CS 189 Introduction to Machine Learning
Fall 2019 Jennifer Listgarten & Stella Yu DIS1

1 More Gradients

Consider the optimization problem $\min_{x\in\mathbb{R}^n} \frac{1}{2}x^\top Ax - b^\top x$, where $A \in \mathbb{R}^{n\times n}$ is a PSD matrix with $0 < \lambda_{\min}(A) \leq \lambda_{\max}(A) < 1$.

(a) Find the optimizer x^*.

(b) Solving a linear system directly using Gaussian elimination takes $O(n^3)$ time, which may be wasteful if the matrix A is sparse. For this reason, we will use gradient descent to compute an approximation to the optimal point x^*. Write down the update rule for gradient descent with a step size of 1.

(c) Show that the iterates $x^{(k)}$ satisfy the recursion $x^{(k)} - x^* = (I - A)(x^{(k-1)} - x^*)$.

(d) Using the previous part and the fact that $||Ax||_2 \leq \lambda_{\max(A)}||x||_2$, show that for some $0 < \rho < 1$,

$$||x^{(k)} - x^*||_2 \leq \rho||x^{(k-1)} - x^*||_2.$$

(e) Let $x^0 \in \mathbb{R}^n$ be the starting value for our gradient descent iterations. If we want a solution $x^{(k)}$ that is $\epsilon > 0$ close to x^*, i.e. $||x^{(k)} - x^*||_2 \leq \epsilon$, then how many iterations of gradient descent should we perform? In other words, how large should k be? Give your answer in terms of $\rho, ||x^{(0)} - x^*||_2$, and ϵ.

2 Least Squares Using Calculus

(a) In ordinary least-squares linear regression, we typically have $n > d$ so that there is no $\mathbf{w}$ such that $\mathbf{Xw} = \mathbf{y}$ (these are typically overdetermined systems — too many equations given the number of unknowns). Hence, we need to find an approximate solution to this problem. The residual vector will be $\mathbf{r} = \mathbf{Xw} - \mathbf{y}$ and we want to make it as small as possible. The most common case is to measure the residual error with the standard Euclidean ℓ^2-norm. So the problem becomes:

$$\min_{\mathbf{w}} ||\mathbf{Xw} - \mathbf{y}||_2^2$$

Where $\mathbf{X} \in \mathbb{R}^{n\times d}, \mathbf{w} \in \mathbb{R}^d, \mathbf{y} \in \mathbb{R}^n$. Derive using vector calculus an expression for an optimal estimate for $\mathbf{w}$ for this problem assuming $\mathbf{X}$ is full rank.

(b) How do we know that $\mathbf{X}^\top\mathbf{X}$ is invertible?

(c) What should we do if $\mathbf{X}$ is not full rank?

3 Regularization and Risk Minimization

(a) Let $\mathbf{A}$ be a $d \times n$ matrix. For any $\mu > 0$, show that $(\mathbf{A}\mathbf{A}^\top + \mu\mathbf{I})^{-1}\mathbf{A} = \mathbf{A}(\mathbf{A}^\top\mathbf{A} + \mu\mathbf{I})^{-1}$.

(b) Let $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$ be a sequence of data points. Each y_i is a scalar and each $\mathbf{x}_i$ is a vector in $\mathbb{R}^d$. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top$ and $\mathbf{y} = [y_1, \ldots, y_n]^\top$. Consider the *regularized* least squares problem.

$$\min_{\mathbf{w}\in\mathbb{R}^d} ||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2 + \mu||\mathbf{w}||_2^2$$

Show that the optimum $\mathbf{w}_*$ is unique and can be written as the linear combination $\mathbf{w}_* = \sum_{i=1}^n \alpha_i \mathbf{x}_i$ for some scalars $\alpha_1, ..., \alpha_n$. What are the coefficients α_i?

(c) More generally, consider the general regularized empirical risk minimization problem

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \text{loss}(\mathbf{w}^\top\mathbf{x}_i, y_i) + \mu||\mathbf{w}||_2^2$$

where the loss function is convex in the first argument. Prove that the optimal solution has the form $\mathbf{w}_* = \sum_{i=1}^n \alpha_i \mathbf{x}_i$. If the loss function is not convex, does the optimal solution have the form $\mathbf{w}_* = \sum_{i=1}^n \alpha_i \mathbf{x}_i$? Justify your answer.

4 Covariance Practice

Let X be a multivariate random variable (recall, this means it is a vector of random variables) with mean vector $\mu \in \mathbb{R}^d$ and covariance matrix $C \in \mathbb{R}^{d\times d}$. Prove that if C is singular, then the space where X takes values with non-zero probability (this space is called the support of X) has dimension strictly less than n.

Hint: use the identity $\text{Var}(\sum\limits_{i=1}^{d} Y_i) = \sum\limits_{i=1}^{d}\sum\limits_{j=1}^{d} \text{Cov}(Y_i, Y_j)$.