SLDM Homework 1

Dave Zachariah

January 24, 2023

- Solution proposals are individual.
- Each solution must be reproducible by your peer. Code should be added to the appendix.
- The solution to each subproblem yield 0 to 2 points.

1

Let us consider the ship localization problem with data drawn from target distribution $p(\mathbf{z})$. The reported ship location is parameterized as

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_x \\ \theta_y \end{bmatrix} \in \Theta$$

We consider the following loss function $\ell_\theta(\mathbf{z}) = \|\mathbf{z} - \boldsymbol{\theta}\|^2$ when reporting a location $\boldsymbol{\theta}$.

a) Show that the risk (aka. expected loss) of $\boldsymbol{\theta}$ can be expressed as

$$L(\boldsymbol{\theta}) = \mathbb{E}[\ell_\theta(\mathbf{z})] = \boldsymbol{\theta}^\top \boldsymbol{\theta} - 2\boldsymbol{\theta}^\top \mathbb{E}[\mathbf{z}] \quad + \mathbb{E}[\mathbf{z}^\top \mathbf{z}],$$

where the last term is a constant, and show therefore there is a unique ship location target

$$\boldsymbol{\theta}_\circ = \mathbb{E}[\mathbf{z}]$$

when the parameter space is the entire plane $\Theta = \mathbb{R}^2$

b) For notational simplicity, let

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{z}] \qquad \boldsymbol{\Sigma} = \mathbb{V}[\mathbf{z}]$$

denote the mean and covariance matrix of $\mathbf{z}$ under p. Plot $\boldsymbol{\theta}_\circ$ and the contours of $L(\boldsymbol{\theta}, p)$ (skip constants) and over a parameter space $\Theta = [0, 400] \times [0, 400]$ in units [m] when

$$\boldsymbol{\mu} = \begin{bmatrix} 200 \\ 200 \end{bmatrix}$$

c) Suppose the target distribution $p(\mathbf{z})$ is a multivariate t-distribution $t_\nu(\boldsymbol{\mu}, \mathbf{C})$. What happens to $L(\boldsymbol{\theta})$ when the unknown degrees of freedom change as $\nu = 3, 2, 1$?

d) Consider i.i.d. training data $\mathbf{z}^n$ in from the target distributon $p(\mathbf{z})$.

Show that when using the empirical distribution, we have

$$L(\boldsymbol{\theta}, p_n) = \mathbb{E}_n[\ell_\theta(\mathbf{z})] = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{z}_i - \boldsymbol{\theta}\|^2$$

and that the empirical risk minimizer is

$$\widehat{\boldsymbol{\theta}}_n = \mathbb{E}_n[\mathbf{z}]$$

when the parameter space is the entire plane $\Theta = \mathbb{R}^2$

e) Suppose the unknown data-generating process corresponds to a multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{bmatrix} 400 & 50 \\ 50 & 400 \end{bmatrix}$$

Draw a sample $\underline{\mathbf{z}}^n$ and plot contours of the resulting empirical risk $L(\boldsymbol{\theta}, p_n)$ along with the best parameter $\boldsymbol{\theta}_\circ$ and the learned parameters

$$\{\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_{10}, \widehat{\boldsymbol{\theta}}_{100}, \widehat{\boldsymbol{\theta}}_{1000}\}$$

f) In the above problem, the learned parameter $\widehat{\boldsymbol{\theta}}_n$ could be numerically evaluated using a closed-form expression. Such expressions are not available in many problems, but instead the minimizer of $L(\boldsymbol{\theta}, p_n)$ must be approximated using a numerical search methods, e.g.:

$$\widehat{\boldsymbol{\theta}}^{(k+1)} = \widehat{\boldsymbol{\theta}}^{(k)} - \gamma_k \mathbf{g}_k \qquad \text{where } \gamma_k \geq 0$$

A common choice is gradient descent methods, where

$$\mathbf{g}_k = \partial_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, p_n)|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(k)}}$$

is the gradient of the empirical risk.

Implement a gradient descent method with a fixed γ to approximate the learned parameter $\widehat{\boldsymbol{\theta}}_n$ and plot the approximations using the same data as above. Initialize at $\widehat{\boldsymbol{\theta}}^{(0)} = \mathbf{0}$.

g) Show that the MSE-matrix of the learned parameter $\widehat{\boldsymbol{\theta}}_n$ over all possible training datasets equals

$$\mathbf{M} = \mathbb{E}[(\boldsymbol{\theta}_\circ - \widehat{\boldsymbol{\theta}}_n)(\boldsymbol{\theta}_\circ - \widehat{\boldsymbol{\theta}}_n)^\top] = \frac{1}{n}\boldsymbol{\Sigma}$$

Tip: Note that the bias of $\widehat{\boldsymbol{\theta}}_n$ is $\mathbf{0}$.

h) We begin to study the large-sample behaviour of the ERM method using the fact that $L(\boldsymbol{\theta}, p)$ admits a Taylor-series approximation. Show that the Hessian of $L(\boldsymbol{\theta}, p)$ at $\boldsymbol{\theta}_\circ$ is

$$\mathbf{Q} = 2\mathbf{I}_2$$

and that the sensitivity score with covariance matrix equals

$$\dot{\boldsymbol{\ell}}_\circ(\mathbf{z}) = 2(\boldsymbol{\theta}_\circ - \mathbf{z}) \qquad \mathbf{L} = 4\boldsymbol{\Sigma}$$

We can conclude that the errors of $\widehat{\boldsymbol{\theta}}_n$ approach a zero-mean Gaussian distribution at a $\sqrt{n}$-rate:

$$\sqrt{n}(\boldsymbol{\theta}_\circ - \widehat{\boldsymbol{\theta}}_n) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

By re-arranging the LHS, this means, somewhat informally, that $\widehat{\boldsymbol{\theta}}_n$ approaches a distribution with mean $\boldsymbol{\theta}_\circ$ (i.e. vanishing bias) and covariance $\frac{1}{n}\boldsymbol{\Sigma}$. Compare this with MSE-matrix above.

2

We want to study the association between blood pressure y and covariates $\mathbf{x}$, specifically to evaluate a particular medication. The data-generating process is a randomized trial, where

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ y \end{bmatrix}$$

a) We consider the target parameters to be those that yield the 'best' linear predictor of outcome y given covariates $\mathbf{x}$. Using the squared-error loss $\ell_\theta(\mathbf{z}) = (y - \mathbf{x}^\top \boldsymbol{\theta})^2$, show that the target parameters are given by

$$\begin{aligned} \Theta_\circ &= \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, p) \\ &= \arg\min_{\boldsymbol{\theta}} \ \boldsymbol{\theta}^\top \mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right] \boldsymbol{\theta} - 2\boldsymbol{\theta}^\top \mathbb{E}[\mathbf{x}y] \\ &= \left\{\boldsymbol{\theta} : \mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right] \boldsymbol{\theta} = \mathbb{E}[\mathbf{x}y]\right\} \end{aligned} \tag{1}$$

b) A generalized inverse $\mathbf{A}^-$ to a matrix $\mathbf{A}$ is any matrix that satisfies $\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$. Show that using any generalized inverse $\mathbb{E}\left[\boldsymbol{\phi}\boldsymbol{\phi}^\top\right]^-$, we have that

$$\Theta_\circ = \left\{\mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right]^- \mathbb{E}[\mathbf{x}y]\right\}$$

Conclude that a target parameter $\boldsymbol{\theta}_\circ$ is point-identifiable when $\mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right]$ has full rank.

Hint: Note that $\mathbb{E}[\mathbf{x}y]$ belongs to the column space of $\mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right]$.

c) Suppose

$$\mathbf{x} = \begin{bmatrix} 1 \\ x \end{bmatrix}$$

where x quantifies a drug dosage (which varies so that indeed $\mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right] \succ \mathbf{0}$). Show that the target parameter can be written as:

$$\boldsymbol{\theta}_\circ = \begin{bmatrix} \mathbb{E}[y] - \frac{\mathbb{V}[x,y]}{\mathbb{V}[x]} \mathbb{E}[x] \\ \frac{\mathbb{V}[x,y]}{\mathbb{V}[x]} \end{bmatrix},$$

where $\mathbb{V}[x, y]$ denotes the *covariance* between drug dosage x and blood pressure y.

d) The target parameter

$$\theta_{\circ,2} = \frac{\mathbb{V}[x, y]}{\mathbb{V}[x]}$$

is often understood as the 'average treatment effect' of the drug on blood pressure. How would you justify this interpretation in this randomized trial process?