Optimization-Based Data Analysis

Recitation 11

1. You evaluated $\log(e^a + e^b + e^c + e^d)$ in Python for $a, b, c, d \in \mathbb{R}$ and got ∞. What happened, and is there a better computation you could perform?

   Solution. Let $m = \max(a, b, c, d)$. Then

   $$\log(e^a+e^b+e^c+e^d) = \log(e^m(e^{a-m}+e^{b-m}+e^{c-m}+e^{d-m})) = m+\log(e^{a-m}+e^{b-m}+e^{c-m}+e^{d-m}).$$

2. Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable but not convex. We want to minimize f by always stepping in a descent direction in order to find the global minimum.

   (a) Suppose $\nabla f(\vec{x}) \neq 0$ for some fixed $\vec{x}$. Which directions are descent directions (i.e., directions such that the function decreases if the step size is sufficiently small)?

   (b) Suppose $\nabla f(\vec{x}) = 0$ for some fixed $\vec{x}$ that is not the global minimum. Give conditions on $\nabla^2 f(\vec{x})$ so that there exists a descent direction at $\vec{x}$. Under those conditions, determine a descent direction.

   Solution.

   (a) Any direction $\vec{p}$ such that $\nabla f(\vec{x})^T \vec{p} < 0$.

   (b) Has at least 1 negative eigenvalue. Corresponding eigenvector.

3. We have a signal $\vec{x} \in \mathbb{C}^n$ and let $\vec{X} = W\vec{x} \in \mathbb{C}^n$ denote its DFT where W is the DFT matrix.

   (a) Suppose we observe noisy frequency samples $\vec{Y} = \vec{X} + \vec{z}$. How could we estimate the true signal $\vec{x}$?

   (b) Suppose we observe noisy frequency samples $\vec{Y} = \vec{X} + \vec{z}$ but we also know the true signal $\vec{x}$ is sparse. How could we estimate $\vec{x}$?

   (c) Suppose we observe some of the exact frequency samples

   $$\vec{X}[k_1], \ldots, \vec{X}[k_p]$$

   for some $p < n$ and we know that $\vec{x}$ is sparse. How could we estimate $\vec{x}$?

   Solution.

   (a)
   $$\text{minimize}_{\vec{x}} \quad \|W\vec{x} - \vec{Y}\|_2^2$$

   (b)
   $$\text{minimize}_{\vec{x}} \quad \|W\vec{x} - \vec{Y}\|_2^2 + \lambda\|\vec{x}\|_1$$

(c)

$$\begin{array}{ll} \text{minimize}_{\vec{x}} & \|\vec{x}\|_1 \\ \text{subject to} & (W\vec{x})[k_i] = \vec{X}[k_i] \quad \text{for } i = 1, \ldots, p \end{array}$$

4. We have a signal $\vec{x} \in \mathbb{C}^{2n}$ but we observe it through samples of its DFT $\vec{X} \in \mathbb{C}^{2n}$. Suppose each sample has a cost, so we instead only sample the even coordinates:

$$\vec{X}[0], \vec{X}[2], \ldots, \vec{X}[2n-2].$$

   (a) Suppose $\vec{x}$ is s-sparse where $s < n$. Can we exactly recover $\vec{x}$ from the given measurements?

   (b) Give a general assumption on $\vec{x}$ that would let us recover it exactly from the given measurements.

   Solution.

   (a) Not necessarily. For example, $e_k + e_{k+n}$ is indistinguishable from $2e_k$. To see this, we can either appeal to aliasing (by the similarity of the DFT and the inverse DFT) or note that

$$e^{-2\pi i(2k)j/(2n)} = e^{-2\pi i(2k)(j+n)/(2n)}.$$

   This shows that $W[:: 2, :] = [W[:: 2, : n]W[:: 2, n :]]$ using python indexing notation. That is, the first n columns of the even rows of W match the last n columns. Thus $e_k - e_{k+n}$ is in the null space of $W[:: 2, :]$ for all k.

   (b) If we knew that it was supported in the first n entries.

5. We observe $\vec{y} \in \mathbb{R}^n$ given by $\vec{y} = X\vec{\beta} + \vec{z}$ where $X \in \mathbb{R}^{n \times p}$, $\vec{\beta} \in \mathbb{R}^p$, and $\vec{z} \in \mathbb{R}^n$. Here all of the columns $X[:, i]$ have zero mean, and $\vec{z}$ is uncorrelated with each of the features.

   (a) Suppose $\vec{\beta}[i] > 0$ what does this imply about the sample correlation between $\vec{y}$ and the ith feature $X[:, i]$?

   (b) Before performing the regression, should we simplify our model by removing all features i that are uncorrelated with $\vec{y}$?

   Solution.

   (a) Nothing. Note that

$$X[:, i]^T \vec{y} = X[:, i]^T X\vec{\beta} = \sum_j X[:, i]^T X[:, j]\vec{\beta}[j],$$

   so it entirely depends on how the features are correlated with one another.

   (b) No. As we saw above, uncorrelated features need not correspond to 0 coefficients (in $\vec{\beta}$).

6. Suppose each row of $\mathbf{X} \in \mathbb{R}^{n \times p}$ is an iid draw from a p-dimensional distribution with mean 0 and covariance Σ. Let $\vec{\mathbf{z}} \in \mathbb{R}^n$ be an iid draws that are independent of $\mathbf{X}$. Let $\vec{\mathbf{y}} \in \mathbb{R}^n$ be defined by $X\vec{\beta} + \vec{\mathbf{z}}$ where $\vec{\beta} \in \mathbb{R}^p$.

   (a) Suppose $p = 4$ and $\vec{\beta} = (1, 3, 0, 0)^T$ and the covariance matrix Σ is given by

$$\Sigma = \begin{bmatrix} 1 & 0 & .99 & 0 \\ 0 & 1 & 0 & .99 \\ .99 & 0 & 1 & 0 \\ 0 & .99 & 0 & 1 \end{bmatrix}.$$

   What effect will this covariance have on the least squares estimated coefficients $\vec{\beta}_{\text{LS}}$?

   (b) Does this effect persist if there is a lot of data?

   Solution.

   (a) The first and third coefficients will be "confused" by the model, along with the second and fourth coefficients. More precisely, the least squares estimate is given by

$$\beta + (X^T X)^{-1} X^T \vec{z} = \beta + V S^{-1} U^T \vec{z},$$

   where USV^T is the SVD of X. Due to the correlation, the difference between the first and third columns, and the second and fourth columns will nearly be in the null space of X. Thus they will approximately correspond to the smallest 2 singular values of X and the largest two of the pseudoinverse (will roughly correspond to the last two columns of V). Thus the noise will produce a magnified effect in those two directions (roughly $e_1 - e_3$ and $e_2 - e_4$).

   (b) As $n \to \infty$ we have

$$(X^T X)^{-1} X^T \vec{z} \to \Sigma^{-1} \frac{X^T \vec{z}}{n} \to 0,$$

   since by the SLLN since X and $\vec{z}$ are uncorrelated.

7. It can be harder to fit a regression model $\vec{y} = X\vec{\beta} + \vec{z}$ when the columns of X are highly correlated. To avoid this, someone suggests to first orthogonalize the data matrix X before performing the regression.

   (a) Suppose X already contains orthogonal columns. What is the least squares estimator $\vec{\beta}_{\text{LS}}$?

   (b) Is first orthogonalizing a good idea?

   Solution.

   (a) $\vec{\beta}[i] = X[:, i]^T \vec{y} / \|X[:, u]\|_2^2$ seen by computing $X[:, i]^T \vec{y}$.

(b) It is exactly what we do when we compute the least squares solution using the QR factorization. It doesn't really solve the correlation issue, since the regression coefficients are in terms of the new orthgonalized features, and usually need to be converted back to the old features to be interpretted (this is what the R matrix encodes).

8. Suppose we are solving the minimization problem

$$\arg\min_{\vec{\beta}} \frac{1}{2}\|X\vec{\beta} - \vec{y}\|_2^2 + \lambda\|\vec{\beta}\|_1.$$

We are going to apply coordinate descent, where we minimize each coordinate separately, and repeat. What is the solution to

$$\arg\min_{\vec{\beta}[i]} \frac{1}{2}\|X\vec{\beta} - \vec{y}\|_2^2 + \lambda\|\vec{\beta}\|_1?$$

Solution. Computing the subdifferential we obtain

$$X^T(X\vec{\beta} - \vec{y}) + \lambda\partial\|\vec{\beta}\|_1.$$

We need to choose $\vec{\beta}[i]$ so that there is a subgradient whose ith coordinate is zero. That is, we need

$$X[:,i]^T(X\vec{\beta} - \vec{y}) + \lambda\,\mathrm{sgn}(\vec{\beta}[i]) = 0.$$

Letting $\hat{\vec{\beta}} = \vec{\beta} - \vec{\beta}[i]e_i$ gives

$$\|X[:,i]\|_2^2\vec{\beta}[i] = X[:,i]^T\vec{y} - X[:,i]^TX\hat{\vec{\beta}} - \lambda\,\mathrm{sgn}(\vec{\beta}[i]).$$

Let $w = X[:,i]^T\vec{y} - X[:,i]^TX\hat{\vec{\beta}}$. Checking the cases of $w \geq \lambda$, $w \leq -\lambda$ and $w \in (-\lambda, \lambda)$ we see the answer is

$$\vec{\beta}[i] = \frac{1}{\|X[:,i]\|_2^2}S_\lambda(w).$$