CS 189 Introduction to Machine Learning
Fall 2017 DIS5

1 Derivation of PCA

In this question we will derive PCA. PCA aims to find the direction of maximum variance among a dataset. You want the line such that projecting your data onto this line will retain the maximum amount of information. Thus, the optimization problem is

$$\max_{u:\|u\|_2=1} \frac{1}{n} \sum_{i=1}^{n} \left(u^T x_i - u^T \hat{x}\right)^2$$

where n is the number of data points and $\hat{x}$ is the sample average of the data points.

(a) Show that this optimization problem can be written in this format:

$$\max_{u:\|u\|_2=1} u^T \Sigma u$$

where $\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x})(x_i - \hat{x})^T$.

(b) Show that the maximizer for this problem is equal to v_1, where v_1 is the eigenvector corresponding to the largest eigenvalue λ_1. Also show that optimal value of this problem is equal to λ_1.

2 Canonical Correlation Analysis

Assume that you have a database of images of the words typed in two different fonts. $X, Y \in \mathbb{R}^{n \times d}$ corresponds to the dataset of font 1 and font 2 respectively. Think of the database X as being composed on n independent draws (samples) from a random variable $\mathbf{X} \in \mathbb{R}^d$, and similarly Y as n draws from a random variable $\mathbf{Y}$. Your goal is to use machine learning to build a text recognition of word images.

(a) Explain why you would want to consider using CCA in this problem.

(b) Assume that X and Y include zero-mean features of the word images. Given two vectors $u, v \in \mathbb{R}^d$, what is the correlation coefficient of the projected variables? Correlation coefficient between two scalar random variables P and Q is computed by:

$$\rho(P,Q) = \frac{cov(P,Q)}{\sigma_P \sigma_Q}$$

(c) Assume that the features of matrix X are rescaled to have values between -1 and 1. How does this change the correlation coefficient?

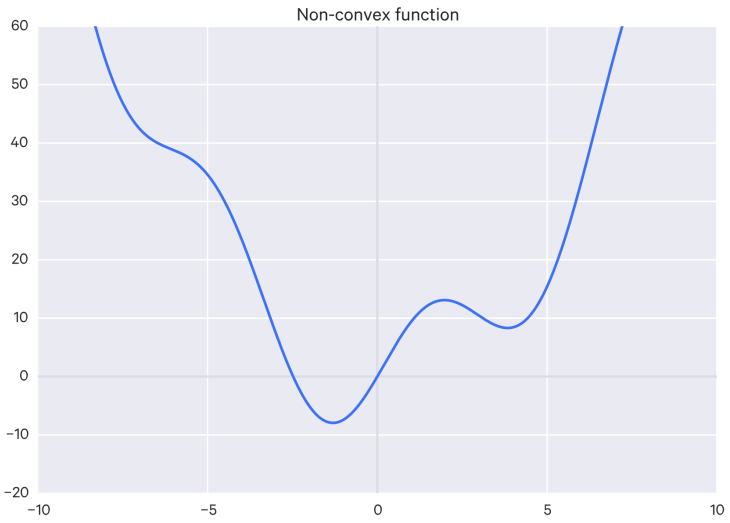

Figure 1: A nonconvex function

3 Convex Functions for Gradient Descent

Convex functions have properties that make them very useful in machine learning and optimization. In this exercise, we will take a look at these properties.

Recall the definition of convexity: a function f is convex if and only if for any x, y and scalar $\alpha \in [0,1]$, the following inequality holds:

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$$

(a) Show that if two functions f and g are convex, then their sum $f+g$ is convex.

(b) Show that l_2 regularized least squares is an optimization problem over a convex function.

(c) Consider the nonconvex function in the figure. How does the choice of initial value affect the result of the gradient descent? What do you suggest to improve that?

4 Extra: PCA Review

In this problem, we fix an $n \times n$ positive semi-definite matrix A. We will derive, from first principles, the best rank-1 approximation to A. Recall the following metric of approximation: for any integer $1 \leq r \leq n$, we define the best rank-r approximation as any minimizer A_r of

$$\min_{M \in R^{n \times n}} \|A - M\|_F : \mathsf{rank}(M) \leq r \tag{1}$$

Note that such a minimizer is not necessarily unique (why?). Since A is positive semi-definite, A_r must be as well: you make take this fact for granted. In this problem, we focus on the special case of $r = 1$. Also assume $A \neq 0$ to avoid any uninteresting, degenerate cases.

(a) Show that $A_1 = mm^T$, where m is any minimizer of the following *unconstrained* optimization problem

$$\min_{m \in R^n} \left\| A - mm^T \right\|_F . \tag{2}$$

(b) Define the function $f : R^n \longrightarrow R$ as $f(m) = \left\| A - mm^T \right\|_F^2$. Compute $\nabla f(m)$.

(c) Argue from part (c) that $A_1 = \lambda_1 v_1 v_1^T$, where λ_1 is the maximum eigenvalue of A and v_1 is a corresponding (unit normalized) eigenvector.