Math Foundations of ML, Fall 2017

Homework #9

Due Monday November 20, at the beginning of class

As stated in the syllabus, unauthorized use of previous semester course materials is strictly prohibited in this course.

1. Using you class notes, prepare a 1-2 paragraph summary of what we talked about in class in the last week. I do not want just a bulleted list of topics, I want you to use complete sentences and establish context (Why is what we have learned relevant? How does it connect with other things you have learned here or in other classes?). The more insight you give, the better.

2. We say that random variable X in $\mathbb{R}$ has a *two-sided Laplacian* distribution if its pdf is

$$f_X(x;\mu) = \frac{1}{2}e^{-|x-\mu|},$$

for some parameter $\mu \in \mathbb{R}$. Suppose we observe a series of independent and identically distributed two-sided Laplacian random variables $X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N$. Given these observations, how do we form the maximum likelihood estimate for μ? Justify your answer thoroughly.

You might want to start thinking about the particular cases $N = 2$ and $N = 3$ (your answer should be slightly different for N even and N odd).

Hint: the function $h(x) = |x - \mu|$ is differentiable everywhere except at $x = \mu$. But to conjure the answer to this problem (which you then might justify in many different ways), image approximating this function by replacing the "sharp point" in a very small interval around $x = \mu$ with something that is differentiable everywhere and has zero derivative at $x = \mu$.

3. Another method we can use to solve a constrained optimization program

$$\underset{\boldsymbol{x}\in\mathbb{R}^D}{\text{minimize}}\ f(\boldsymbol{x}) \quad \text{subject to} \quad \boldsymbol{x} \in \mathcal{C}$$

is through a method called *projected gradient descent*. Above, $f(\boldsymbol{x})$ is a differentiable convex function and $\mathcal{C}$ is a (convex) feasibility set. The projected gradient descent iteration is

$$\boldsymbol{x}^{(k+1)} = \mathrm{P}_{\mathcal{C}}\left(\boldsymbol{x}^{(k)} - \alpha_k \nabla f(\boldsymbol{x}^{(k)})\right),$$

where $\mathrm{P}_{\mathcal{C}} : \mathbb{R}^D \to \mathcal{C}$ is the "closet point in $\mathcal{C}$" operator:

$$\mathrm{P}_{\mathcal{C}}(\boldsymbol{x}) = \arg\min_{\boldsymbol{v}\in\mathcal{C}}\ \|\boldsymbol{x} - \boldsymbol{v}\|_2.$$

For example, if $D = 4$, and

$$\mathcal{C} = \{\boldsymbol{x} \in \mathbb{R}^4\ :\ x[2] = 0,\ x[3] = 0\},$$

then

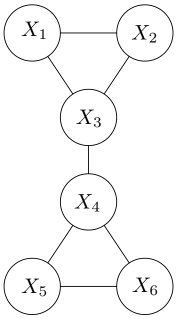
$$\mathrm{P}_{\mathcal{C}}(\boldsymbol{x}) = \begin{bmatrix} x[1] \\ 0 \\ 0 \\ x[4] \end{bmatrix}.$$

(That makes perfect sense, but you can formally justify this using the orthogonality principle since in this case $\mathcal{C}$ is a subspace.)

Suppose that we observe iid random vectors $X_1, X_2, \ldots, X_N$ in $\mathbb{R}^6$ distributed as

$$X_n \sim \mathrm{Normal}(\boldsymbol{0}, \boldsymbol{R}),$$

where $\boldsymbol{R}$ is unknown except for the inverse covariance structure indicated by this graph:



Write code that finds the MLE for the data vectors in the file `hw9p3_data.mat`. (That file contains a 6×1000 matrix `X` whose columns are the X_n referred to above.) Compare your answer to the sample covariance (i.e. the MLE in the unconstrained case).

4. Suppose that X is a random vector in $\mathbb{R}^D$ that depends on another random vector $\Theta \in \mathbb{R}^D$. With $\Theta = \boldsymbol{\theta}$ fixed, the conditional distribution for X is

$$f_X(\boldsymbol{x}|\Theta = \theta) = (2\pi)^{-D/2} \det(\boldsymbol{R})^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\theta})^{\mathrm{T}} \boldsymbol{R}^{-1} (\boldsymbol{x} - \boldsymbol{\theta})\right),$$

and the prior distribution for Θ is

$$f_\Theta(\boldsymbol{\theta}) = (2\pi)^{-D/2} \det(\boldsymbol{W})^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{W}^{-1} \boldsymbol{\theta}\right).$$

(a) Find the MAP estimate of $\boldsymbol{\theta}$ given a single observation $X_1 = \boldsymbol{x}_1$.

(b) Find the MAP estimate of $\boldsymbol{\theta}$ given N iid observations $X_1 = \boldsymbol{x}_1, \ldots, X_N = \boldsymbol{x}_N$.

5. Three friends, Aaron, Blake, and Colin, meet together every week to play poker. They each buy in for \$100, and play until one of them has it all. Poker is a game of skill, but also a game of luck — the winner each week is modeled as a discrete random variable X with distribution parameterized by θ_a, θ_b, with

$$\mathrm{P}\left(X = A\right) = \theta_a, \quad \mathrm{P}\left(X = B\right) = \theta_b, \quad \mathrm{P}\left(X = C\right) = 1 - \theta_a - \theta_b,$$

where

$$\theta_a, \theta_b \geq 0, \quad \text{and} \quad \theta_a + \theta_b \leq 1. \tag{1}$$

Above, event A corresponds to Aaron winning, B corresponds to Blake winning, and C corresponds to Colin winning.

The parameters θ_a and θ_b are unknown, and we want to infer them after observing the winners each week for many weeks. We have no idea of the relative skill of the players at the beginning of this experiment, so our prior is uniform on the triangular region specified by the constraints in (1):

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_a \\ \theta_b \end{bmatrix}, \quad f_\Theta(\boldsymbol{\theta}) = \begin{cases} 2, & \boldsymbol{\theta} \in \mathcal{S}, \\ 0, & \boldsymbol{\theta} \notin \mathcal{S}, \end{cases} \quad \mathcal{S} = \left\{ \boldsymbol{\vartheta} \in \mathbb{R}^2 \ : \ \vartheta[1], \vartheta[2] \geq 0, \ \ \vartheta[1] + \vartheta[2] \leq 1 \right\}.$$

(You might, at this point, want to sketch the set $\mathcal{S}$ in $\mathbb{R}^2$.)

(a) Show that after N weeks, where we have observed N_a wins for Aaron, N_b wins for Blake, and $N_c = N - N_a - N_b$ wins for Colin, the posterior for Θ is given by the *Dirichlet distribution*

$$f_\Theta(\boldsymbol{\theta} | X_1 = x_1, \ldots, X_N = x_n) \propto \theta_a^{N_a} \theta_b^{N_b} (1 - \theta_a - \theta_b)^{N - N_a - N_b}.$$

(The constant in front of the expression on the right turns out to be

$$\frac{\Gamma(N+3)}{\Gamma(N_a+1)\Gamma(N_b+1)\Gamma(N-N_a-N_b+1)},$$

which is the integral of the expression on the right over the constraint set $\mathcal{S}$.)

(b) Using MATLAB (or Python), plot the posterior density if after a year of play, we are at

$$N_a = 14, \quad N_b = 28, \quad N_c = 10.$$