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 $\mathbf{X}_t = \left[\mathbf{x}_{j^1}, \dots, \mathbf{x}_{j^t}\right]$ 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?

Solution: For a 1-sparse signal, we should converge in one step with OMP and getting the argmax of the following. So at $t = 1$, we are solving for

$$\begin{aligned} j^1 &= \arg\min_i \min_v \|\mathbf{r}^{(1)} - \mathbf{v}\mathbf{x}_i\|^2 \\ &= \arg\min_i \|\mathbf{y} - \frac{(\mathbf{x}_i^\top y)}{\|\|^2}\mathbf{x}_i\|^2 \\ &= \arg\min_i \|\mathbf{y}\|^2 + \frac{(\mathbf{x}_i^\top \mathbf{y})^2}{\|\mathbf{x}_i\|^4}\|\mathbf{x}_i\|^2 - 2\frac{(\mathbf{x}_i^\top \mathbf{y})^2}{\|\mathbf{x}_i\|^2} \\ &= \arg\max_i \frac{(\mathbf{x}_i^\top \mathbf{y})^2}{\|\mathbf{x}_i\|^2} \end{aligned}$$

where the second step follows from knowing from our study of OLS that the value of $v = \frac{\mathbf{x}_i^\top \mathbf{y}}{\|\mathbf{x}_i\|^2} = \mathbf{x}_i^\top \mathbf{y}$. And the later step are just algebra.
Note that this makes sense, since we are trying to find the feature that best aligns with the target vector $\mathbf{y}$.

Now OMP works accurately when $j^1 = 1$ and for that to happen we need

$$\frac{(\mathbf{x}_1^\top \mathbf{y})^2}{\|\mathbf{x}_1\|^2} > \frac{(\mathbf{x}_i^\top \mathbf{y})^2}{\|\mathbf{x}_i\|^2}, \quad \text{for all } i = 2, \dots d.$$

Substituting $\mathbf{y} = \mathbf{x}_1 w_1^*$, we obtain that for exact recovery we need

$$\frac{(\mathbf{x}_1^\top \mathbf{x}_1)^2}{\|\mathbf{x}_1\|^2}(w_1^*)^2 > (w_1^*)^2 \frac{(\mathbf{x}_i^\top \mathbf{x}_1)^2}{\|\mathbf{x}_i\|^2} \iff \|\mathbf{x}_1\|\|\mathbf{x}_i\| > |\mathbf{x}_i^\top \mathbf{x}_1|,$$

which is true by Cauchy-Schwarz's inequality unless $\mathbf{x}_i$ perfectly aligns with $\mathbf{x}_1$. However, note that we don't know which of the w_i^*'s is non-zero and hence a stronger condition of the form

$$\mu = \max_{i \neq j} \frac{|\mathbf{x}_i^\top \mathbf{x}_j|}{\||\mathbf{x}_i\|\mathbf{x}_j\|} < 1$$

is necessary here. As it is clear, we do not need normalization for exact recovery since OLS solutions are not affected by the norms of the features.

More generally, for the case when the true signal is k-sparse and we have noiseless observation, we need $\mu < \frac{1}{2k-1}$. That is for more sparsity, we want the columns to span approximately orthogonal directions and have small inner product between each other.

(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

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

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).

Solution: Borrowing computations from the previous part and using the fact that $\|\mathbf{x}_i\| = 1$ and $\mathbf{y} = \mathbf{x}_1 w_1 + \mathbf{Z}$, we have

$$j^1 = \arg\max_i \frac{(\mathbf{x}_i^\top \mathbf{y})^2}{\|\mathbf{x}_i\|^2}$$

$$
\begin{aligned}
&= \arg\max_i (\mathbf{x}_i^\top \mathbf{y})^2 \qquad (\text{since } \|\mathbf{x}_i\| = 1)\\
&= \arg\max_i \left| \mathbf{x}_i^\top \left(\sum_{j=1}^{d} \mathbf{x}_j w_j^* + \mathbf{Z} \right) \right|\\
&= \arg\max_i \left| \mathbf{x}_i^\top \mathbf{x}_1 w_1^* + \sum_{j=2}^{d} \mathbf{x}_i^\top \mathbf{x}_j w_j^* + \mathbf{x}_i^\top \mathbf{z} \right\|\\
&= \arg\max_i \left| \mathbf{x}_i^\top \mathbf{x}_1 w_1^* + \underbrace{\mathbf{x}_i^\top \mathbf{z}}_{\xi_i} \right| \qquad (\text{since } w_j^* = 0, \text{ for } j \neq 1).
\end{aligned}
$$

The difference between this case and the noiseless case is the term $\mathbf{x}_i^\top \mathbf{z} = \xi_i$. Note that $\xi_i \sim \mathcal{N}(0, \sigma^2 \|\mathbf{x}_i\|^2) = \mathcal{N}(0, \sigma^2)$ and hence we can use the hint to conclude that

$$
\max_{i \in [d]} |\xi_i| \leq 2\sigma \sqrt{\log d} \quad \text{with probability at least} 1 - \frac{1}{d}.
$$

Let us condition on this event. When will OMP recover approximately the right signal (at least the right index) conditional to the event defined above. This will happen if $j^1 = 1$ which will happen if

$$
\left| \mathbf{x}_1^\top \mathbf{x}_1 w_1^* + \xi_1 \right| > \max_{i \neq 1} \left| \mathbf{x}_i^\top \mathbf{x}_1 w_1^* + \xi_i \right| .
$$

A sufficient condition to ensure this would be

$$
\begin{aligned}
& |\mathbf{x}_1^\top \mathbf{x}_1 w_1^*| - |\xi_1| > \max_{i \neq 1} |\mathbf{x}_i^\top \mathbf{x}_1| |w_1^*| + \max_{i \in [d]} |\xi_i|\\
\iff \qquad & |w_1^*|(1 - \max_{i \neq 1} |\mathbf{x}_i^\top \mathbf{x}_1|) > \max_{i \in [d]} |\xi_i| + |\xi_1|.
\end{aligned}
$$

Assume $\mu = |\max_{i \neq j} \mathbf{x}_i^\top \mathbf{x}_j|$, then if the signal strength is as strong as

$$
|w_1^*| \geq \frac{2}{(1-\mu)} \max_{i \in [d]} |\xi_i|
$$

then we have correct recovery of the index.

Using the high probability bound, we conclude that if

$$
|w_1^*| > \frac{1}{(1-\mu)} 4\sigma \sqrt{\log d},
$$

then the noisy corruptions in the signal are still overcome by the signal with probability at least $1 - 1/d$ since $|w_1^*| \geq \frac{2}{(1-\mu)} \max_{i \in [d]} |\xi_i|$ with probability at least $1 - 1/d$ and as a result, MP finds the right index with probability at least $1 - 1/d$.

Intuitively our computations quantify how large w_1^* needs to be as a function of the mutual coherence between columns, the noise variance and the dimension of the space.

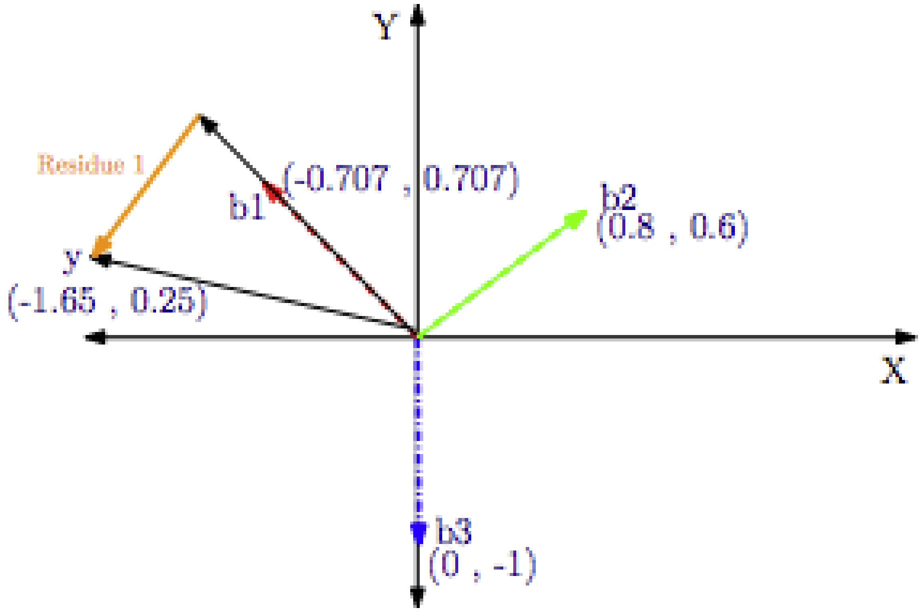

Figure 1: (Optional figures to be drawn on the board). Figure depicting y and residue 1 where b1, b2, and b3 are basis vectors of matrix X.

Ideally, if $\mu \ll 1$, i.e., the columns are approximately-orthogonal to each other or *columns are mutually incoherent* the signal needs to scale simply with $\sqrt{\log d}$ for exact recovery of the support. But when the signal align closely (that is $\mu \approx 1$), we really need a strong signal to have exact recovery of the support because the noise in the observations will confuse us between the closely aligned features.

(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?

Solution: Note that since we fit residuals, it is equivalent to improving upon the past weak learners. Also we only pick one learner at a time (hence we can consider our learners to be weak). However, the estimation update step, refits using all the features discovered so far, and hence strictly speaking the weak learner at next step is one new feature plus all the features identified so far.

In matching pursuit, we just have the index update step and hence the weak learner at each step

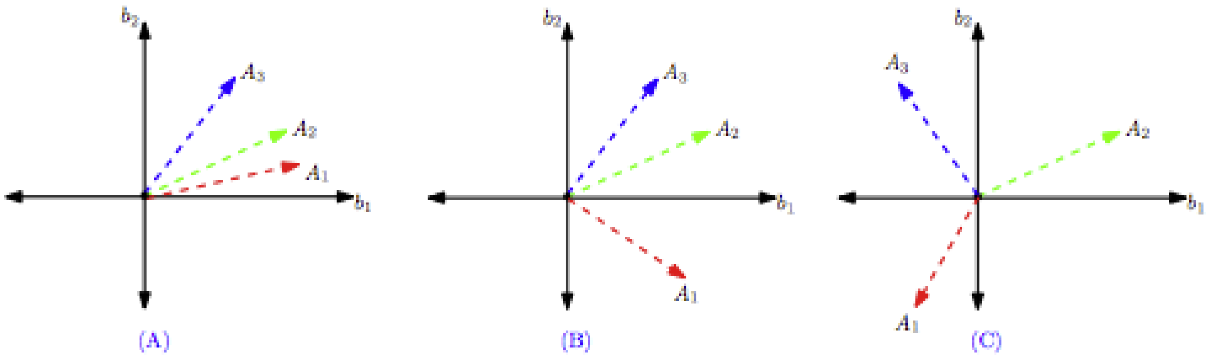

Figure 2: (Optional figures to be drawn on the board.) Figure showing decreasing μ

is just a new feature.