Math Foundations of ML, Fall 2022 **Instructor: Justin Romberg**

Homework 3

Due: Friday, October 7, 2022 at 5:00 pm EST

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

1. The file `hw3p1_data.mat` contains two variables: `udata` and `ydata`. We will use this data to estimate a function $f : \mathbb{R}^2 \to \mathbb{R}$. The columns of `udata` contain sample locations, of which there are $M = 100$. The entries of $\boldsymbol{y}$ are the corresponding responses. We want to estimate f such that

$$f(\boldsymbol{u}_m) \approx y_m, \quad m = 1, \ldots, M, \quad \text{where} \quad \boldsymbol{u}_m = \begin{bmatrix} s_m \\ t_m \end{bmatrix}.$$

We will restrict f to be a second-order polynomial on $[0,1] \times [0,1]$:

$$f(s,t) = \alpha_1 s^2 + \alpha_2 t^2 + \alpha_3 st + \alpha_4 s + \alpha_5 t + \alpha_6, \tag{1}$$

which means that f lies in a six dimensional subspace of $L_2([0,1]^2)$.

   (a) Explain how to compute the 100×6 matrix $\boldsymbol{A}$ so that $\boldsymbol{y} \approx \boldsymbol{A}\boldsymbol{\alpha}$, where $\boldsymbol{y}$ contains the 100 response values in `ydata`. Write the code to compute $\boldsymbol{A}$ and turn it in.

   (b) Solve

$$\underset{\boldsymbol{\alpha} \in \mathbb{R}^6}{\text{minimize}} \;\; \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\alpha}\|_2^2.$$

   Turn in your code and the numerical value of your solution $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^6$.

   (c) Make a contour plot of the corresponding

$$\hat{f}(s,t) = \hat{\alpha}_1 s^2 + \hat{\alpha}_2 t^2 + \hat{\alpha}_3 st + \hat{\alpha}_4 s + \hat{\alpha}_5 t + \hat{\alpha}_6.$$

   Include 50 contour lines, just so we have a very clear picture of what this function looks like.

2. Consider the space $\mathcal{P}_2$ of second-order polynomials on $[0,1]^2$ specified by $\boldsymbol{\alpha} \in \mathbb{R}^6$ as in (1) above.

   (a) At every point (s,t), the gradient $\nabla f(s,t)$ of a function $f \in \mathcal{P}_2$ is a vector in $\mathbb{R}^2$. As every $f \in \mathcal{P}_2$ is specified by a vector $\boldsymbol{\alpha} \in \mathbb{R}^6$, we can think of the gradient at (s,t) as a mapping from $\mathbb{R}^6$ to $\mathbb{R}^2$. Show that this mapping is linear, which means, for a specified (s,t), there is a 2×6 matrix $\boldsymbol{G}_{s,t} \in \mathbb{R}^{2\times 6}$ such that

$$\nabla f(s,t) = \boldsymbol{G}_{s,t}\boldsymbol{\alpha}$$

(b) Find the 6×6 matrix $\boldsymbol{H}_{s,t} \in \mathbb{R}^{6\times 6}$ such that[1]

$$\|\nabla f(s,t)\|_2^2 = \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{H}_{s,t} \boldsymbol{\alpha}.$$

What kinds of functions f are in the null space of $\boldsymbol{H}_{s,t}$ for all s and t? Why?

(c) Compute the matrix

$$\boldsymbol{Q} = \int_0^1 \int_0^1 \boldsymbol{H}_{s,t}\; ds\; dt.$$

(This is done simply by integrating each entry individually.)

(d) Describe how to set up and solve the optimization program

$$\underset{\boldsymbol{f} \in \mathcal{P}_2}{\text{minimize}} \sum_{m=1}^{M} (y_m - f(s_m, t_m))^2 + \delta \int_0^1 \int_0^1 \|\nabla f(s,t)\|_2^2 \, ds dt.$$

What is the regularizer above penalizing? What kinds of solutions do we expect for large δ?

(e) Apply your answer to part (d) to the data set from Problem 1. Play around with the value of δ, and produce estimates for three different δ that are interesting. Discuss why you think those values are indeed "interesting".

3. Let $\boldsymbol{A}$ be an $M \times N$ matrix with $\text{rank}(\boldsymbol{A}) < N$. We have seen in this case that the least-squares problem

$$\underset{\boldsymbol{x} \in \mathbb{R}^N}{\text{minimize}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 \tag{2}$$

has an infinite number of solutions. We have also seen, however, that the regularized least squares problem

$$\underset{\boldsymbol{x} \in \mathbb{R}^N}{\text{minimize}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \delta \|\boldsymbol{x}\|_2^2 \tag{3}$$

has a unique solution for every $\delta > 0$. In this problem, we will show that as $\delta \to 0$, the regularized solution goes to the minimum norm solution of

$$\underset{\boldsymbol{x} \in \mathbb{R}^N}{\text{minimize}} \|\boldsymbol{x}\|_2^2 \quad \text{subject to} \quad \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}. \tag{4}$$

(a) Start by showing that if $\boldsymbol{x}_1 \in \text{Row}(\boldsymbol{A})$ and $\boldsymbol{x}_2 \in \text{Row}(\boldsymbol{A})$ then $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}_1 \neq \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}_2$ unless $\boldsymbol{x}_1 = \boldsymbol{x}_2$.

(b) Use part (a) to argue that the solution to (4) is always unique.

(c) In fact, something stronger than what we showed in part (a) is true. There exists a constant $C > 0$ such that

$$\|\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{x}_1 - \boldsymbol{x}_2)\|_2 \;\geq\; C\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 \quad \text{for all} \quad \boldsymbol{x}_1, \boldsymbol{x}_2 \in \text{Row}(\boldsymbol{A}).$$

(This follows very easily from work we do later in the course, so we will defer its proof for now.) Use this fact to show that the solution of (3) goes to the solution of (4) as $\delta \to 0$. In particular, if $\boldsymbol{x}^\star$ is the (always unique) minimizer of (4), and $\hat{\boldsymbol{x}}_n$ is the (always unique) minimizer of (3) with[2] $\delta = 1/n$, show that

$$\lim_{n\to\infty} \hat{\boldsymbol{x}}_n = \boldsymbol{x}^\star,$$

i.e. $\lim_{n\to\infty} \|\boldsymbol{x}^\star - \hat{\boldsymbol{x}}_n\|_2 = 0$.

[1] Hint: $\|\boldsymbol{G}_{s,t}\boldsymbol{\alpha}\|_2^2 = \langle \boldsymbol{G}_{s,t}\boldsymbol{\alpha}, \boldsymbol{G}_{s,t}\boldsymbol{\alpha}\rangle = \cdots$.

[2] There is nothing special about taking $\delta = 1/n$ ... your argument should work for any sequence of δs that goes to zero.