# Compressed Sampling and Sparse Recovery

Zhiwei Ying

Used for the reading group in spring semester, 2024

Main Technique: $\ell_1/\ell_1$ Guarantee, Restricted Isometry Property (RIP).

The contents are mainly from *Beyond the Worst-Case Analysis of Algorithm*[3] by Tim Roughgarden.

## 1 Preliminary

**Definition 1.1.** *A vector $x$ is **$k$-sparse** if the number of non-zero entries of $x$ is $k$.*

If the entry of $x$ is not zero but so small that we can regard it as 0 and the number of meaningful entries is $k$ ($x$ is close to a $k$-sparse vector), we say that $x$ is **approximately $k$-sparse**.

**Definition 1.2.** *$H_k(x)$ denotes the k-sparse vector of $x$ in $\mathbb{R}^n$ that sets all but the largest $k$ entries of $x$ to zero.*

**Definition 1.3.** *For $I \subset [n]$, $x_I$ is the vector that only has the coordinates in $I$ of $x$.*

## 2 Streaming Algorithms

First, consider the scenario of a vector $u \in \mathbb{R}^n$, and $u$ starting as the 0 vector. Then we see a sequence of updates $(i, \Delta)$ each causing the change $u_i \leftarrow u_i + \Delta$, where $i \in [n]$ and $\Delta \in \mathbb{Z}$.

### 2.1 Model 1: Election Counting (For each update, $\Delta = 1$.)

For a giant stack of ballots of $n$ candidates, there are only $k$ real candidates and all the others are rare. If using a sheet of paper to record the count for each candidate, the recording sheet would be very long and most of the space would be wasted on rare candidates.

Formally, consider an insertion-only data stream: $u_1, u_2, \ldots, u_N \in [n]$, the count vector $x \in \mathbb{R}^n$: $x_i = \left|\{j : u_j = i\}\right|$. Our goal is to approximate $x$ when scanning through $u$ while storing much less than $n$ or $\|x\|_1$ space.

You should notice a special case $aabbbccd (k = 2, n = 4) \longrightarrow bd$, from which we can see that the algorithm is not always correct.

**Theorem 2.1.** *For Frequent Element Algorithm, the estimates $\hat{x}_u = d[u]$ satisfy that for every $u$,*

$$x_u - \frac{1}{k+1}\sum_i x_i \le \hat{x}_u \le x_u.$$

**Algorithm 1:** Frequent Element (Stream, Space $d$ of Size $k$) [For Insertion-Only Setting]

```
for $u$ in Stream do
  if $u$ in $d$ then
    $d[u] += 1$ (Addition step)
  else if $d$ is not full then
    $d[u] \leftarrow 1$ (Addition step)
  else
    $d[u'] -= 1\ \forall u' \in d$ and remove keys of $d$ that now map to zero (Subtraction step)
  end if
end for
return $d$
```

*Proof.* Consider $\sum_u d[u]$, the sum of $d[u]$ in the statistical space $d$. The addition step adds one to $\sum_u d[u]$ each time. The subtraction step subtracts $\sum_u d[u]$ by $k$ at a time because only when the statistical space is full, the subtraction step will be processed. Since $\sum_u d[u]$ is always non-negative, subtraction is done at most $\frac{1}{k+1}\sum_i x_i$ times. □

**Theorem 2.2.** *For Frequent Element Algorithm, the estimates $\hat{x}_u = d[u]$ satisfy that for every $u$,*

$$x_u - \frac{2}{k}\sum_{x_i \notin H_{k/2}(x)} x_i \le \hat{x}_u \le x_u.$$

*Proof Sketch.* Consider the sum $\sigma$ of $d[u]$ for the smaller $\frac{k}{2}$ term in the statistical space $d$. The $\sigma$ performs at most $\sum_{x_i \notin H_{k/2}(x)} x_i$ additions. Presumably, if we add in the wrong place, we can always find one that wasn't added before. Each subtraction of $\sigma$ subtracts at least $k/2$. Since $\sigma$ is non-negative, subtraction is done at most $\frac{2}{k}\sum_{x_i \notin H_{k/2}(x)} x_i$ times. □

**Claim 2.3.** *Theorem 2.2 is better when $x$ is sparser, and Theorem 2.1 is better when $x$ is more even.*

## 2.2 Model 2: Strict turnstile ($\Delta$ can be negative, but $\forall i, x_i \ge 0$ at all times.)

Consider the scenario of counting the number of people in each building, each person can enter and exit the building at any time, but a person must enter a building before they exit the building, which ensures that at any time, the number of people in each building is non-negative. Three algorithms are given below to conquer this model. Please note that Count Median Sketch can handle the problem where $\exists i, x_i < 0$ for some time.

Naturally, the first idea comes out that if we have limited space and then meet collisions, we can ignore them and just store in each hash cell the total number of elements that hash there. A simple analysis is that a cell containing the true count may contain other colliding elements. And any other element has only $\frac{1}{B}$ chance of colliding in this cell, where $B$ is the space of the hash table. Therefore, the expected error is at most $\|x\|_1 / B$. Undoubtedly, some elements will have much higher errors. Thus, we can repeat the process several times to get the minimum estimate. This is the basic idea of Count Min Sketch.

**Theorem 2.4.** *For Count Min Sketch Algorithm, if $B \ge 4k$ and $R \ge 2\log_2 n$, then for all $u$ with $1 - 1/n$ probability,*

$$x_u \le \hat{x}_u \le x_u + \frac{1}{k}\|x - H_k(x)\|_1 .$$

**Algorithm 2:** Count Min Sketch/ Count Median Sketch/ Count Sketch (Turnstile Stream, Space Size $B$, # of Round $R$ ) [For Insertion-Deletion Setting]

```
Pick $R$ pairwise independent hash functions $h_1, \dots, h_R : [n] \to [B]$.
(For Count Sketch) Pick $R$ pairwise independent sign function $s_1, \dots, s_R : [n] \to \{-1, 1\}$.
$y_i^{(r)} \leftarrow 0 \quad \forall i \in [B], r \in [R]$
for $(u, a)$ in Stream do
  for $r \in [R]$ do
    (For Count Sketch) $y_{h_r(u)}^{(r)} += a \cdot s_r(u)$.
    (For Count Min Sketch and Count Median Sketch) $y_{h_r(u)}^{(r)} += a$.
  end for
end for
for $u \in [n]$ do
  (For Count Sketch) $\hat{x}_u \leftarrow \text{median}_{r \in [R]} y_{h_r(u)} \cdot s_r(u)$.
  (For Count Min Sketch) $\hat{x}_u \leftarrow \min_{r \in [R]} y_{h_r(u)}$.
  (For Count Median Sketch) $\hat{x}_u \leftarrow \text{median}_{r \in [R]} y_{h_r(u)}$.
end for
return $\hat{x}_u$
```

Recall the Markov's Inequality. For a non-negative random variable $X$, $\forall a > 0. \Pr[X \geq a] \leq \frac{\mathbb{E}[X]}{a}$.

*Proof.* Let $H \subset [n]$ be the set of largest $k$ entries of $x$. Define $\hat{x}_u^{(r)} = y_{h_r(u)}^{(r)}$. Then,

$$0 \leq \hat{x}_u^{(r)} - x_u = \sum_{\substack{h_r(v) = h_r(u) \\ v \neq u}} x_v = \sum_{\substack{h_r(v) = h_r(u) \\ v \neq u,\ v \in H}} x_v + \sum_{\substack{h_r(v) = h_r(u) \\ v \neq u,\ v \notin H}} x_v = E_H + E_L,$$

where $E_H = \sum_{\substack{h_r(v) \neq h_r(u) \\ v \neq u,\ v \in H}} x_v$ and $E_L = \sum_{\substack{h_r(v) \neq h_r(u) \\ v \neq u,\ v \notin H}} x_v$. For these two parts, we can obtain that

$$\begin{aligned}\Pr[E_H > 0] &\leq \Pr[\exists v \in H \backslash \{u\}.\ h_r(v) = h_r(u) \wedge x_v > 0] \\ &\leq \Pr[\exists v \in H \backslash \{u\}.\ h_r(v) = h_r(u)] \leq \frac{k}{B}.\end{aligned}$$

$$\begin{aligned}\Pr[E_L \geq \frac{1}{k} \|x - H_k(x)\|_1] &\leq \frac{k \cdot \mathbb{E}[E_L]}{\|x - H_k(x)\|_1} \quad \text{(by Markov's Inequality)} \\ &= \frac{k}{\|x - H_k(x)\|_1} \sum_{v \neq u,\ v \notin H} x_v \cdot \Pr[h_r(v) = h_r(u)] \\ &= \frac{k}{B \cdot \|x - H_k(x)\|_1} \sum_{v \neq u,\ v \notin H} x_v \leq \frac{k}{B}.\end{aligned}$$

Therefore,

$$\begin{aligned}\Pr[\hat{x}_u^{(r)} - x_u \geq \frac{1}{k} \|x - H_k(x)\|_1] &= \Pr[E_H + E_L \geq \frac{1}{k} \|x - H_k(x)\|_1] \\ &\leq \Pr[E_H > 0] + \Pr[E_L \geq \|x - H_k(x)\|_1] \\ &\leq \frac{2k}{B}.\end{aligned}$$

$$\begin{aligned}\Pr[\hat{x}_u - x_u \ge \frac{1}{k}\|x - H_k(x)\|_1] &= \Pr[\min_r\{\hat{x}_u^{(r)}\} - x_u \ge \frac{1}{k}\|x - H_k(x)\|_1] \\ &= \Pr[\hat{x}_u^{(r)} - x_u \ge \frac{1}{k}\|x - H_k(x)\|_1]^R \\ &\le (\frac{2k}{B})^R \le \frac{1}{n^2}.\end{aligned}$$

□

**Theorem 2.5.** *For Count Median Sketch Algorithm, if $B \ge 16k$ and $R \ge 4\log_2 n$, then for all $u$ with $1 - 1/n$ probability,*

$$\hat{x}_u - x_u \le \frac{1}{k}\|x - H_k(x)\|_1 .$$

*Proof.* Since $2^R = C_R^0 + C_R^1 + \cdots + C_R^R > C_R^{\frac{R}{2}}$, the probability of that $\hat{x}_u - x_u \ge \frac{1}{k}\|x - H_k(x)\|_1$ happens in at least $R/2$ of the $R$ repetitions is then at most $C_R^{\frac{R}{2}} \cdot (\frac{2k}{B})^{\frac{R}{2}} < (\frac{8k}{B})^{\frac{R}{2}} \le \frac{1}{n^2}$. □

**Theorem 2.6.** *For Count Sketch Algorithm, if $B \ge 16k$ and $R \ge 4\log_2 n$, then for all $u$ with $1-1/n$ probability,*

$$(\hat{x}_u - x_u)^2 \le \frac{1}{k}\|x - H_k(x)\|_2 .$$

*Proof.* Let $H \subset [n]$ be the set of largest $k$ entries of $x$. Define $\hat{x}_u^{(r)} = y_{h_r(u)}^{(r)}$. Then,

$$\hat{x}_u^{(r)} - x_u = \sum_{\substack{h_r(v)\neq h_r(u)\\ v\neq u}} x_v s_r(u) s_r(v) = \sum_{\substack{h_r(v)\neq h_r(u)\\ v\neq u,\ v\in H}} x_v s_r(u) s_r(v) + \sum_{\substack{h_r(v)\neq h_r(u)\\ v\neq u,\ v\notin H}} x_v s_r(u) s_r(v) = E'_H + E'_L,$$

where $E'_H = \sum_{\substack{h_r(v)\neq h_r(u)\\ v\neq u,\ v\in H}} x_v s_r(u) s_r(v)$ and $E'_L = \sum_{\substack{h_r(v)\neq h_r(u)\\ v\neq u,\ v\notin H}} x_v s_r(u) s_r(v)$. Since $s_r$ is pairwise independent and mean zero, all cross terms in $\mathbb{E}_{s_r}[(\hat{x}_u^r - x_u)^2]$ disappear. Then,

$$\mathbb{E}_{h_r,s_r}[E_L'^2] = \sum_{\substack{v\in[n]\setminus H\\ v\neq u}} x_v^2 \cdot \Pr[h_r(v) = h_r(u)] \le \frac{\|x - H_k(x)\|_2^2}{B}.$$

Then, we can obtain that $\Pr[E_H'^2 > 0] = \Pr[E_H > 0] \le \frac{k}{B}$ and $\Pr[E_L'^2 > \frac{1}{k}\|x - H_k(x)\|_2^2] \le \frac{k}{B}$. Combine these two inequalities, $\Pr[(\hat{x}_u - x_u)^2 \le \frac{1}{k}\|x - H_k(x)\|_2] \le \frac{2k}{B}$. The probability of that $(\hat{x}_u - x_u)^2 \ge \frac{1}{k}\|x - H_k(x)\|_2$ happens in at least $R/2$ of the $R$ repetitions is then at most $C_R^{\frac{R}{2}} \cdot (\frac{2k}{B})^{\frac{R}{2}} < (\frac{8k}{B})^{\frac{R}{2}} \le \frac{1}{n^2}$. □

Note that the square of $\ell_1$ norm $\frac{1}{k^2}\|x - H_k(x)\|_1^2 = \frac{1}{k^2}(\sum_{x_i \notin H_k(x)} x_i)^2$, and the $\ell_2$ norm $\frac{1}{k}\|x - H_k(x)\|_2^2 = \frac{1}{k}\sum_{x_i\notin H_k(x)} x_i^2$. An observation is that the $\ell_2$ bound is stronger than the $\ell_1$ bound for every vector $x$. A simple fact is that by counting the number of $x_i^2$, the number in the $\ell_1$ norm is $n^2/k^2$ while in the $\ell_2$ norm it is $n/k$. Below we are going to examine it in detail. First, we can use $\|\hat{x} - x\|_\infty$ to denote $\max_u\{\hat{x}_u - x_u\}$.

Consider power-law distributions where the $i$th largest element has frequency proportional to $i^{-a}$, and the number of coordinates $n \gg k$ is finite. Then the $\ell_1$ guarantee is

$$\begin{aligned}\|\hat{x}-x\|_\infty &\le \frac{1}{k}\sum_{i=k+1}^{n} x_i \approx \frac{1}{k}x_1\sum_{i=k+1}^{n} i^{-\alpha} \approx \frac{1}{k}x_1\int_k^n x^{-\alpha}\mathrm{d}x \\ &\approx \begin{cases}\frac{1}{\alpha-1}x_k, & \text{for sharply decaying distributions of } \alpha > 1 \\ \frac{1}{k(1-\alpha)}x_1 n^{1-\alpha}, & \text{for intermediate decaying distributions of } 0 < \alpha < 1\end{cases}\end{aligned}$$

while the $\ell_2$ guarantee for $\alpha > 0.5$ is

$$\begin{aligned}\|\hat{x}-x\|_\infty &\le \sqrt{\frac{1}{k}\sum_{i=k+1}^{n} x_i^2} \approx \sqrt{\frac{1}{k}x_1^2\sum_{i=k+1}^{n} i^{-2\alpha}} \approx \sqrt{\frac{1}{k}x_1^2\int_k^n x^{-2\alpha}\mathrm{d}x} \\ &\approx \begin{cases}\frac{1}{\sqrt{2\alpha-1}}x_k, & \text{for sharply decaying distributions of } \alpha > 0.5 \\ \frac{1}{\sqrt{1-2\alpha}}\sqrt{\frac{n}{k}}x_1 n^{-\alpha}, & \text{for intermediate decaying distributions of } 0 < \alpha < 0.5\end{cases}\end{aligned}$$

## 3 Restricted Isometry Property

Let's focus on compressed sampling now. A general form is that $Ax = b$, where $A \in \mathbb{R}^{m\times n}$ is a matrix, $x \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$ are a column vector. If $r(A) \ll n$, $x$ can be solved to a space of rank $n - \min\{m, r(A)\}$.

But if we know that $x$ is $k$-sparse ahead of all, maybe a much less $m$ is enough to recover $x$. For example, assume $x$ is a $k$-sparse 0-1 vector and each element of $b$ represents one bit of information. So what we need to recover $x$ is at most $C_n^k$. Thus $m = O(\log C_n^k) = O(k\log n)$. If $k$ is $O(\sqrt{n})$, $O(\log n)$ or much smaller, then $m \ll n$ is enough in this scenario. A natural problem is that: is $O(k\log n)$ is the lower bound of sparsity recovery problem?

If all non-zero elements in $A$ are concentrated in the first several columns, then the bigger $m$ cannot bring us more information about the last elements of $x$. Informally, we can believe that if we want to recover $x$, then $m \ge k\log n$ and $A$ could not be too sparse. Let's look backward to the matrix multiplication. In a way, it's a compression from data with $n$ dimension to the other data with $m$ dimension. And somehow if $\|x\|_2$ and $\|b\|_2 = \|Ax\|_2$ are close, the compression is good enough, or we can say $A$ is "good enough".

**Definition 3.1.** *For any $k$, the restricted isometry constant $\delta_k = \delta_k(A)$ of a matrix $A \in \mathbb{R}^{m\times n}$ is the smallest $\delta$ s.t. for all $k$-sparse $x$, $(1-\delta)\|x\|_2^2 \le \|Ax\|_2^2 \le (1+\delta)\|x\|_2^2$. An equivalent formulation is that $\|(A^\top A - I)_{S\times S}\| \le \delta$ for all $S \subset [n]$, $|S| \le k$.*

Informally, we say that $A$ satisfies the Restricted Isometry Property if $\delta_{Ck} < c$ for some sufficiently large constant $C \ge 1$ and sufficiently small $c < 1$. We'll show that RIP implies that approximate $k$-sparse recovery is possible afterward. Here we're going to discuss how can we construct such an RIP matrix.

**Theorem 3.2.** *Let $0 < \varepsilon < 1$ and $k > 1$ be parameters. If $A \in \mathbb{R}^{m\times n}$ has i.i.d. Gaussian entries of variance $\frac{1}{m}$, and $m > C\frac{1}{\varepsilon^2}k\log\frac{n}{k}$ for a sufficiently large constant $C$, then $A$ has RIP constant $\delta_k < \varepsilon$ with $1 - e^{-\Omega(\varepsilon^2 m)}$ probability.*

To prove Theorem 3.2, we should start from some lemmas. A basic lemma is about the operator norm of matrices.

**Lemma 3.3.** *The spectral/operator norm of a real symmetric matrix $M$ is the maximum of its eigenvalues, i.e.*

$$\|M\| = \max_{\|x\|_2=1} \|Mx\| = \sup_{\|x\|_2=1} x^\top Mx = \lambda_{max}.$$

*Proof.* $\exists \alpha_1, \ldots, \alpha_n \in \mathbb{R}$, *s.t.* $x = \sum_{i=1}^n \alpha_i v_i$, where $\{v_1, \ldots, v_n\}$ is a orthogonal basis of $M$. Therefore,

$$\|Mx\| = \left\| \sum_{i=1}^n \alpha_i \lambda_i v_i \right\| \leq \sum_{i=1}^n \|\alpha_i \lambda_i v_i\| \leq \lambda_{max} \sum_{i=1}^n \|\alpha_i v_i\| = \lambda_{max}.$$

□

The lemma below considers the size of a cover of a unit ball.

**Lemma 3.4.** *For any $\varepsilon > 0$, the number $N(\varepsilon)$ of $\ell_2$ balls of radius $\varepsilon$ that can cover a unit $\ell_2$ ball satisfies that $(\frac{1}{\varepsilon})^n \leq N(\varepsilon) \leq (\frac{2}{\varepsilon}+1)^n$.*

*Proof.* For the lower bound, assume a unit ball's volume is $V_u$. Then a ball of radius $\varepsilon$'s volume is $\varepsilon^n V_u$. For the upper bound, note that the minimum number of closed balls of radius $\varepsilon$ that can form a cover for the unit ball is less than or equal to the maximum number of disjoint closed balls of radius $\frac{\varepsilon}{2}$. (A detailed proof is in Appendix A.) And all these disjoint closed balls are in a ball of radius $1 + \frac{\varepsilon}{2}$. Thus, the upper bound is $\frac{(1+\frac{\varepsilon}{2})^n}{(\frac{\varepsilon}{2})^n}$. □

**Lemma 3.5.** *There exists a set $T \subset \mathbb{R}^n$ of $5^n$ unit vectors such that, for any symmetric matrix $M \in \mathbb{R}^{n\times n}$, $\|M\| \leq 4 \max_{x\in T} x^\top Mx$.*

*Proof.* By Lemma 3.4, we can choose $T$ to be a 1/2-cover of the unit $\ell_2$ ball $\mathcal{B}$, *i.e.* for every $x \in \mathcal{B}$, there exists an $x'$ in $T$ such that $\|x' - x\|_2 \leq \frac{1}{2}$. Without loss of generality, assume $\|M\| = x^\top Mx$. Since $M$ is a symmetric matrix, $M$ is orthogonally diagonalizable, *i.e.* $M = Q\Lambda Q^{-1}$, where $Q = [v_1, \ldots, v_n]$ is a orthogonal matrix and for all $i$, $v_i$ is the unit eigenvector corresponding to eigenvalue $\lambda_i$. (We can get these eigenvectors by Gram-Schmidt Process on $M$.) Thus, $\exists \alpha_1, \ldots, \alpha_n \in \mathbb{R}$, *s.t.* $x' = \sum_{i=1}^n \alpha_i v_i$. Assume $x^\top Q = [\beta_1, \ldots, \beta_n]$, $\sum_{i=1}^n \beta_i = 1$. First, we will show that $(x-x')^\top Mx' \leq \frac{1}{4}\|M\|$.

$$\begin{aligned}
(x - x')^\top Mx' &= (x - x')^\top Q\Lambda Q^{-1}x' \\
&= (x^\top Q - [\alpha_1, \ldots, \alpha_n])\Lambda[\alpha_1, \ldots, \alpha_n]^\top \\
&= [\beta_1 - \alpha_1, \ldots, \beta_n - \alpha_n] \cdot [\lambda_1\alpha_1, \ldots, \lambda_n\alpha_n]^\top \\
&= \sum_{i=1}^n \lambda_i \alpha_i (\beta_i - \alpha_i) \\
&\leq \frac{1}{4}\lambda_{max} \sum_{i=1}^n \beta_i^2 \leq \frac{1}{4}\lambda_{max} = \frac{1}{4}\|M\|.
\end{aligned}$$

Then we have that

$$\begin{aligned}
\|M\| = x^\top Mx &= (x - x' + x')^\top M(x - x' + x') \\
&= (x - x')^\top M(x - x') + x'^\top Mx' + 2(x - x')^\top Mx' \\
&\leq \frac{1}{4}\|M\| + \max_{x\in T} x^\top Mx + 2(x - x')^\top Mx' \\
&\leq \frac{3}{4}\|M\| + \max_{x\in T} x^\top Mx.
\end{aligned}$$

□

**Lemma 3.6** (Johnson-Lindenstrauss)**.** *Let $\mathcal{X}$ be a set of $N$ points in $\mathbb{R}^n$ and $\varepsilon > 0$. Assume that $m \geq O(\frac{1}{\varepsilon^2}\log N)$. If $A \in \mathcal{R}^{m\times n}$ has i.i.d. Gaussian entries of variance $\frac{1}{m}$,*

$$\Pr\left[\left|\|Ax - Ay\|_2 - \|x - y\|_2\right| > \varepsilon \|x - y\|_2\right] < 2e^{-\Omega(\varepsilon^2 m)}.$$

Lemma 3.6 gives a high-level intuition that regardless of how many dimensions $N$ vectors were, we can reduce them to $O(\log N)$ dimensions and keep the relative distances. The proof of Lemma 3.6 is based on two intuitive lemma.

First, we show that the $\ell_2$ norm of a random vector sampling from the normal distribution with variance $\frac{1}{n}$ is likely 1.

**Lemma 3.7.** *Let $u \in \mathbb{R}^n$ be a vector of independent samples from $\mathcal{N}(0, \frac{1}{n})$ and $\varepsilon \in (0, 1)$ be a given constant, then we have $\Pr[\left|\|u\|_2^2 - 1\right| \geq \varepsilon] \leq 2e^{-\frac{\varepsilon^2 n}{8}}$.*

*Proof.* $\Pr[\|x\|_2^2 - 1 \geq \varepsilon] = \Pr[e^{\lambda\|x\|_2^2 - \lambda} \geq e^{\lambda\varepsilon}] \leq \min_{\lambda>0} e^{-\lambda\varepsilon-\lambda}E[e^{\lambda\|x\|_2^2}] = \min_{\lambda>0} e^{-\lambda\varepsilon-\lambda}\Pi_i E[e^{\lambda x_i^2}] = \min_{\lambda>0} e^{-\lambda\varepsilon-\lambda}\Pi_i \int_{-\infty}^{\infty} e^{\lambda\frac{x_i^2}{n}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}} dx_i = \min_{\lambda>0} e^{-\lambda\varepsilon-\lambda}(\frac{n}{n-2\lambda})^{\frac{n}{2}} = e^{\frac{n}{2}(\ln 1+\varepsilon-\varepsilon)} \leq e^{-\frac{n\varepsilon^2}{8}}$. Similarly, we can prove that $\Pr[1 - \|x\|_2^2 \geq \varepsilon] \leq e^{-\frac{n\varepsilon^2}{8}}$. □

The second lemma tells us that if $n$ is large, the product of two random vectors is likely 0, *i.e.* these two vectors are almost orthogonal.

**Lemma 3.8.** *Let $u, v \in \mathbb{R}^n$ be two vectors independently sampled from $\mathcal{N}(0, \frac{1}{n})$ and $\varepsilon \in (0, 1)$ be a given constant, then we have $\Pr[|\langle u, v\rangle| \geq \varepsilon] \leq 4e^{-\frac{\varepsilon^2 n}{8}}$.*

*Proof.* Note that adding $\|\frac{u+v}{\sqrt{2}}\|_2^2 - 1 \geq \varepsilon$ and $1 - \|\frac{u-v}{\sqrt{2}}\|_2^2 \geq \varepsilon$ up, we can get $\langle u, v\rangle \geq \varepsilon$. Therefore, $\Pr[\langle u, v\rangle \geq \varepsilon] \leq \Pr[\|\frac{u+v}{\sqrt{2}}\|_2^2 - 1 \geq \varepsilon] + \Pr[1 - \|\frac{u-v}{\sqrt{2}}\|_2^2 \geq \varepsilon] \leq 2e^{-\frac{\varepsilon^2 n}{8}}$. Similarly, we can prove that $\Pr[-\langle u, v\rangle \geq \varepsilon \leq 2e^{-\frac{\varepsilon^2 n}{8}}$. □

With these two lemmas, we yield that a random matrix $M \in \mathbb{R}^{n\times n}$ that uniformly sampling from a normal distribution $\mathcal{N}(0, \frac{1}{n})$ can be regarded as an orthogonal matrix. Now we can begin the proof of Lemma 3.6.

*Proof of Lemma 3.6.* With Lemma 3.7 and Lemma 3.8, we obtain that

$$\Pr[\left|\left\|\frac{A(v_i - v_j)}{\|v_i - v_j\|}\right\|_2^2 - 1\right| \geq \varepsilon] \leq 2e^{-\frac{\varepsilon^2 m}{8}}.$$

Then,

$$\Pr[\exists(i, j) : \left|\left\|\frac{A(v_i - v_j)}{\|v_i - v_j\|}\right\|_2^2 - 1\right| \geq \varepsilon] \leq 2C_N^2 e^{-\frac{\varepsilon^2 m}{8}}.$$

□

Now we can continue the proof of Theorem 3.2, which implies that a random matrix is good enough with high probability.

*Proof of Theorem 3.2.* By Lemma 3.5, there exist a set $T \subset \mathbb{R}^k$ of $5^k$ unit vectors such that for every set $S \subset [n]$ of size $k$,

$$\left\|(A^\top A - I)_{S\times S}\right\| \leq 4 \max_{x\in T} x^\top (A^\top A - I)_{S\times S} x.$$

By Lemma 3.6,

$$\Pr[x^\top (A^\top A - I)_{S\times S} x \leq \frac{\varepsilon}{4} \|x\|_2^2] \geq 1 - 2e^{-\varepsilon^2 m}.$$

Thus, take a union bound over all $S$ and $x \in T$, (using Stirling's approximation $n! \approx \sqrt{2\pi n}(\frac{n}{e})^n$,)

$$\begin{aligned}\Pr[\delta_k \leq \varepsilon] &\geq \Pr[\delta_k \leq 4 \max_S \max_{x\in T} x^\top (A^\top A - I)_{S\times S} x] \\ &\geq 1 - 2C_n^k 5^k e^{-\Omega(\varepsilon^2 m)} = 1 - 2e^{-(\Omega(\varepsilon^2 m) - k \log \frac{n}{k})} \geq 1 - 2e^{-\Omega(\varepsilon^2 m)}.\end{aligned}$$

□

# 4 Recovery Algorithms

## 4.1 Definition of Sparse Recovery Problem

Given $Ax$, recover a $k$-sparse vector $\hat{x}$ such that

$$\|x - \hat{x}\|_p \leq C \min_{k\text{-sparse } x'} \|x - x'\|_q$$

for some norm parameters $p$ and $q$ and an approximation factor $C = C(k)$.

## 4.2 Iterative Methods

Suppose $y = Ax^*$, which means that there is no noise (both postmeasurement noise and premeasurement noise, since $x^*$ is an exactly $k$-sparse vector). With matrix $A$ that satisfies RIP, we have that $A^\top y = A^\top A x^* \approx x^*$. A nice property we are going to show is $\|H_k(A^\top y) - x^*\|_2 \leq O(\delta_{2k}) \|x^*\|_2$. Therefore, $x^{(1)} = H_k(A^\top y)$ is a good first step in recovering $x$. Then, to take the residual error into account, since $y - Ax^{(1)} = A(x^* - x^{(1)})$, we compute $x^{(2)} = H_k(x^{(1)} + A^\top(y - Ax^{(1)}))$. The recovery algorithm is given below. We'll find that even with noise ($y = Ax^* + e$), the algorithm still works well.

**Algorithm 3:** Iterative Hard Thresholding (IHT)

```
$x^{(0)} \leftarrow 0$
for $r \leftarrow 0, 1, \ldots, R-1$ do
    $x^{(r+1)} \leftarrow H_k\left(x^{(r)} + A^\top(y - Ax^{(r)})\right)$
end for
return $x^{(R)}$
```

To analyze this algorithm, we should start with a basic lemma.

**Lemma 4.1.** *$x \in \mathbb{R}^n$ is $k$-sparse with support $S$, $z \in \mathbb{R}^n$ and the largest $k$ entries of $z$ is in $T \subset [n]$. Then*

$$\|x - z_T\|_2^2 \leq 3 \|(x - z)_{S\cup T}\|_2^2.$$

*Proof.* For every $i \in S \backslash T$ we can assign a unique $j \in T \backslash S$ such that $|z_j| \geq |z_i|$. Therefore,

$$x_i^2 \leq (|x_i - z_i| + |z_i|)^2 \leq (|x_i - z_i| + |z_j|)^2 \leq 2(x_i - z_i)^2 + 2z_j^2.$$

Thus,

$$\begin{aligned}
\|x - z_T\|_2^2 &= \sum_{i \in S \cap T} (x_i - z_i)^2 + \sum_{i \in S \backslash T} x_i^2 + \sum_{i \in T \backslash S} z_i^2 \\
&\leq \sum_{i \in S \cap T} (x_i - z_i)^2 + 2 \sum_{i \in S \backslash T} (x_i - z_i)^2 + 3 \sum_{i \in T \backslash S} z_i^2 \\
&\leq 3( \sum_{i \in S \cap T} (x_i - z_i)^2 + \sum_{i \in T \backslash S} (x_i - z_i)^2) \\
&= 3 \sum_{i \in S \cup T} (x_i - z_i)^2 \\
&= 3 \, \|(x - z)_{S \cup T}\|_2^2 .
\end{aligned}$$

□

**Lemma 4.2.** *In each iteration of IHT,*

$$\|x^{(r+1)} - x^*\|_2 \leq \sqrt{3}\delta_{3k}\|x^{(r)} - x^*\|_2 + \sqrt{6}\|e\|_2,$$

*where* $e = y - Ax^*$.

*Proof.* Define $x' := x^{(r)} + A^\top(y - Ax^{(r)}) = x^* + (A^\top A - I)(x^* - x^{(r)}) + A^\top e$. Let $S = \mathrm{supp}(x^{(r+1)}) \cup \mathrm{supp}(x^{(r)}) \cup \mathrm{supp}(x^*)$, so $|S| \leq 3k$. Note that the RIP implies that $\|A_S^\top\|^2 = \|(A^\top A)_{S \times S}\| \leq 1 + \delta_{3k}$. Therefore,

$$\begin{aligned}
\|(x' - x^*)_S\|_2 &\leq \|((A^\top A - I)(x^* - x^{(r)}))_S\|_2 + \|(A^\top e)_S\|_2 \\
&\leq \|(A^\top A - I)_{S \times S}\| \cdot \|x^* - x^{(r)}\|_2 + \|A_S^\top\| \cdot \|e\|_2 \\
&\leq \delta_{3k}\|x^* - x^{(r)}\|_2 + \sqrt{1 + \delta_{3k}}\|e\|_2.
\end{aligned}$$

Finally, with Lemma 4.1,

$$\|x^{(r+1)} - x^*\|_2 \leq \sqrt{3}\|(x' - x^*)_S\|_2 = \sqrt{3}\delta_{3k}\|x^* - x^{(r)}\|_2 + \sqrt{3 + 3\delta_{3k}}\|e\|_2.$$

□

We get the theorem below by applying Lemma 4.2.

**Theorem 4.3.** *If* $\delta_{3k} < \frac{1}{4\sqrt{3}}$*, the output* $x^{(R)}$ *of IHT will have* $\|x^{(R)} - x^*\|_2 \leq \sqrt{24}\|e\|_2$ *after* $R = \log_2 \frac{\|x^*\|_2}{\|e\|_2}$ *iterations.*

## 4.3 L1 Minimization

L1 minimization is another good method for sparsity recovery. In the field of machine learning, it is also called the **LASSO**. The intuition is that since the true $x^*$ is $k$-sparse, one would like to find the sparsest vector $\hat{x}$ that satisfies $\|y - A\hat{x}\|_2 \leq R$ for some external estimate $R$ on the noise $\|e\|_2$. However, finding the sparsest $\hat{x}$ is a hard non-convex optimization problem, so we settle for minimizing its convex relaxation $\|\hat{x}\|_1$.

**Theorem 4.4.** *Let* $A \in \mathbb{R}^{m \times n}$ *has RIP constant* $\delta_{2k} < 0.62$*. For any* $k$*-sparse* $x \in \mathbb{R}^n$ *and any* $e \in \mathbb{R}^m$*, and any* $R \geq \|e\|_2$*, there exists a constant* $C > 0$*, such that the L1 minimization result* $\hat{x}$ *(where the input* $y = Ax + e$*) satisfies* $\|\hat{x} - x^*\|_2 \leq CR$*.*

You can find the proof of Theorem 4.4 in [1].

```
Algorithm 4: L1 Minimization
    $\hat{x} \leftarrow \arg\min_{\|y-Ax'\|_2 \le R} \|x'\|_1$
    return $\hat{x}$
```

## 5 Lower Bound of Linear Measurement for $\ell_1/\ell_1$ Guarantee[2]

The $\ell_1/\ell_1$ guarantee for a linear sparse recovery algorithm refers to that $\|\hat{x}-x\|_1 \le O(1)\cdot\|x-H_k(x)\|_1$. Note that in the preceding sections, we have proved that both Count Min Sketch and Count Median Sketch achieve this guarantee with $O(k\log n)$ linear measurements, and Iterative Hard Thresholding and L1 Minimization achieve this with $O(k\log\frac{n}{k})$ Gaussian linear measurements. We are going to show that $O(k\log\frac{n}{k})$ is the lower bound for sparse recovery with $\ell_1/\ell_1$ guarantee.

**Corollary 5.1** (Corollary of Lemma 3.4)**.** *Take an $m\times n$ real matrix $A$, positive reals $\varepsilon, p, \lambda$, and $Y \subset B_p^n(\lambda)$, where we use $B_p^n(r)$ to denote the $\ell_p$ ball of radius $r$ in $\mathbb{R}^n$. If $|Y| > (1+\frac{1}{\varepsilon})^m$, then there exist $z, \bar{z} \in B_p^n(\varepsilon\lambda)$ and $y, \bar{y} \in Y$ with $y \ne \bar{y}$ and $A(y+z) = A(\bar{y}+\bar{z})$.*

**Lemma 5.2.** *For any $q, k \in \mathbb{Z}^+, \varepsilon \in \mathbb{R}^+$ with $\varepsilon < 1-\frac{1}{q}$, there exists a set $Y \subset \{0,1\}^{qk}$ of binary vectors with exactly $k$ ones, s.t. $Y$ has minimum Hamming distance $2\varepsilon k$ and $|Y| > e^{(1-H_q(\varepsilon))k\log q}$, where the $q$-ary entropy function $H_q(x) = -x\log_q\frac{x}{q-1} - (1-x)\log_q(1-x)$, and by convention, we define $H_q(0) = H_q(1) = 0$.*

**Definition 5.3** (Volume of a Hamming ball)**.** *Let $q \ge 2$ and $n \ge r \ge 1$ be integers. The volume of a Hamming ball of radius $r$ is given by*

$$Vol_q(r,n) = |B_q(\mathbf{0},r)| = \sum_{i=0}^{r} C_n^i (q-1)^i.$$

**Lemma 5.4.** *For $0 < \varepsilon < 1-\frac{1}{q}$, $q^{nH_q(\varepsilon)-o(n)} \le Vol_q(\varepsilon n, n) \le q^{nH_q(\varepsilon)}$.*

Hint: For the RHS, note that $1 = (\varepsilon + (1-\varepsilon))^k$; for the LHS, use Stirling's approximation.

*Proof of Lemma 5.2.* We will construct a codebook $T$ of block length $k$, alphabet $q$, and minimum Hamming distance $\epsilon k$. Replacing each character $i$ with the $q$-long standard basis vector $e_i$ will create a binary $qk$-dimensional codebook $S$ with minimum Hamming distance $2\epsilon k$ of the same size as $T$, where each element of $S$ has exactly $k$ ones.

The Gilbert-Varshamov bound, based on volumes of Hamming balls, states that a codebook of size $L$ exists for some

$$L \ge \frac{q^k}{\sum_{i=0}^{\epsilon k-1}\binom{k}{i}(q-1)^i}.$$

Using the Lemma 5.4 that for $\epsilon < 1-1/q$,

$$\sum_{i=0}^{\epsilon k}\binom{k}{i}(q-1)^i < q^{kH_q(\epsilon)},$$

we have that $\log L > (1-H_q(\epsilon))k\log q$, as desired. □

**Theorem 5.5.** *Any $\ell_1/\ell_1$ linear sparse recovery algorithm with constant approximation factor $C$ and a fixed matrix $A$ requires at least $\frac{1-H_{\lfloor\frac{n}{k}\rfloor}(\frac{1}{2})}{\log(4+2C)}k\log\lfloor\frac{n}{k}\rfloor = \Omega(k\log\frac{n}{k})$ linear measurements.*

*Proof.* By Lemma 5.2, let $q = \lfloor \frac{n}{k} \rfloor$, $\varepsilon = \frac{1}{2}$, we have $\log|Y| > (1 - H_{\lfloor \frac{n}{k} \rfloor}(\frac{1}{2}))k\log\lfloor \frac{n}{k} \rfloor$. Let $\gamma = \frac{1}{3+2C}$. Suppose that the theorem is not true; then $m < \frac{\log|Y|}{\log(1+\frac{1}{\gamma})}$, or $|Y| > (1+\frac{1}{\gamma})^m$. Then by Corollary 5.1, we can find some $z, \bar{z} \in B_p^n(\gamma k)$ and $y, \bar{y} \in Y$ with $A(y+z) = A(\bar{y}+\bar{z})$.

Let $w$ be the result of running the recovery algorithm on $A(y+z)$. We have

$$\begin{aligned}
&\|y+z-w\|_1 \le C\min_{y'}\|y+z-H_k(y')\|_1 \\
\Rightarrow&\|y-w\|_1 - \|z\|_1 \le C\|z\|_1 \\
\Rightarrow&\|y-w\|_1 \le (1+C)\|z\|_1 \le \frac{1+C}{3+2C}k.
\end{aligned}$$

and similarly, $\|\bar{y}-w\|_1 \le \frac{1+C}{3+2C}k$. Thus $\|\bar{y}-y\|_1 \le \frac{2+2C}{3+2C} < k$, which contradicts the definition of $Y$. □

**Theorem 5.6.** *Any $\ell_1/\ell_1$ linear sparse recovery algorithm with constant approximation factor $C$, a randomized matrix $A$ and constant success probability requires $\Omega(k\log\frac{n}{k})$ linear measurements.*

You can find the proof of Theorem 5.6 in [2].

## References

[1] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. Communications on Pure and Applied Mathematics, 59(8):1207–1223, 2006. doi: https://doi.org/10.1002/cpa.20124. 9

[2] Khanh Do Ba, Piotr Indyk, Eric Price, and David P. Woodruff. Lower bounds for sparse recovery. In Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '10, page 1190–1197, 2010. ISBN 9780898716986. 10, 11

[3] Eric Price. Sparse Recovery, page 140–164. 1

## A Minimum Covering and Maximal Packing

You can read some reference materials in `https://en.wikipedia.org/wiki/Delone_set`.

If $(M,d)$ is a metric space, and $X$ is a subset of $M$, then the packing radius, $r$, of $X$ is half of the smallest distance between distinct members of $X$. Open balls of radius $r$ centered at the points of $X$ will all be disjoint from each other. The covering radius, $R$, of $X$ is the smallest distance such that every point of $M$ is within distance $R$ of at least one point in $X$; that is, $R$ is the smallest radius such that closed balls of that radius centered at the points of $X$ have all of $M$ as their union.

An $\varepsilon$-packing is a set $X$ of packing radius $r \ge \frac{\varepsilon}{2}$ (equivalently, minimum distance $\ge \varepsilon$), and an $\varepsilon$-covering is a set $X$ of covering radius $R \le \varepsilon$. Let $N(M,\varepsilon) = \min\{card(X)|X \text{ is an } \varepsilon\text{-covering of } M\}$, and $P(M,\varepsilon) = \max\{card(X)|X \text{ is an } \varepsilon\text{-packing of } M\}$.

**Theorem A.1.** *$P(M,\varepsilon) \ge N(M,\varepsilon)$.*

*Proof.* For any $X$, which is a maximal $\varepsilon$-packing of $M$, and any point $x \in M$, we will show that $\exists x_0 \in X,\ d(x_0,x) \le \varepsilon$. If $x \in X$, then choose $x_0 = x$. Otherwise, if $x \notin X$, since $X \cup \{x\}$ is not an $\varepsilon$-packing of $M$, there exists a point $x' \in X$ *s.t.* $d(x,x') \le \varepsilon$. □