1 Overview

In today’s lecture we will discuss algorithms for sparse approximation with respect to an arbitrary basis. We will see that the LP from last lecture can easily be generalized to this case, at least for approximation in $L_2$ norm. For reference, here is the LP we considered last lecture:

$$\min ||x^*||_1 \text{ subject to}$$

$$Ax^* = Ax.$$ 

Here $Ax$ is the set of measurements we take, and $x^*$ is the reconstruction. The main result from last lecture is that there exists an $m \times n$ matrix $A$ with $m = \Theta(k \log(\frac{n}{k}))$ that has the following property:

Given $Ax$, the optimal solution $x^*$ to the LP above satisfies $||x - x^*||_2 \leq C \frac{\text{Err}_k^1(x)}{\sqrt{k}}$, for some fixed constant $C$. Here $\text{Err}_k^p(x)$ denotes the optimal error achievable by $k$-sparse vectors in the $L_p$ norm, i.e., $\text{Err}_k^p(x) = \min ||x''||_{0 \leq k} ||x - x''||_p$.

2 Remark About LP From Last Lecture

Before we discuss new problems, we want to make a remark about the construction from last lecture. One question we did not address last time is whether $x^*$ is actually $k$-sparse. The formula $||x - x^*||_2 \leq C \frac{\text{Err}_k^1(x)}{\sqrt{k}}$ immediately implies that $x^*$ in general will not be $k$-sparse. To see this, consider the case where $x$ is has exactly $k + 1$ nonzero entries. Then, $\text{Err}_k^1(x) = \text{Err}_k^2(x) = x^{(k+1)}$, where $x^{(k+1)}$ represents the smallest of the $k + 1$ nonzero entries in $x$. Thus, the formula above implies that for such $x$, the LP finds an $x^*$ that is better than the best $k$-sparse approximation, so clearly $x^*$ cannot be $k$-sparse.

In practice, it is often not important that $x^*$ be $k$-sparse. For example, in image processing, it may not be important that our reconstruction have exactly $k$ non-zero frequency components, so reconstructing $x$ to $x^*$ may be fine. However, if we do want to construct a $k$-sparse approximation $\tilde{x}$ to $x$, a simple application of the triangle inequality shows that setting $\tilde{x}$ to be the optimal $k$-sparse approximation to $x^*$ produces a good $k$-sparse approximation to $x$. Formally, let $x'$ denote the optimal $k$-sparse approximation to $x$. Then,

$$||x - \tilde{x}||_2 \leq ||x - x^*||_2 + ||x^* - \tilde{x}||_2 \leq ||x - x^*||_2 + ||x^* - x'||_2 \leq ||x - x^*||_2 + ||x^* - x||_2 + ||x - x'||_2 \leq \text{Err}_k^2(x) + 2C \frac{\text{Err}_k^1(x)}{\sqrt{k}}.$$ 

1Calculating $\tilde{x}$ is trivial once $x^*$ has been computed - just take the $k$ entries in $x^*$ with largest absolute value.
The first inequality is just the triangle inequality. The second inequality uses the fact that \( x' \) is \( k \)-sparse, and hence the approximation error when using \( x' \) to approximate \( x^* \) must be at least as large as the error when the optimal \( x \) is used. The third inequality uses the triangle inequality again, and the last inequality uses the definition of Err and the bound we proved last lecture. Thus, we see that constructing a good \( k \)-sparse approximation to \( x \) is simple once \( x^* \) has been found.

3 Compressed sensing in an arbitrary basis

Our LP from last lecture was formulated specifically for the standard basis. However, many signals in practice may be sparse in another basis, e.g., the Fourier basis or a wavelet basis. Today we will try to generalize our results from last lecture to an arbitrary basis \( B \). We would like to produce good \( k \)-sparse approximations for any \( L_p \) norm. For the standard basis, this was simple - the optimal \( k \)-sparse representation was always given by taking the largest \( k \) entries of \( x \), regardless of the value of \( p \). For a general basis \( B \), this is no longer true, i.e., we cannot simply choose the \( k \) vectors of \( B \) with the largest coefficients as our approximation. Here is an example of this for \( p = \infty \) in \( \mathbb{R}^3 \):

\[
\begin{align*}
    b_1 &= \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, & b_2 &= \begin{bmatrix} 0 \\ \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} \end{bmatrix}, & b_3 &= \begin{bmatrix} 0 \\ -\sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} \end{bmatrix} \\
    x &= .9b_1 + b_2
\end{align*}
\]

Consider the optimal 1-sparse approximation of \( x \). The naive approach would pick \( b_2 \) since this \( b_2 \) has the largest coefficient, and this results in an \( l_\infty \) error of .9 because the first entry of \( x \) is .9 while the first entry of \( b_2 \) is 0. Now, if we instead choose .9\( b_1 \) as our approximation, the \( l_\infty \) norm of the error is only \( \sqrt{\frac{1}{2}} < .9 \), showing that the naive approach does not work.

The reason our naive approach does not work for arbitrary \( L_p \) norms is that the \( L_p \) norm for \( p \neq 2 \) is not rotation invariant, so changing basis can have a big effect. However, we know that \( L_2 \) is invariant under rotation, so intuitively it would appear that the situation should be much simpler for \( L_2 \). As we will see, this intuition is correct.

It is fairly obvious what we should do if the signal is sparse in basis \( B \) instead of the standard basis. Let \( B = [b_1 b_2 \ldots b_n] \) be a representation of the basis as a matrix with columns \( b_1, b_2, \ldots, b_n \), where \( b_1, b_2, \ldots, b_n \) are the (orthonormal) vectors forming the basis. If we define \( u = B^{-1}x \), then \( u \) will have the same sparsity in the standard basis as \( x \) has in \( B \). So, if we apply the LP to \( u \), we should get a good approximation \( u^* \) to \( u \), and then we can set \( x^* = Bu^* \) as our approximation to \( x \). Let us write out the final LP:

\[
\min ||u^*||_1 \text{ subject to } A u^* = A u = AB^{-1} x
\]

Note that we are using the matrix \( AB^{-1} \) as a measurement matrix, so \( AB^{-1} x \) is the measurement vector.

\(^2\text{For the purposes of this lecture, whenever we say basis we mean an orthonormal basis.}\)
To see what are the guarantees for this approach, let us assume that the following hypothesis is true:

**Lemma 1.** (Hypothetical) Let $x$ be an arbitrary vector in $\mathbb{R}^n$, and let $M$ be an $m \times n$ matrix where each entry is chosen i.i.d. from $N(0,1)$. Then, the “usual” LP:

$$\min ||x^*||_1 \text{ subject to } Mx^* = Mx$$

has the following guarantee: with high probability over the choice of $A$, $||x - x^*||_2 \leq C\text{Err}_k^2(x)$ for some fixed constant $C$.

Note the order of the quantifiers: if $x$ is fixed, then most $A$’s will work. This does not mean that there is one $A$ that will work for all $x$ (which is what we have seen in the previous lecture, for the mixed L2/L1 guarantee). Also, it turns out that the hypothetical lemma is actually true! However, we will not cover the proof in lecture, as it has not been published yet.

Given the lemma, it is not hard to prove that our new LP (with $u^*$) solves the $L_2$ approximation problem. Specifically, consider the following chain of inequalities:

$$||Bu^* - x||_2 = ||u^* - B^{-1}x||_2 \leq C\text{Err}_k^2(B^{-1}x) = C \min_{||u'||_{0 \leq k}} ||u' - B^{-1}x||_2 \leq C \min_{||u'||_{0 \leq k}} ||Bu' - x||_2.$$

The first equality just says that $L_2$ is rotation invariant. The next inequality comes from the hypothetical lemma applied to the random matrix $M = AB^{-1}$. The third equality just expands out the definition of Err. Finally, the last equality uses the rotation-invariance of $L_2$ again.

The only technicality is that we need to show is that if $A$ has entries that are distributed i.i.d. $N(0,1)$, then so does $M = AB^{-1}$. This follows immediately from the fact that a collection of i.i.d. Gaussian random variables has a joint distribution that is spherically symmetric.

### 4 Mixed $L_1/L_2$ guarantee

Let us repeat the same analysis as in the previous section, but using the mixed $L_1/L_2$ guarantee from the previous lecture, instead of the hypothetical lemma. Then, the chain of inequalities becomes

$$||Bu^* - x||_2 = ||u^* - B^{-1}x||_2 \leq C\frac{\text{Err}_k^1(B^{-1}x)}{\sqrt{k}} = C\frac{\min_{||u'||_{0 \leq k}} ||u' - B^{-1}x||_1}{\sqrt{k}} \leq C \frac{\min_{||u'||_{0 \leq k}} ||Bu' - x||_{L_1(B)}}{\sqrt{k}}.$$
The first equality is the same as before, and in the second inequality we have used the $L_1/L_2$ guarantee proved in the last lecture. In the last equality, we have introduced the notation $\| \cdot \|_{L_1(B)}$. $\| \cdot \|_{L_1(B)}$ means the $L_1$ norm with respect to $B$, i.e., the sum of the absolute values of the coefficients when the vector is expressed in basis $B$. Thus, the last equality is not really saying anything new, i.e., it is just a tautology.

At first glance, it may seem that we have not really accomplished anything other than renaming the error using $\| \cdot \|_{L_1(B)}$. However, a little thought shows that $\| \cdot \|_{L_1(B)}$ is actually the useful norm for our problem. Specifically, the reason we are trying to generalize our results to $B$ is because we believe that the signal $x$ has a sparse approximation in basis $B$. If this is the case, then it is $\| \cdot \|_{L_1(B)}$, as opposed to the usual $\| \cdot \|_1$, that should be small. Thus, our translation of the mixed $L_1/L_2$ guarantee into a mixed $\| \cdot \|_{L_1(B)}/L_2$ guarantee is useful in that it captures the relevant notion of error for our problem.

5 Universality vs. Uniformity

An issue with our LP reconstruction algorithm is that we need to know the basis $B$ in advance, so that we can calculate the measurement matrix $AB^{-1}$. What we would really like to have is universality. This means that we should be able to make one set of measurements, and choose the basis used for reconstruction afterwards. The following argument due to Richard Baraniuk shows why this is a useful property. The Haar wavelets were one of the first wavelets to be discovered, and have been used in applications for many years. Starting about 15 years ago, however, the Daubechies wavelets have become more popular due to their superior performance. Who knows what future basis may be even better? If our reconstruction algorithm had the universality property, we could take one set of measurements and then reconstruct $x$ using the current “best” basis.

Now that we know that universality is desirable, how do we actually implement it? We have essentially already seen the answer. We previously observed that $A$ and $AB^{-1}$ have the same distribution. So, to achieve universality, we just pick the measurement matrix $M$ at the beginning, and pretend that the matrix we have chosen is actually equal to $AB^{-1}$. Then, we can make the measurements with respect to $M$, and we only need to know $B$ when the time comes to actually solve the LP and reconstruct an approximation to $x$.

Before finishing, it is important to highlight the distinction between universality and uniformity. Uniformity refers to the type of guarantee we proved in the last lecture, namely that there exists one $A$ that works simultaneously for all values of $x$. In some sense, universality and uniformity are mutually exclusive. In particular, universality is inherently a probabilistic guarantee. In other words, the following statement is false:

There exists an algorithm and an $m \times n$ matrix $A$ (with $m < n$) such that for all $x \in \mathbb{R}^n$ and all bases $B$, given the measurement $Ax$, the algorithm can recover an approximation $x^*$ with the property that $\|x^* - x\|_2 \leq \min_{x'} k$-sparse in $B$ $C\|x' - x\|_2$.

The choice of $L_2$ norm is not important, i.e., the statement remains false no matter what norm is chosen. There are many ways to see that the statement is false, but here we will mention only one. Set $k = 1$. Clearly, every vector $x$ is 1-sparse in a basis $B$ that includes a normalized version of $x$. Therefore, by choosing $B$ properly, we can make the right hand side of the inequality equal
to 0 for any $x$. Thus, the purported reconstruction algorithm must be able to recover exactly any $x \in \mathbb{R}^n$ from fewer than $n$ linear measurements, which is impossible.

The above false statement shows that universality and uniformity need to be interpreted carefully. Universality says that if we fix a vector $x$ and a basis $B$, then with high probability (over the choice of measurement matrix $A$) our LP can recover a “good” approximation to $x$, where “good” means the error with respect to sparse approximations in the (fixed) basis $B$. Uniformity lets us reverse the quantifiers: for any fixed basis, there exists a single $A$ which works for every $x \in \mathbb{R}^n$. Thus, by carefully specifying the quantifiers, we see that there is no contradiction with the previously mentioned false statement.

6 Conclusion

We showed that by modifying the LP from last lecture, we can perform compressed sensing with respect to an arbitrary basis and get provable reconstruction guarantees for the $L_2$ norm. Several generalizations are possible. For example, $A$ does not have to be chosen i.i.d. $\mathcal{N}(0,1)$. Any distribution satisfying the Johnson-Lindenstrauss lemma will also work [BDDW07]. In the next lecture, we will see combinatorial algorithms for compressed sensing.

References