Department of Mathematics, University of Utah

Introduction to Optimization

MATH 5770/6640, ME EN 6025 – Section 001 – Fall 2021

Homework 3

Least squares and gradient descent

Due October 5, 2021

Submit your homework assignment as a scanned copy **ON CANVAS**, to the `Homework 3` assignment.

Some of the exercises below are computational. The book problems are explained in Matlab. You need not use Matlab to complete the assignment; numerical simulation with any programming language is acceptable.
Text: *Introduction to Nonlinear Optimization*, Amir Beck,

Exercises: # 3.1,
3.2,
4.3 (only the first 3 parts, ignore the *diagonally scaled* portions)

Additional problems:

P1. (Maximum likelihood estimation) Let $\{y_1, \ldots, y_M\} \subset \mathbb{R}$ denote M data points on the real line. The overall goal of this problem is to "fit" a probability distribution to these data points.
In particular, we assume that this data arose as (independent, identically distributed) samples from an unknown probability distribution with density $p(y)$. In order to find $p(y)$, we assume further that p corresponds to a normal distribution, i.e., a distribution having density

$$p(y; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-(y-\mu)^2/(2\sigma^2)),$$

where μ and σ are the unknown mean and standard deviation of the distribution. We will choose the *parameters* (μ, σ) of this distribution as those parameters that maximize the "likelihood" of the data. In particular, given (μ, σ) and the data $\{y_m\}_{m=1}^M$, the likelihood is formally defined as

$$\mathcal{L}(\mu, \sigma) \coloneqq \prod_{m=1}^M p(y_m; \mu, \sigma) = \prod_{m=1}^M \frac{1}{\sigma\sqrt{2\pi}} \exp(-(y_m-\mu)^2/(2\sigma^2)),$$

which is the probability of seeing independent data $\{y_m\}_{m=1}^M$ conditioned on their distribution having parameters (μ, σ). (It is not necessary for you to understand probability to complete this problem.)
The *maximum likelihood estimate* is the parameter choice that maximizes the likelihood:

$$(\mu_*, \sigma_*) = \underset{\mu\in\mathbb{R}, \sigma\in\mathbb{R}_++}{\text{argmax}} \; \mathcal{L}(\mu, \sigma).$$

Show that a strict global maximum of this optimization problem is given by

$$\mu_* = \frac{1}{M}\sum_{m=1}^{M} y_m, \qquad \sigma_*^2 = \frac{1}{M}\sum_{m=1}^{M}(y_m - \mu_*)^2.$$

(You may find it convenient to (i) use the logarithm function to monotonically transform the likelihood, (ii) convert the maximization problem into a minimization problem.)
6000-level students only: Simulate this result – with $M = 100$, choose some fixed value of μ, σ and generate data $\{y_m\}_{m=1}^{100}$ from a normal distribution with your prescribed (μ, σ). Compare a histogram of the data against the density $p(\cdot; \mu_*, \sigma_*)$ computed as the maximum likelihood estimate above.

P2. (Maximum likelihood for coin flips) Suppose that you are given the result of 100 flips of a two-sided coin. Let H denote the number of heads observed, and T the number of tails (so that $H + T = 100$). Assume that $H, T > 0$. The coin may not be fair; it has probability $p \in [0, 1]$ that a heads is observed (and $1 - p$ for tails). The goal is determine the parameter p that maximizes the likelihood of having observed (H, T). Given that the likelihood is equal to

$$\mathcal{L}(p) = \binom{100}{H} p^H (1-p)^T,$$

compute a maximum likelihood estimate for p. Is your computed value a global maximum?

P3. (**6000-level students only**) Consider the optimization problem,

$$\min_{\boldsymbol{x} \in S \subset \mathbb{R}^n} f(\boldsymbol{x}),$$

where S is a given subset of $\mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$ is a given function. Prove that if $g : \mathbb{R} \to \mathbb{R}$ is a strictly monotonic increasing function, then

$$\operatorname*{argmax}_{\boldsymbol{x} \in S \subset \mathbb{R}^n} f(\boldsymbol{x}) = \operatorname*{argmax}_{\boldsymbol{x} \in S \subset \mathbb{R}^n} g(f(\boldsymbol{x})), \qquad \operatorname*{argmin}_{\boldsymbol{x} \in S \subset \mathbb{R}^n} f(\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{x} \in S \subset \mathbb{R}^n} g(f(\boldsymbol{x})).$$