MATH FOUNDATION OF DATA SCIENCE HOMEWORK 3

$>>$**NAME:** $<<$
$>>$HKU ID: $<<$

Instructions: Please complete the homework in LaTeX, produce a pdf, and submit it on time as indicated on Moodle.

1 Optimization basics

Q1 [10 points] Let $A \in R^{n\times n}$ be a symmetric matrix. Use an algebraic method to solve the following problem

$$\begin{array}{ll} \text{Minimize} & \mathbf{x}^T A\mathbf{y} \\ \text{subject to} & \mathbf{x}^T\mathbf{x} = 1 \\ & \mathbf{y}^T\mathbf{y} = 1. \end{array}$$

Solution:

Q2 [10 points] Consider the problem

$$\begin{array}{ll} \text{Minimize} & f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} + \mathbf{b}^T\mathbf{x} \\ \text{subject to} & -1 \leq x_i \leq 1 \ \text{ for } i = 1, 2, 3, \end{array}$$

where

$$A = \begin{pmatrix} 13 & 12 & -2 \\ 12 & 17 & 6 \\ -2 & 6 & 12 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} -22.0 \\ -14.5 \\ 13.0 \end{pmatrix}.$$

(a) [2 points] Find the gradient ∇f and the Hessian matrix $\nabla^2 f$ of f.

(b) [8 points] Prove that $\mathbf{x}^* = (1, \frac{1}{2}, -1)^\top$ is an optimal solution.

Solution:

Q3 [10 points] Consider the problem of minimizing (locally) the function

$$f(x, y) = p(x^2 + y^2 - 2x - 2y) + (xy - 1)^2,$$

where x and y are real numbers and p is a real parameter. Answer the following questions, justifying your answers rigorously.

(a) [2 points] What are values x_0 and y_0 such that f has a stationary point at (x_0, y_0) for *every* value of p?

(b) [8 points] For which value(s) of p can you be certain that f is convex in a neighborhood of (x_0, y_0)?

Solution:

2 Constrained Optimization

Q4 [10 points] Let A be an $m \times n$ matrix of rank m and let b be a vector in $\mathbb{R}^n$ such that

$$b^T[I - A^T(AA^T)^{-1}A]b > 0.$$

Consider the following problem:

$$\begin{aligned} \min \quad & b^T x \\ \text{s.t.} \quad & Ax = 0 \\ & x^T x \leq 1. \end{aligned}$$

(a) [4 points] Demonstrate that LICQ holds at every feasible point.

(b) [6 points] Determine a x^* that satisfies the KKT conditions.

Solution:

Q5 [10 points] Consider the problem discussed in Question 4. Suppose that $b \neq 0$ and $b^T[I - A^T(AA^T)^{-1}A]b > 0$.

(a) [6 points] Formulate its dual problem.

(b) [4 points] Solve the dual problem obtained in (a), and then find an optimal solution to the primal.

Solution:

3 Gradient Descent and Variants

Q6 [10 points] In weighted least squares, each data point has a weight $w_i > 0$, and the objective is:

$$\min_{\theta \in \mathbb{R}^d} L(\theta) = \frac{1}{2} \sum_{i=1}^{n} w_i (y_i - \theta^\top x_i)^2$$

(a) (2 points) Write this objective in matrix form using a diagonal weight matrix $W = \text{diag}(w_1, \ldots, w_n)$.

(b) (2 points) Derive the closed-form solution θ^*.

(c) (2 points) Compute the gradient $\nabla L(\theta)$ and Hessian $\nabla^2 L(\theta)$.

(d) (2 points) If we use gradient descent with step size η, what is the condition on η for convergence? Express your answer in terms of the weights w_i and the data matrix X.

(e) (2 points) Explain when weighted least squares is preferred over ordinary least squares. Give a practical example.

Solution:

Q7 [10 points] Consider the quadratic function $f(\theta) = \frac{1}{2}\theta^T A\theta$ where

$$A = \begin{bmatrix} 26 & 24 \\ 24 & 26 \end{bmatrix}$$

(a) (3 points) What is the condition number κ of this problem?

(b) (4 points) Using gradient descent with optimal step size $\eta = \frac{2}{L+\mu}$, how many iterations are needed to reduce the error $\|\theta^{(t)} - \theta^*\|$ by a factor of 10^{-6}? (Recall: GD converges as $\left(\frac{\kappa-1}{\kappa+1}\right)^t$)

(c) (3 points) Can we design a better optimization algorithm to make the above convergence faster? Please give some detailed explanations.

Solution:

Q8 [10 points) Consider SGD with mini-batch size B on a finite-sum problem with n samples: $L(\theta) = \frac{1}{n}\sum_{i=1}^{n} L_i(\theta)$.
Recall the convergence bound:

$$\mathbb{E}[L(\theta^{(t+1)})] \leq L(\theta^{(t)}) - \left(\eta - \frac{L\eta^2}{2}\right) \|\nabla L(\theta^{(t)})\|^2 + \frac{L\eta^2\sigma^2}{2}$$

where σ^2 denotes an upper bound of the variance of stochastic gradients: $\mathbb{E}_i[\|\nabla L_i(\theta) - \nabla L(\theta)\|_2^2] \leq \sigma^2$.

(a) (5 points) What happens if we use a constant step size η and run SGD for a very long time (as $t \to \infty$)? Does it converge to θ^*?

(b) (5 points) Why is variance σ^2 smaller for larger mini-batch sizes B? (Hint: Think about averaging independent random variables)

Solution:

Q9 [10 points] In SVRG, the gradient estimator is:

$$g_t = \nabla L_i(\theta^{(t)}) - \nabla L_i(\tilde{\theta}) + \nabla L(\tilde{\theta})$$

where $\tilde{\theta}$ is a historical model parameter.

(a) (3 points) Show that $\mathbb{E}[g_t] = \nabla L(\theta^{(t)})$ (i.e., g_t is unbiased).

(b) (3 points) Explain intuitively why $\text{Var}(g_t)$ is small when $\theta^{(t)} \approx \tilde{\theta}$.

(c) (4 points) What is the main computational cost of SVRG per epoch, i.e., all computations in one outer loop? (Count both outer and inner loop costs, assuming the inner loop takes n steps, i.e., the same as the number of data points).

Solution:

Q10 [10 points] Consider the momentum update:

$$\begin{aligned} m_t &= \beta m_{t-1} + \eta \nabla L(\theta^{(t)}) \\ \theta^{(t+1)} &= \theta^{(t)} - m_t \end{aligned}$$

(a) (5 points) Expand m_t to show that it's an exponentially weighted sum of past gradients.

(b) (5 points) If gradients are constant at g for all iterations, what is the effective step size in the direction of g? (Hint: Compute $\sum_{k=0}^{\infty} \beta^k$)

Solution: