CS 189 Introduction to Machine Learning
Summer 2018

DIS12

1 Orthogonal Matching Pursuit

Consider the problem setting, where you are given $\mathbf{X}$ and $\mathbf{y}$ where $\mathbf{X} \in \mathbb{R}^{n\times d}$ and $\mathbf{y} \in \mathbb{R}^n$. That is we have n observations that are given by linear combinations of d-features. How can we find the original $\mathbf{w}$ such that $\mathbf{y} = \mathbf{X}\mathbf{w}^*$? We've learned many techniques for this so far, but suppose you have an under-determined system $n \ll d$ and we want a k-sparse solution (a signal $\mathbf{w}$ is called k-sparse if $|\mathbf{w}|_0 = k$). In the lecture, we discussed Lasso (ℓ_1-regularized regression) as one of the powerful methods to solve such problems. And furthermore, we saw that coordinate descent can be employed to solve the lasso-penalized regression problem. With coordinate descent, we make progress along a randomly chosen coordinate or feature. A natural question arises: can we choose the coordinates in a smart way? And what if we optimize (rather than taking simply a step) along that coordinate? Let's review one technique, that attempts a greedy coordinate selection procedure: orthogonal matching pursuit. Let us introduce some notation: we use $\mathbf{x}_j$ to denote the j-th column of the matrix $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_d \\ | & | & & |. \end{bmatrix}$$

Note that the vector $\mathbf{x}_j \in \mathbb{R}^n$ denotes the j-th feature for all n data points.

OMP Algorithm:

1. Initialize the residue $\mathbf{r}^0 = \mathbf{y}$ and initialize the index set to be $I^0 = \emptyset$. Set your estimate $\hat{\mathbf{w}}^0 = 0$.

2. Repeat for $t = 1, \dots$ **until** $\|r^t\|_2 <$ threshold or $t > k$(sparsity budget):

   - **Index update**: Find an index j^t which reduces the residue most. In other words, find the best one-feature linear fit to the existing residual vector, and then update the index set by including the index corresponding to the best feature. Mathematically this update can be written as:

$$\begin{aligned} &\mathbf{r}^t = \mathbf{y} - \mathbf{X}\hat{\mathbf{w}}^{t-1} \\ &j^t = \arg\min_i(\min_v \|\mathbf{r}^t - v\mathbf{x}_i\|_2^2) \\ &I^t = I^{t-1} \cup \{j^t\} \end{aligned}$$

   - **Estimate update**: Estimate the best linear fit of the target $\mathbf{y}$ using the features obtained so far. Given that we have found t good features, we now find the best linear fit for the target $\mathbf{y}$ using these t-features. Define $\mathbf{X}_t = \left[\mathbf{x}_{j^1}, \ldots, \mathbf{x}_{j^t}\right]$ made up of these t-features. Then we determine $\hat{\mathbf{w}}^t$ as the solution for the following least-squares problem:

$$\hat{\mathbf{w}}^t = \arg\min_{\mathbf{w}\in\mathbb{R}^t} \|\mathbf{y} - \mathbf{X}_t\mathbf{w}\|_2^2$$

We now discuss under what conditions can we expect such a greedy-coordinate-finding procedure to provide us a good solution. To keep the discussion centered around key ideas, we discuss the simplest case possible: recovery of a one-sparse signal.

(a) 1**-sparse noiseless case**: Suppose that the true signal is given by

$$\mathbf{w}^* = \begin{bmatrix} w_1^* \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad w_1^* \neq 0$$

(but we don't know the true signal) and we are given *noiseless* observations

$$\mathbf{y} = \mathbf{X}\mathbf{w}^* = \sum_{i=1}^{d} \mathbf{x}_i w_i^* = \mathbf{x}_1 w_1^*.$$

When will OMP work for such a setting? First consider the case when columns of $\mathbf{X}$ are normalized to have unit norm, that is $\|\mathbf{x}_i\|_2 = 1$. **Is it necessary to have normalized columns for exact recovery in this setting?** Note that since we have one-sparse signal, OMP works correctly if it finds the right coordinate in the first $(t = 1)$ step.

Hint: Let's define $\mu = \max_{i\neq j} \frac{|\mathbf{x}_i^\top \mathbf{x}_j|}{\|\mathbf{x}_i\|\mathbf{x}_j\|}$. What condition on μ do we need for exact recovery?

(b) 1**-sparse noisy case**: For simplicity, let us assume that we have normalized columns and that the observations are corrupted by Gaussian noise. We continue to assume that the true signal is 1-sparse:

$$\mathbf{y} = \mathbf{X}\mathbf{w}^* + \mathbf{Z} = \sum_{i=1}^{d} \mathbf{x}_i w_i^* + \mathbf{Z} = \mathbf{x}_1 w_1^* + \mathbf{Z}$$

where $w_1^* \neq 0$ and $\mathbf{Z} \sim \mathcal{N}(0, \sigma^2 I)$ is an n-dimensional Gaussian random vector. **When will OMP recover the true support of the signal?** Note that true signal magnitude can not be recovered exactly due to noise, but here we investigate if we can find the right index using OMP.

Hint: We have to again consider the quantity $\mu = \max_{i\neq j} \frac{|\mathbf{x}_i^\top \mathbf{x}_j|}{\|\|\mathbf{x}_i\|\mathbf{x}_j\|}$. Furthermore, the Gaussian tail bound from HW12 (Question 3d) might be useful here. For $Z_i \sim \mathcal{N}(0, \sigma^2)$ (not necessarily independent), we have

$$\Pr\left\{\max_{i\in\{1,2,\ldots,d\}} |Z_i| \geq 2\sigma\sqrt{\log d}\right\} \leq \frac{1}{d}.$$

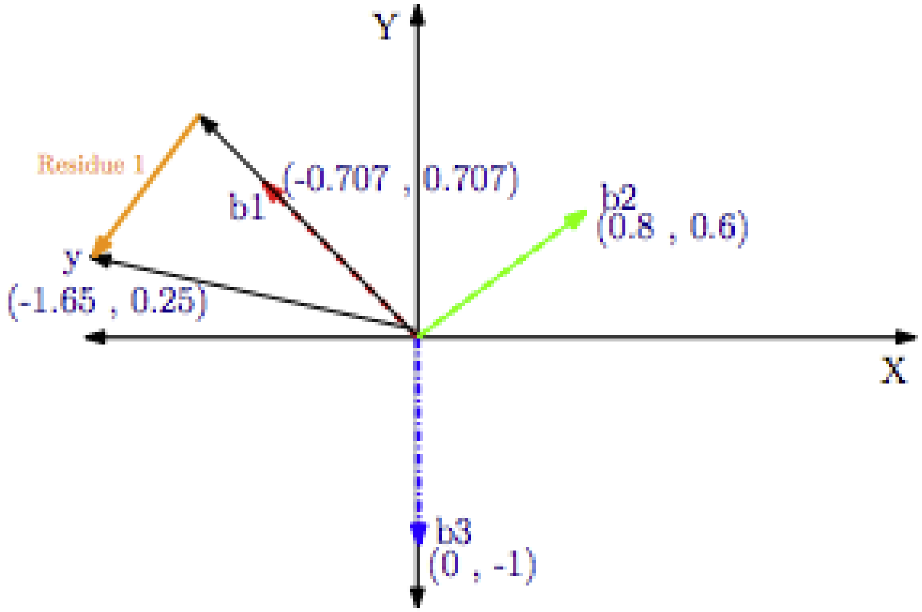

Figure 1: (Optional figures to be drawn on the board). Figure depicting y and residue 1 where b1, b2, and b3 are basis vectors of matrix X.

In other words, we have that the minimum and maximum of d Gaussian random variables with zero mean and σ^2 variance are bounded between $[-2\sigma\sqrt{\log d}, 2\sigma\sqrt{\log d}]$ with high probability (when d is large).

(c) OMP is also related to the classical idea of Boosting. As you have already seen (or may see in the next few lectures) boosting is a powerful recipe of training a set of weak learners (that do not fit the data very well by themselves) and combining them to find a strong learner (that fits the data well). The general idea is as follows: In the first step, we use one weak learner to fit a given dataset. In the next step, we use another weak learner to improve upon the first learner, by putting more weights on the data points that the first learner was unable to fit well (wrong classification or large squared error in regression). We repeat this process until a desirable accuracy is achieved.

Can you see OMP as an illustration of boosting for regression?

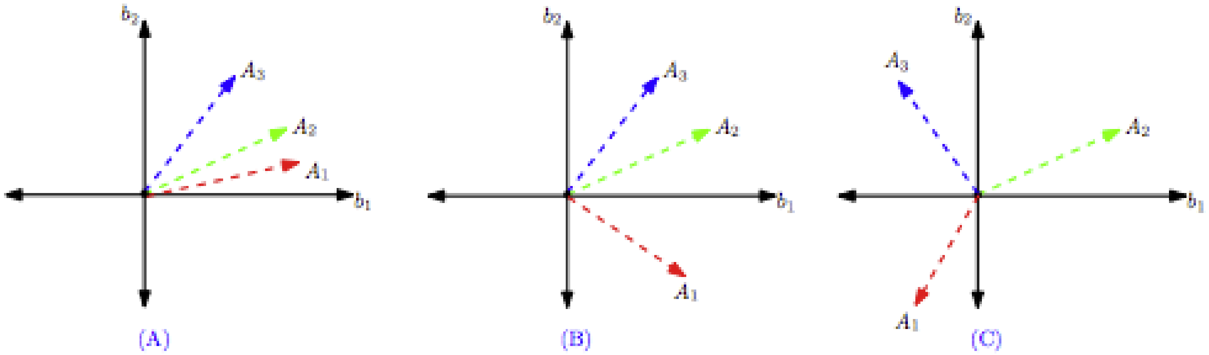

Figure 2: (Optional figures to be drawn on the board.) Figure showing decreasing μ