Lecture 9: Sparse Recovery, JL Lemma

09-23-2025 Lecturer: Ali Vakilian | Scribe: Shih-Han Huang | Editor: Ali Vakilian

1 Sparse Recovery

In our last lecture, we introduced the concept of *sparsity*, a key structure in modern data analysis where data is dominated by a few significant values. We formalized this with the *sparse recovery problem*: given a vector x, find the best k-sparse approximation z (meaning z has at most k non-zero entries) that minimizes the error $\|x - z\|$. We concluded by showing that the optimal offline solution is a simple greedy algorithm: select the k entries in x with the largest absolute values and set all other entries to zero.

More formally, given a vector $x \in \mathbb{R}^n$ and an integer $k \geq 1$, the *optimal k-sparse approximation error* of x under the ℓ_2-norm is defined as $\mathrm{err}_k(x) = \min_{z:\|z\|_0 \leq k} \|x - z\|_2$. The optimal solution, denoted x_k, is the best k-sparse approximation of x. It can be found via **hard-thresholding**, which involves keeping the k entries of x with the largest absolute values and setting the rest to zero.

The error of this optimal solution is the norm of the "tail" of the vector, which contains the $n - k$ entries that were zeroed out: $\mathrm{err}_k^{(2)}(x) = \|\mathrm{tail}_k(x)\|_2$.

1.1 Sparse Recovery in the Streaming Model

Next, we focus on solving the sparse recovery problem in the streaming model. Formally,

Theorem 1.1 (Streaming ℓ_2-sparse recovery)**.** *There is a linear sketch of size $O\left(\frac{k}{\varepsilon^2}\,\mathrm{polylog}(n)\right)$ returning z with $\|z\|_0 \leq k$ such that w.h.p. $\|x - z\|_2 \leq (1+\varepsilon)\mathrm{err}_k^{(2)}(x)$.*

In particular, if x is exactly k-sparse, the algorithm recovers it exactly.

CountSketch Recap. CountSketch uses d rows, each with width w. For each row ℓ: $h_\ell : [n] \to [w]$, and $s_\ell : [n] \to \{-1, +1\}$. Moreover, the sketch maintains counters $C[\ell, 1..w] = 0$. Update on (i, Δ) in the stream is as follows: $C[\ell, h_\ell(i)] \leftarrow C[\ell, h_\ell(i)] + s_\ell(i) \cdot \Delta$.

Then, the estimate of the frequency of item i is computed as $\hat{x}_i = \mathrm{median}_{\ell \in [d]} \left(s_\ell(i) \cdot C[\ell, h_\ell(i)]\right)$.

Sparse Recovery via CountSketch. At a high level, the hope is that once we have estimated the frequency of each item, we can find a good k-sparse approximation by simply keeping the k coordinates with the largest absolute values. The CountSketch algorithm is a natural candidate for this approach.

However, the analysis of CountSketch from Lecture 6, which resulted in an error guarantee in terms of $\|x\|_2$, is insufficient for this task. To overcome this, we will now provide a tighter analysis showing that CountSketch can estimate frequencies with an additive error proportional to $\|\mathrm{tail}_k(x)\|_2$, which is precisely the optimal k-sparse recovery error, $\mathrm{err}_k(x)$.

Theorem 1.2 (Tighter analysis of CountSketch)**.** *If $w = \Theta(k/\varepsilon^2)$, $d = \Theta(\log n)$, then w.h.p.*

$$|\hat{x}_i - x_i| \leq \tfrac{\varepsilon}{k} \cdot \mathrm{err}_k^{(2)}(x), \quad \forall i \in [n].$$

Proof. Fix an index i. For a single row ℓ, define the row estimate

$$Z_\ell = s_\ell(i) \cdot C[\ell, h_\ell(i)] = x_i + \sum_{j \neq i} s_\ell(i) s_\ell(j) Y_j x_j,$$

where indicator random variable $Y_j = \mathbf{1}\{h_\ell(j) = h_\ell(i)\}$ indicates collision with item i in row ℓ. By pairwise independence of h_ℓ, $\mathbb{E}[Z_\ell] = x_i$.

To bound the error $\mathrm{err}_k(x)$, we will analyze collisions separately for two groups of items: those corresponding to the k largest coordinates of x and the remaining items in the tail. Let T_{big} be the indices of the k largest $|x_j|$ and $T_{\text{small}} = [n] \setminus T_{\text{big}}$. Note that $\sum_{j \in T_{\text{small}}} x_j^2 = \mathrm{err}_k^{(2)}(x)^2$.

For $j \in T_{\text{big}} \setminus \{i\}$, $\Pr[Y_j = 1] = 1/w$. Let $Y = \sum_{j \in T_{\text{big}} \setminus \{i\}} Y_j$. Then $\mathbb{E}[Y] \leq k/w$. Choosing width $w = \frac{3k}{\varepsilon^2}$, by Markov's inequality $\Pr[Y \geq 1] \leq \mathbb{E}[Y] \leq \frac{\varepsilon^2}{3}$. Thus with probability at least $1 - \varepsilon^2/3$, no large coordinate collides with i.

Conditioning on no collision with large coordinates/frequencies, for $j \in T_{\text{small}}$, $s_\ell(i)s_\ell(j)Y_j x_j$ has mean 0 and variance $\mathbb{E}[Y_j] = x_j^2/w$. Hence

$$\mathrm{Var}\Big(\sum_{j \in T_{\text{small}}} s_\ell(i)s_\ell(j)Y_j x_j\Big) = \frac{1}{w}\sum_{j \in T_{\text{small}}} x_j^2 = \frac{\mathrm{err}_k^{(2)}(x)^2}{w}.$$

By Chebyshev,

$$\Pr\left(|Z_\ell - x_i| > \tfrac{\varepsilon}{k}\mathrm{err}_k^{(2)}(x) \;\middle|\; \text{no large collision}\right) \leq \frac{\mathrm{err}_k^{(2)}(x)^2/w}{(\varepsilon\,\mathrm{err}_k^{(2)}(x)/k)^2} = \frac{k^2}{\varepsilon^2} \cdot \frac{1}{w} = \frac{1}{3}.$$

By union bound, for sufficiently small values of ε, the failure probability per row is

$$\Pr\left(|Z_\ell - x_i| > \tfrac{\varepsilon}{k} \cdot \mathrm{err}_k^{(2)}(x)\right) \leq \frac{\varepsilon^2}{3} + \frac{1}{3} \leq \tfrac{2}{5}.$$

Thus each estimate deviates from the true frequency by at most $\frac{\varepsilon}{k} \cdot \mathrm{err}_k^{(2)}(x)$, with probability at least $3/5$. Take $d = \Theta(\log n)$ independent rows. The median of $Z_1, \ldots, Z_d$ is within the desired error bound with probability $\geq 1 - n^{-2}$ by Chernoff bounds. By union bound over all $i \in [n]$, the failure probability is at most $1/n$. Thus w.h.p. simultaneously for all i,

$$|\widehat{x}_i - x_i| \leq \tfrac{\varepsilon}{k} \cdot \mathrm{err}_k^{(2)}(x).$$

□

Theorem 1.3. *If* $\|x-y\|_\infty \leq \frac{\varepsilon}{k} \cdot \mathrm{err}_k^{(2)}(x)$ *and* z *keeps the* k *largest* $|y_i|$, *then* $\|x-z\|_2 \leq (1+O(\varepsilon)) \cdot \mathrm{err}_k^{(2)}(x)$.

Proof. Let T be the indices of the k largest $|x_i|$, and S be the indices of the k largest $|y_i|$. Let $\delta = \frac{\varepsilon}{k}\mathrm{err}_k^{(2)}(x)$. If $i \in T \setminus S$, then some $j \in S \setminus T$ must replace it, and $|x_i| - |x_j| \leq 2\delta$ by triangle inequality since $|y_i - x_i| \leq \delta$ and $|y_j - x_j| \leq \delta$. Thus swaps occur only among nearly equal magnitudes. So,

$$\|x - z\|_2^2 \leq \|x_{T\cap S} - y_{T\cap S}\|_2^2 + \|x_{T\setminus S}\|_2^2 + \|y_{S\setminus T}\|_2^2 + \|x_{\text{tail}(k)}\|_2^2.$$

Each term is bounded by either $\|x - y\|_\infty$ (on shared indices) or δ (on swapped indices). Hence the total additional error is at most $O(\varepsilon)\,\mathrm{err}_k^{(2)}(x)^2$. Thus $\|x - z\|_2 \leq (1 + O(\varepsilon))\,\mathrm{err}_k^{(2)}(x)$. □

Algorithm 1 Sparse Recovery via CountSketch (ℓ_2)

```
Require: Stream updates $(i_t, \Delta_t)$; sparsity $k$; accuracy $\varepsilon$; universe size $n$
Ensure: $k$-sparse $z$ with $\|x - z\|_2 \le (1+\varepsilon) \cdot \mathrm{err}_k^{(2)}(x)$
$w \leftarrow \lceil c_2\, k/\varepsilon^2 \rceil$, and $d \leftarrow \lceil c_1 \log n \rceil$
maintain a CountSketch $C$ with $d$ rows and $w$ width and updates it in the stream.
                                        ▷ estimate coordinates after the stream
for $i \in [n]$ do
    for $\ell = 1$ to $d$ do
        $\widehat{x}_i^{\ell} \leftarrow g_\ell(i) \cdot C[\ell, h_\ell(i)]$
    $\widehat{x}_i \leftarrow \mathrm{median}\{\widehat{x}_i^1, \dots, \widehat{x}_i^d\}$
$\widehat{S} \leftarrow$ indices of the top-$k$ values of $|\widehat{x}_i|$
for $i \in [n]$ do
    if $i \in \widehat{S}$ then $z_i \leftarrow \widehat{x}_i$ else $z_i \leftarrow 0$
return $z$
```

We remark that the space complexity of the proposed algorithm is $O(\frac{k}{\varepsilon^2} \cdot \log n)$.

2 Dimensionality Reduction

In many modern data science and machine learning applications, we work with data that lives in a very high-dimensional space (large d). For example, an image can be represented as a vector of pixel values, or a document as a vector of word counts. Processing and analyzing such high-dimensional data can be computationally expensive and statistically challenging, a phenomenon often called the "curse of dimensionality".

Dimensionality reduction. is a set of techniques aimed at transforming data from a high-dimensional space into a lower-dimensional space (small k) while preserving some essential structure of the original data. The primary motivation is to make computations more tractable and to remove irrelevant or redundant features, which can lead to better performance in downstream tasks like classification, clustering, and visualization. A key question is: can we find a mapping $f : \mathbb{R}^d \to \mathbb{R}^k$ (with $k \ll d$) that preserves geometric properties, such as distances between points?

2.1 The Johnson-Lindenstrauss (JL) Property

The Johnson-Lindenstrauss (JL) Lemma provides a remarkable answer to this question, showing that it's possible to project points into a much lower dimension while approximately preserving their pairwise Euclidean distances. There are two main flavors of this result.

Distributional JL Lemma. The distributional version considers a single vector and guarantees that its norm is preserved with high probability when projected by a random matrix.

Theorem 2.1 (Distributional JL). *Fix a vector $x \in \mathbb{R}^d$. Let $\Pi \in \mathbb{R}^{k \times d}$ be a random matrix whose entries are drawn independently from a standard normal distribution, $\mathcal{N}(0,1)$. If the target dimension is $k = \Omega(\varepsilon^{-2} \log(1/\delta))$, then with probability at least $1-\delta$, the squared norm of the projected vector is preserved up to a factor of $(1 \pm \varepsilon)$:*

$$\Pr\left[\frac{1}{k}\|\Pi x\|_2^2 = (1 \pm \varepsilon)\|x\|_2^2\right] \ge 1 - \delta.$$

Metric JL Lemma. The metric version extends this idea from a single vector to a set of n points, guaranteeing that all pairwise distances are simultaneously preserved by a single projection.

Theorem 2.2 (Metric JL). *For any set of n points $v_1, \ldots, v_n \in \mathbb{R}^d$, there exists a linear map $f(v) = \Pi v$ that projects them into a space of dimension $k \leq 8\ln(n)/\varepsilon^2$ such that for all pairs of distinct points (v_i, v_j):*

$$(1-\varepsilon)\|v_i - v_j\|_2 \leq \|f(v_i) - f(v_j)\|_2 \leq (1+\varepsilon)\|v_i - v_j\|_2.$$

This result is incredibly powerful because the required dimension k depends only logarithmically on the number of points n, and not at all on the original dimension d.

Proof. Consider the $N = \binom{n}{2}$ difference vectors $u_{ij} = v_i - v_j$. By the DJL lemma, for a fixed u,

$$\frac{1}{m}\|\Pi u\|^2 = (1 \pm \varepsilon)\|u\|^2$$

with probability at least $1 - \delta'$, if $m = \Omega(\varepsilon^{-2}\log(1/\delta'))$. Applying this to each of the N vectors, the union bound gives overall success probability at least $1 - N\delta'$. Choosing $\delta' = 1/n^2$ gives $N\delta' \leq 1/2$. Thus with positive probability all distances are preserved, and $m = O(\varepsilon^{-2}\log n)$. □