

This factorizes as

$$f_{X,Y}(x,y) = \left(\frac{1}{\sqrt{2\pi}} e^{-x^2/2}\right)\left(\frac{1}{\sqrt{2\pi}} e^{-y^2/2}\right) \overset{law}{=} \mathcal{N}(0,1) \otimes \mathcal{N}(0,1),$$

so $X, Y \overset{law}{=} \mathcal{N}(0,1)$, $Y \sim \mathcal{N}(0,1)$, and are independent.

2.2 2.2

We recall that R follows a Rayleigh distribution with density

$$f_R(r) = r e^{-r^2/2} \mathbf{1}_{\{r>0\}}.$$

The corresponding cumulative distribution function is

$$F_R(r) = \int_0^r s e^{-s^2/2}\, ds = 1 - e^{-r^2/2}.$$

To simulate R, we use the inverse transform method. Let $V \overset{law}{=} \mathcal{U}(0,1)$, then

$$V = F_R(r) = 1 - e^{-r^2/2} \quad \Rightarrow \quad r = \sqrt{-2\ln(1-V)}.$$

Since $(1-V) \overset{law}{=} V$, we can equivalently write

$$R \overset{law}{=} \sqrt{-2\ln(V)}.$$

We now draw another independent variable $\Theta \overset{law}{=} \mathcal{U}(0, 2\pi)$ and define

$$X = R\cos(\Theta), \qquad Y = R\sin(\Theta).$$

Then, as proved previously, X and Y are independent standard normal random variables.

Algorithm 1 Simulation of $(X,Y) \overset{law}{=} \mathcal{N}(0,1) \otimes \mathcal{N}(0,1)$

Require: n ▷ Number of Gaussian pairs to generate

for $i = 1$ to n **do**

  Simulate $U_1, U_2 \overset{law}{=} \mathcal{U}(0,1)$ independently

  Compute

$$\Theta = 2\pi U_1, \qquad R = \sqrt{-2\ln(U_2)}.$$

  Then set

$$X_i = R\cos(\Theta), \qquad Y_i = R\sin(\Theta).$$

  Output (X_i, Y_i)

end for

2.3

The couple (V_1, V_2) lives in the unit ball : $B(0,1)$, so it follows $\mathcal{U}(B(0,1))$.

2.4

Set

$$M := \inf\Big\{n \geq 1 : \ \|(V_1^{(n)}, V_2^{(n)})\|_2 \leq 1 \text{ and } \|(V_1^{(i)}, V_2^{(i)})\|_2 > 1, \ \forall i < n\Big\}.$$

Each trial is accepted with probability

$$p = \mathbb{P}\big(\|(V_1, V_2)\|_2 \leq 1\big) = \frac{\text{Area}(B(0,1))}{\text{Area}([-1,1]^2)} = \frac{\pi}{4}.$$

Hence $M \sim \text{Geom}(p)$ and

$$\mathbb{E}[M] = \frac{1}{p} = \frac{4}{\pi} \approx 1.27.$$

On average, about 2 iterations are required before acceptance.

2.5

We know that (V_1, V_2) is uniformly distributed over the unit disk :

$$f_{V_1,V_2}(v_1, v_2) = \frac{1}{\pi}\mathbf{1}_{\{v_1^2+v_2^2 \leq 1\}}.$$

We introduce the polar transformation

$$R = \sqrt{V_1^2 + V_2^2}, \qquad \Theta = \text{atan2}(V_2, V_1),$$

so that $V_1 = R\cos\Theta$ and $V_2 = R\sin\Theta$. The Jacobian of this transformation is R, as shown previously. Therefore, the joint density of (R, Θ) is

$$f_{R,\Theta}(r, \theta) = f_{V_1,V_2}(r\cos\theta, r\sin\theta)\,|J| = \frac{1}{\pi} r\,\mathbf{1}_{0\leq r\leq 1}\,\mathbf{1}_{0\leq\theta<2\pi}.$$

We can derive the marginal densities:

$$f_R(r) = \int_0^{2\pi} f_{R,\Theta}(r,\theta)\,d\theta = \frac{2\pi}{\pi}\,r\,\mathbf{1}_{[0,1]}(r) = 2r\,\mathbf{1}_{[0,1]}(r),$$

and

$$f_\Theta(\theta) = \int_0^1 f_{R,\Theta}(r,\theta)\,dr = \frac{1}{2\pi}\,\mathbf{1}_{[0,2\pi)}(\theta).$$

Hence, R and Θ are independent, with

$$R \text{ having density } f_R(r) = 2r \text{ on } [0,1], \qquad \Theta \overset{law}{=} \mathcal{U}([0, 2\pi)).$$

Now, define

$$V = V_1^2 + V_2^2 = R^2.$$

We obtain the density of V through the transformation $v = r^2$:

$$f_V(v) = f_R(\sqrt{v})\left|\frac{dr}{dv}\right| = 2\sqrt{v}\,\frac{1}{2\sqrt{v}} = \mathbf{1}_{(0,1)}(v).$$

Thus,

$$V \overset{law}{=} \mathcal{U}(0, 1).$$

Next, we set

$$T_1 = \frac{V_1}{\sqrt{V_1^2 + V_2^2}}.$$

Since $\Theta \overset{law}{=} \mathcal{U}([0, 2\pi))$, we can compute the density of T_1 by the change of variable $t = \cos\theta$. For $t \in]-1, 1[$, let's be careful since the two solutions are $\theta_1 = \arccos t$ and $\theta_2 = 2\pi - \arccos t$. Hence,

$$f_{T_1}(t) = \sum_{k=1}^{2} f_\Theta(\theta_k)\left|\frac{d\theta_k}{dt}\right| = \frac{1}{2\pi}\left(\frac{1}{\sqrt{1-t^2}} + \frac{1}{\sqrt{1-t^2}}\right) = \frac{1}{\pi\sqrt{1-t^2}}\,\mathbf{1}_{(-1,1)}(t).$$

Therefore, T_1 has the same distribution as $\cos(\Theta)$ with $\Theta \overset{law}{=} \mathcal{U}([0, 2\pi))$.

Finally, since $V = R^2$ depends only on R and $T_1 = \cos\Theta$ depends only on Θ, and R and Θ are independent, we conclude that

$$T_1 \perp V.$$

2.6

Define

$$S := \sqrt{-2\log V}.$$

Since $V \overset{\text{law}}{=} \mathcal{U}(0,1)$, the inverse-transform method yields

$$S \overset{\text{law}}{=} \text{Rayleigh}, \qquad S \perp\!\!\!\perp \Theta.$$

Moreover, we have $T_1 \overset{\text{law}}{=} \cos\Theta$, and $T_1^2 + T_2^2 = 1$, the algorithm outputs :

$$X = S\,\frac{V_1}{\sqrt{V_1^2 + V_2^2}} = S\,T_1 \overset{\text{law}}{=} S\cos\Theta, \qquad Y = S\sqrt{1 - T_1^2} \overset{\text{law}}{=} S\sin\Theta.$$

so we necessarily have by first questions that : $(X,Y) \overset{law}{=} \mathcal{N}(0,1) \otimes \mathcal{N}(0,1)$.

3 Exercise 3:

3.1

We aim to minimize the empirical loss

$$L_n(w) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \langle w, x_i\rangle)^2,$$

where each observation (x_i, y_i) satisfies $x_i \in [0,1]^d$ and $y_i \in \{-1, 1\}$. Since $[0,1]^d$ is compact, the sequence $\{x_i\}$ is bounded by $\|x_i\| \leq R = \sqrt{d}$. The gradient of L_n is given by

$$\nabla L_n(w) = -\frac{2}{n}\sum_{i=1}^{n}(y_i - \langle w, x_i\rangle)x_i^T,$$

and its Hessian is constant, $\nabla^2 L_n(w) = \frac{2}{n}\sum_{i=1}^{n} x_i x_i^\top$. The spectral norm of each $x_i x_i^\top$ is bounded by $\|x_i\|^2 \leq R^2$, hence $\|\nabla^2 L_n(w)\|_2 \leq 2R^2$. Therefore, ∇L_n is $2R^2$-Lipschitz, which ensures that a stochastic gradient descent procedure is stable. Moreover, since both y_i and x_i are bounded, the gradient itself is bounded on $\mathbb{R}^d$ by

$$\|\nabla L_n(w)\| \leq \frac{2}{n}\sum_{i=1}^{n} |y_i - \langle w, x_i\rangle|\,\|x_i\| \leq 2(1 + \|w\|R)R$$

Moreover, $\nabla_w L$ is convex. We use the mini-batch version of the stochastic gradient descent algorithm, which updates the iterate as

$$w_{k+1} = w_k - \eta_k\, g_k, \qquad g_k = -\frac{2}{|B_k|}\sum_{i\in B_k}(y_i - \langle w_k, x_i\rangle)x_i,$$

where B_k denotes a batch of indices uniformly drawn without replacement among $\{1, ..., n\}$. The step size sequence $\eta_k = k^{-0.8}$ satisfies $\sum_k \eta_k^2 < \infty$ and $\sum_k \eta_k = \infty$, which are the classical conditions ensuring almost sure convergence toward a minimizer $w^\star$ of L_n, under the assumptions that the gradient estimator is unbiased and its variance is bounded. We also have :

$$\mathbb{E}[\nabla L_n(w_k)] = \nabla L_n(w_k), \quad \mathbb{E}\big[\|\nabla \ell_i(w_k)\|^2\big] < \infty$$

The resulting mini-batch SGD algorithm can thus be written as follows:

Initialize: $w_0 = 0, \quad k = 1.$

Repeat:

Draw a random batch $B_k \subset \{1, \ldots, n\}, \ |B_k| = m.$

Compute $g_k = -\frac{2}{m}\sum_{i\in B_k}(y_i - x_i^\top w_k)x_i.$

Update $w_{k+1} = w_k - \eta_k g_k, \quad \eta_k = k^{-0.8}.$

$k \leftarrow k + 1.$

Until convergence.