

Lema de Johnson-Lindenstrauss y Aplicaciones

Manuel Horn

DM, UBA Exactas, Buenos Aires, Argentina.

manuhorn1910@gmail.com;

Resumen

Se presenta el lema de Johnson-Lindenstrauss y sus aplicaciones en el área de pre-procesamiento de datos para poder aplicar en diversos ámbitos como en *Clustering*, del aprendizaje no supervisado, o en sistemas de recomendación mediante la aproximación/completitud de matrices de bajo rango.

Probabilidad y Machine Learning
Junio/Julio 2024

1

En el análisis de datos, trabajar con dimensiones altas de ellos es habitual. Esto se traduce a tener n puntos en \mathbb{R}^p , con p grande.

Esto mismo puede traer problemas ya que vamos a tener que trabajar con matrices de dimensiones elevadas, que es computacionalmente costoso. Otro problema que viene de la mano de tener dimensión grande es la renombrada maldición de la dimensión, por lo que nos va a interesar reducir la dimensión a un espacio razonablemente manejable. Muchas veces los datos suelen vivir en una superficie de dimensión (digamos k), con $k \ll p$, por lo que nos va a interesar transformarlos a dicho espacio.

Necesitamos una noción de similitud entre puntos, que nos va a permitir discernir entre grupos con características similares; por ejemplo en el problema de clustering, del aprendizaje no supervisado.

Veamos como influye la maldición de la dimensión en la noción de distancia entre puntos:

Sea en \mathbb{R}^D la bola de radio unitario y su volumen $V_D(1) = \omega_D$, tenemos que el volumen de la bola de radio r es $V_D(r) = \omega_D r^D$. Luego, podemos ver que el cociente de la franja entre las bolas de radios $r = 1$ y $r = 1 - \varepsilon$ es

$$\frac{V_D(1) - V_D(1 - \varepsilon)}{V_D(1)} = 1 - (1 - \varepsilon)^D \xrightarrow{D \rightarrow \infty} 1$$

Esto nos dice que a medida de que aumentamos la dimensión, la concentración de puntos está en dicha franja (cáscara) de la bola unitaria. Si sampleamos puntos con alguna distribución, las distancias de ellos al origen se concentran en dicha franja, que se achica arbitrariamente con D tendiendo a infinito. Es decir, las distancias entre puntos dejan de ser distinguibles a dimensiones altas. Más detalle en [1], [2].

Luego, bajo las hipótesis de que los datos provienen de una distribución $f : \mathcal{M} \subset \mathbb{R}^D \rightarrow \mathbb{R}_{\geq 0}$, con M una superficie de dimensión d , contamos con el lema de Johnson-Lindenstrauss, que nos permite reducir la dimensión de los datos logarítmicamente.

Lema 1.1. *Johnson-Lindenstrauss:*

Sean $x_1, x_2, \dots, x_n \in \mathbb{R}^D$ puntos arbitrarios y $\varepsilon > 0$. Luego, para $d = \mathcal{O}(\log(n)/\varepsilon^2)$ existen puntos $y_1, y_2, \dots, y_n \in \mathbb{R}^d$ tal que

$$(1 - \varepsilon)\|x_i - x_j\| \leq \|y_i - y_j\| \leq (1 + \varepsilon)\|x_i - x_j\| \quad \forall i, j$$

Más aún, es posible encontrar una transformación lineal $L : \mathbb{R}^D \rightarrow \mathbb{R}^d$ tal que $L(x_i) = y_i$

Demostración del lema y cotas en [3]

Nos encontramos ante el problema de aprender una representación en menor dimensión que refleje la estructura de los datos (*Manifold Learning*) y esto está relacionado con aprender una distancia (*Metric Learning*).

Existen muchas técnicas que permiten reducir la dimensionalidad de los datos, tanto lineales (*PCA, etc.*), como no lineales (*t-SNE, Isomap, etc.*). El objetivo de este artículo es hablar un poco acerca de la reducción aleatoria de la dimensión, junto con un resultado interesante aplicado a la completitud de matrices de bajo rango, usado en sistemas de recomendación.

1.1 The Fast Johnson-Lindenstrauss Transform (FJLT)

Esta es una reversión de las proyecciones aleatorias, que en tiempo polinomial encuentran un subespacio sobre el cual los datos se proyectan. Esto en cada paso necesita computar el producto matricial Mx , con M una matriz de proyección aleatoria, y esto se hace en $\mathcal{O}(dp) = \mathcal{O}(\varepsilon^{-2} \log(n))$, por J-L. Esto puede ser costoso en muchos

casos, y la alternativa, la FJLT, permite computar el producto matricial en $\mathcal{O}(p \log(p) + d) = \mathcal{O}(p \log(p) + \varepsilon^{-2} \log(n))$. Aquí definimos p como la dimensión original del problema (n puntos en \mathbb{R}^p) y d la dimensión reducida.

Teorema 1. *Fast Johnson-Lindenstrauss Transform:*

Existe una matriz aleatoria Φ de dimensión $d \times p$ con $d = \mathcal{O}(\log(n/\delta) \log(1/\delta) / \varepsilon^2)$ tal que, para cada $x \in \mathbb{R}^p$,

$$\|\Phi x\|_2 \in [1 - \varepsilon, 1 + \varepsilon] \cdot \|x\|_2$$

se cumple con probabilidad al menos $1 - \delta$. El cómputo de multiplicación matrix-vector es $\mathcal{O}(p \log(p) + d)$

Veamos primero porqué es esto así.

Definamos una matriz S dispersa en $d \times p$, donde cada fila de S tiene una entrada no nula de valor $\sqrt{p/d}$ en una ubicación aleatoria (notar que la probabilidad de que una posición en la fila i -ésima de S es $1/p$). Así, para un punto $x \in \mathbb{R}^p$, tenemos

$$\mathbb{E}[(Sx)_i^2] = \sum_{j=1}^p \mathbb{P}(i = j) (\sqrt{p/d} x_j)^2 = \frac{1}{d} \|x\|_2^2$$

Consecuentemente:

$$\mathbb{E}(\|Sx\|_2^2) = \|x\|_2^2$$

Esto es favorable en el sentido de la esperanza de la transformación, pero en la varianza no. Para garantizar que $\|Sx\|_2^2 \neq 0$ necesitamos $d \sim \mathcal{O}(p)$; y si x tiene un valor mucho más grande que el resto en valor absoluto, necesitamos d grande para que $\|Sx\|_2 \simeq \|x\|_2$.

Vamos a medir cuán “puntiagudo” es un vector: para esto introducimos un cociente llamado “peak-to-average ratio”:

$$\frac{1}{\sqrt{p}} \leq \frac{\|x\|_\infty}{\|x\|_2} \leq 1$$

por lo que si $\frac{\|x\|_\infty}{\|x\|_2} \simeq \frac{1}{\sqrt{p}}$, un sampleo raro de x nos preserva la norma Euclídea.

Con estas ideas en mente, primero vamos a mapear vectores raros a no raros en la nueva base, reduciendo la norma infinito y manteniendo constante la norma Euclídea bajo rotaciones. Pero como necesitamos que los vectores densos se mapeen a densos también, randomizamos la transformación mediante la FJLT:

Definición 1. *Fast Johnson-Lindenstrauss Transform:*

Tenemos el mapa $\Phi : \mathbb{R}^p \rightarrow \mathbb{R}^d$ definido por $\Phi := SFD$, con S y D dos matrices

aleatorias y F determinística. Particularmente:

S es una matriz en $d \times p$, donde cada fila de S tiene una entrada no nula de valor $\sqrt{p/d}$ cuya ubicación es elegida uniformemente.

F es la matriz $p \times p$ de Fourier o la de Hadamard (si existe), normalizada por $\frac{1}{\sqrt{p}}$, para obtener una matriz unitaria.

D es una matriz $p \times p$ diagonal con sus entradas en $-1, 1$ con probabilidad $1/2$

Vemos que podemos computar FDx en $\mathcal{O}(p \log p)$ y el producto restante con S en $\mathcal{O}(d)$, resultando en $\mathcal{O}(d + p \log p)$.

En comparación al algoritmo estándar de reducción de la dimensión, la FJLT es más eficiente porque las matrices involucradas presentan estructura dispersa. Vamos a mostrar que el teorema 1 se cumple. Para eso vamos a enunciar dos lemas que nos van a ayudar:

Lema 1.2. Sea $y := FDx$, donde F y D son las matrices de la FJLT, entonces

$$\mathbb{P}\left(\frac{\|y\|_\infty}{\|y\|_2} \geq \frac{2 \log(4p/\delta)}{p}\right) \leq \frac{\delta}{2}$$

Proof. Como FD es unitaria, preserva la norma euclídea de x , por lo que podemos suponer que $\|y\|_2 = \|x\|_2 = 1$, luego $\|y\|_\infty / \|y\|_2 = \|y\|_\infty$. Como las entradas de D son 1 o -1 , tenemos $y_i = \sum_{j=1}^p \varepsilon_j F_{ij} x_j$, y como los términos son iid., con esperanza 0 y acotadas, usando la desigualdad de Hoeffding:

$$\mathbb{P}(|y_i| > \sqrt{\frac{2 \log(4p/\delta)}{p}}) \leq 2 \exp - \frac{2 \log(4p/\delta)/p}{2/p} = \delta/2p$$

$$\text{Con } \sum_{j=1}^p F_{ij}^2 x_j^2 = 1/p \|x\|_2^2 = 1/p$$

Observar que las entradas de la matriz de Fourier elevadas al cuadrado tienen magnitud $1/p$

Luego, un uso del “union bound” termina la demostración. \square

Lema 1.3. Como ya podemos asumir que $\|y\|_\infty > \frac{2 \log(4p/\delta)}{p}$, se tiene:

$$\mathbb{P}(|\|Sy\|_2^2 - 1| \leq \varepsilon) \leq 1 - \frac{\delta}{2}$$

Proof. Usando el “Chernoff bound”: dadas (X_1, X_2, \dots, X_n) , $X = \sum_i X_i$, $\mu = \mathbb{E}(X)$, $\sigma^2 = \mathbb{V}(X)$ y $|X_i| \leq K$ con probabilidad 1,

$$\mathbb{P}(|X - \mu| > t) \leq \max(\exp(-ct^2/\sigma^2), \exp(-ct/K))$$

Haciendo un par de cuentas podemos ver que σ^2 y K se pueden acotar por $\frac{\log(p/\delta)}{d}$. Metiendo esto en la cota de Chernoff, llegamos a que:

$$\mathbb{P}(\|Sy\|_2^2 - 1 \leq \varepsilon) \leq \max(\exp(-\frac{c\varepsilon d}{\log(p/\delta)}), \exp(-\frac{c\varepsilon^2 d}{\log(p/\delta)}))$$

Luego, tomando $d \sim \varepsilon^{-1} \log(p/\delta) \log(1/\delta)$ sirve para obtener la cota de $\delta/2$ \square

1.2 Gordon’s Theorem

Vimos que para preservar la propiedad de JL, nos es suficiente proyectar los datos de un conjunto $S \subset \mathbb{R}^p$ en $\Theta(\varepsilon^{-2} \log(|S|))$ dimensiones ¿Podemos mejorar esto si conocemos la estructura de S ?

Vemos mediante el teorema de Gordon (*Gordon’s Theorem*) que podemos, mediante una transformación con entradas Gaussianas, proyectar a un subespacio de dimensión $\Theta(\log|S|)$ con probabilidad alta.

Primero vamos a establecer un “framework” sobre el que vamos a estar trabajando.

Vamos a trabajar con vectores unitarios en la bola unitaria en \mathbb{R}^p , es decir, sampleando datos en un conjunto $S \subset \mathcal{S}^{p-1}$.

Estos datos van a ser proyectados por transformaciones lineales aleatorias con componentes Gaussianas, es decir, tomando $G : \mathbb{R}^p \rightarrow \mathbb{R}^d$, que son matrices con entradas Gaussianas iid.

Dado esto, definamos la amplitud Gaussiana de S :

Definición 2. *Amplitud Gaussiana* Dado S un cerrado \mathbb{R}^p , definimos la amplitud Gaussiana $\omega(S)$ de S como:

$$\omega(S) = \mathbb{E}(\max_{x \in S} g_p^T x)$$

con $g_p \sim \mathcal{N}(0, I_{p \times p})$

Proposición 2. Dado $g_d \sim \mathcal{N}(0, I_{d \times d})$, definiendo

$$a_d := \mathbb{E}(g_d)$$

entonces $\sqrt{\frac{d}{d+1}}\sqrt{(d)} \leq a_d \leq \sqrt{d}$

Teorema 3. *Gordon's Theorem*

Sea $G \in \mathbb{R}^{d \times p}$ una matriz aleatoria con entradas $\mathcal{N}(0, 1)$ independientes y $S \subset \mathcal{S}^{p-1}$ un cerrado en la bola unitaria en dimensión p . Sea $x \in \mathbb{R}^p$, entonces:

$$\mathbb{E}(\max_{x \in S} \|\frac{1}{a_d} Gx\|) \leq 1 + \frac{\omega(S)}{a_d},$$

y

$$\mathbb{E}(\min_{x \in S} \|\frac{1}{a_d} Gx\|) \geq 1 - \frac{\omega(S)}{a_d}$$

Usando la concentración Gaussiana, viendo que las funciones $F_1(G) = \min_{x \in S} \|Gx\|$ y $F_2(G) = \max_{x \in S} \|Gx\|$ son Lipschitz de constante 1, concluimos que:

$$\forall t > 0 \\ \mathbb{P}(F_1(G) \leq a_d - \omega(S) - t) \leq \exp(-\frac{t^2}{2}) \text{ y } \mathbb{P}(F_2(G) \geq a_d + \omega(S) + t) \leq \exp(-\frac{t^2}{2})$$

Juntando estos dos resultados, llegamos a un resultado que nos permitirá explotar la estructura del espacio donde viven nuestros datos y así, poder trabajar con un algoritmo más eficiente.

Teorema 4. Sea $G \in \mathbb{R}^{d \times p}$ una matriz aleatoria con entradas independientes $\mathcal{N}(0, 1)$ y $S \subset \mathcal{S}^{p-1}$ un conjunto cerrado de la esfera unitaria de dimensión p , entonces, para $\varepsilon > (\frac{\omega(S)}{a_d})$, con probabilidad $\geq 1 - 2 \exp - \frac{a_d^2}{2} (\varepsilon - \frac{\omega(S)}{a_d})^2$:

$$(1 - \varepsilon)\|x\| \leq \|\frac{1}{a_d} Gx\| \leq (1 + \varepsilon) \|x\| \quad \forall x \in S$$

Usando los teoremas 1.2 y 1.3 y tomando $\varepsilon = \frac{\omega(S)+t}{a_d}$, uno puede ver fácilmente que la condición enunciada recién se cumple.

Observar que, dado que el máximo entre n Gaussianas iid. $\mathcal{N}(0, 1)$, podemos ver que la esperanza del máximo entre ellas es $\leq \sqrt{2 \log n}$. Por lo que tomando una dimensión de llegada d del orden de $\log |S|$, podemos garantizar, con un poco de álgebra, que $\frac{1}{a_d} G$ satisface la propiedad de Johnson-Lindenstrauss.

Esto es una mejora considerable en comparación con $d \sim \mathcal{O}(\varepsilon^{-2} \log |S|)$

Para más detalle y demostraciones, ver [4]

1.3 Matrices de bajo rango

Las matrices de bajo rango se presentan en muchos problemas de modelado matemático-computacional. Tienen diversas aplicaciones en machine learning, como en el análisis de componentes principales (PCA) o procesamiento del lenguaje natural (NLP).

La reducción aleatoria de la dimensión tiene aplicaciones en el área de “matrix completion”, donde, dadas algunas entradas de una matriz, queremos reconstruir una matriz de rango menor que mejor la aproxime. Un ejemplo muy famoso es el del *Netflix Prize*, donde se busca crear un sistema de recomendación basado unicamente en puntajes existentes de películas de usuarios de la plataforma (filas: usuarios, columnas: películas). Otro problema es la reconstrucción de señales, donde uno quiere reconstruir una señal discretizada a partir de un sampleo aleatorizado de la señal original.

Veamos un par de ejemplos de esto.

Primero enunciemos el problema de aproximación de una matriz $X_0 \in \mathbb{R}^{m \times n}$ por una matriz de rango $r < \min(m, n)$. Este problema clásico se resuelve con la descomposición *SVD* de X_0 .

Dada $X_0 \in \mathbb{R}^{m \times n}$, buscamos

$$\min_X \|X - X_0\|_F^2$$

$$\text{tq. } \text{rango}(X) = r$$

donde $\|X\|_F^2 = \sum_{i,j} (X[i, j])^2$ (norma de Frobenius de X, al cuadrado)

Luego, la solución está dada por $\hat{X} = U_r \Sigma_r V_r^T = \sum_{i=1}^r \sigma_i u_i v_i^T$, tomando los valores singulares en forma decreciente. (ver Teorema de Eckart Young)

Pero nosotros vamos a abordar el problema de la reconstrucción de una matriz X_0 , por lo que vamos a asumir que, en vez de estar observando X_0 , vamos a observar $y = \mathcal{A}(X_0) + z$ donde z es ruido y $\mathcal{A}: \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^l$ es un “operador de medidas” que toma productos internos de X_0 contra matrices predefinidas A_1, \dots, A_l :

$$y_i = \langle \mathbf{X}_0, \mathbf{A}_i \rangle + z_i = \text{Tr}(A_i^T X_0) + z_i = \sum_{k=1}^m \sum_{j=1}^n X_0[k, j] A_i[k, j] + z_i$$

Vamos a considerar dos modelos:

primero \mathcal{A} consiste de matrices A_i cuyas entradas son iid. variables aleatorias y, segundo, \mathcal{A} devuelve un subconjunto de entradas de la matriz original, donde cada A_i contiene solo una entrada no nula. En el último caso, recuperar información de este sampleo se llama “matrix completion”.

Ahora, el problema se redefine (análogamente) y pasa a ser buscar:

$$\min_X \|y - \mathcal{A}(X)\|_2^2$$

$$\text{tq. } \text{rango}(X) = r$$

Este problema es NP-Hard, pero hay una relajación convexa sin restricciones:

$$\min_X \|y - \mathcal{A}(X)\|_F^2 + \lambda \|X\|_*$$

donde $\|X\|_* = \sum_{i=1}^n \sigma_i$ (suma de valores singulares de X , *Norma nuclear*)

Esto se puede encontrar usando algoritmos proximales.

1.3.1 Modelo Gaussiano

Veamos el caso de recuperación de matrices de bajo rango mediante proyecciones aleatorias Gaussianas:

Tenemos el operador \mathcal{A} , donde cada $A_i (i = 1..l)$ tiene entradas iid. $\sim \mathcal{N}(0, l^{-1})$. Se puede demostrar mediante una cuenta sencilla que esto se toma para que $\mathbb{E}(\|\mathcal{A}(X)\|_2^2) = \|X\|_F^2$ para cualquier matriz fija $X \in \mathbb{R}^{m \times n}$

Para ver que efectivamente podemos diferenciar dos matrices pasadas por el operador \mathcal{A} , se tendría que cumplir la *Matrix Restricted Isometry Property*:

Definición 3. *Matrix Restricted Isometry Property (MRIP)* $\exists 0 \leq \delta < 1$ tq.

$$(1 - \delta)\|X_1 - X_2\|_F^2 \leq \|\mathcal{A}(X_1) - \mathcal{A}(X_2)\|_2^2 \leq (1 + \delta)\|X_1 - X_2\|_F^2$$

con X_1, X_2 de rangos $\leq r$

Esto es una consecuencia del teorema de Johnson-Lindenstrauss. Observar que si \mathcal{A} cumple la MRIP, luego es inyectiva.

Si tenemos \mathcal{A} Gaussiano como especificamos arriba, luego, se puede probar que con alta probabilidad, cuando $l = \Omega(r(n + m))$ (recordar que l es la cantidad de observaciones de X_0), el operador cumplirá la MRIP. La demostración de esto es análoga a la del **teorema 4**.

Para más información y profundización teórica ver [5]

1.3.2 Modelo determinístico

El segundo modelo mencionado, donde tenemos solo una entrada no nula por matriz A_i (para la entrada que estamos observando) del operador \mathcal{A} , se adecua más a lo que serían los sistemas de recomendaciones. Como no podemos asumir una distribución

para el operador, la justificación teórica para este caso no será análoga a la del operador Gaussiano.

Definamos primero la *coherencia* de un subespacio r -dimensional de \mathbb{R}^n .

Definición 4. *Coherencia* Dado U un subespacio r -dimensional de \mathbb{R}^n , la coherencia de U es:

$$\mu(U) := \frac{n}{r} \max_{1 \leq i \leq n} \|\mathcal{P}_U e_i\|_2^2$$

donde e_i es el vector canónico de \mathbb{R}^n y \mathcal{P}_U es la proyección a U

Observar que $1 \leq \mu(U) \leq \frac{n}{r}$ se maximiza cuando un canónico está en la imagen de U y se minimiza cuando U es el subespacio generado por los vectores de magnitud constante $\frac{1}{\sqrt{n}}$.

Las matrices con subespacios de filas y columnas con baja coherencia se dicen “incoherentes” y representan una clase de matrices donde cada entrada contiene una cantidad comparable de información tal que completarlas a partir de un número bajo de observaciones es potencialmente factible.

Se puede ver que para las siguientes hipótesis:

Bajo el problema de minimización con norma nuclear, teniendo de X_0 es relativamente incoherente o satisface condiciones similares; también, siendo $X_0 = U\Sigma V^T$, con $\mu(U), \mu(V) \leq \mu_0, \mu_1, \|E\|_\infty \leq \mu_1 \sqrt{(r/mn)}$, con $E = UV^T$ y $m > n$, tenemos que si observamos l entradas de X_0 con ubicaciones sampleadas uniformemente con $l = \Omega(\max(\mu_0, \mu_1^2)r(m+m)\log^2(m))$, X_0 será solución del problema con alta probabilidad. [5]

Observar que esta cota es superior a la cantidad de observaciones que necesitamos en el caso Gaussiano.

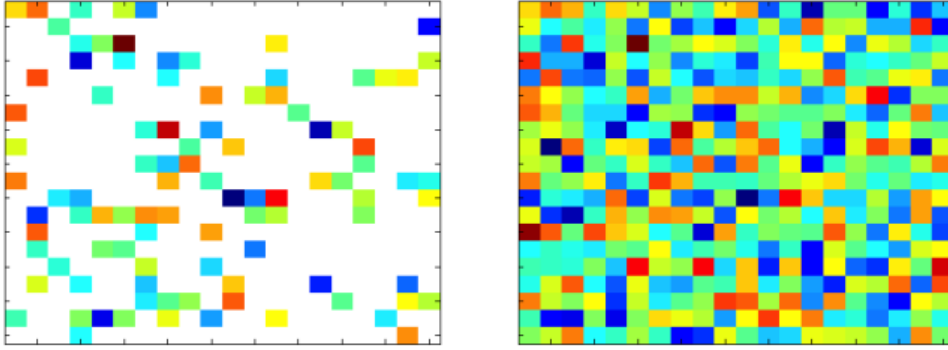


Fig. 1 Ejemplo de *matrix completion*

1.3.3 Ejemplo extra - proyecciones aleatorias en Compressed Sensing

De acuerdo al teorema de Shannon-Nyquist de sampleo de frecuencias, si queremos reconstruir una señal de ancho de banda B a partir de sus samples, necesitamos una frecuencia de sampleo de $2B$, sino caemos en el “Aliasing”.

Consideremos el problema de reconstrucción de señales:

Tenemos una señal temporal discretizada $f \in \mathbb{R}^n$, también asumamos que $\Phi \in \mathbb{R}^{n \times n}$ es la matriz de base donde $\phi_i \in \mathbb{R}^n$. Así, toda señal f puede ser representada como una combinación lineal de las columnas de Φ :

$$\sum_{i=1}^n \phi_i x_i = \Phi x,$$

donde $x \in \mathbb{R}^n$ es un vector disperso que determina lo que aporta cada vector de la base.

Podemos medir a f tomando mediciones aleatorias, análogas al problema anterior:

$$y_j = \phi_j^T f = \phi_j^T \Phi x$$

Podemos deducir que si no hay ruido, entonces al menos necesitamos n vectores ψ_j para reconstruir la señal f .

Si solo podemos usar $m \ll n$ mediciones usando una matriz $\Psi \in \mathbb{R}^{m \times n}$, entonces podemos escribir el problema como

$$y = \Psi f = \Psi \Phi x = Ax$$

donde y es el vector de mediciones y $A = \Psi \Phi$.

Tenemos una cota inferior en la cantidad de mediciones necesarias para poder reconstruir la señal, $m = \Omega(S \log(n/S))$, donde S es la cantidad de entradas no nulas en x [4].

Así, usando la *RIP*, podemos definir una δ_S -isometría tal que

$$(1 - \delta_S)\|x\| \leq \|Ax\| \leq (1 + \delta_S)\|x\|$$

Si una matriz tiene *RIP* de orden $2S$ (ver definición de [4])

$$y_j = \phi_j^T f = \phi_j^T \Phi x$$

podemos tomar dos mediciones $y^{(1)}, y^{(2)} \in \mathbb{R}^m$ y obtener la siguiente igualdad

$$(1 - \delta_S)\|x^{(1)} - x^{(2)}\| \leq \|A(x^{(1)} - x^{(2)})\| \leq (1 + \delta_S)\|x^{(1)} - x^{(2)}\|$$

\Longleftrightarrow

$$(1 - \delta_S) \|x^{(1)} - x^{(2)}\| \leq \|y^{(1)} - y^{(2)}\| \leq (1 + \delta_S) \|x^{(1)} - x^{(2)}\|$$

donde $x^{(1)} - x^{(2)}$ es un vector con a lo sumo $2S$ -disperso. Por lo tanto, si podemos enumerar todos los vectores $2S$ -dispersos y comparar cada uno contra $\|y^{(1)} - y^{(2)}\|$, entonces podemos reconstruir facilmente la seal original.

Ya probamos que una matriz con entradas Gaussianas efectivamente cumple con el *RIP*, y esto se puede usar para este problema. [6]

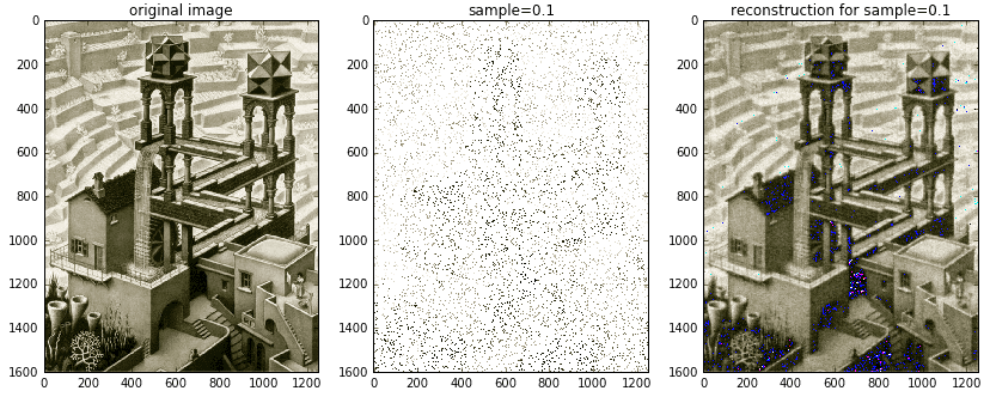


Fig. 2 Ejemplo de reconstruccin de una imagen a partir de un sampleo del 0,1

Referencias

- [1] Sapienza, F.: Distancia de fermat y geodésicas en percolación euclídea: teoría y aplicaciones en machine learning. Tesis de licenciatura, 9 (2018)
- [2] Aggarwal, C.C., Hinneburg, A., Keim, D.A.: On the surprising behavior of distance metrics in high dimensional space. In Database Theory (2001)
- [3] Dasgupta, S., Gupta, A.: An elementary proof of a theorem of johnson and lindenstrauss. ATT Labs; Lucent Bell Labs (2002)
- [4] Bandeira, Singer, Strohmer: Mathematics of Data Science, (2020). Chap. 9, 10
- [5] Davenport, M.A., Romberg, J.: An overview of low-rank matrix recovery from incomplete observations, (2016)
- [6] Nabil, M.: Random Projection and Its Applications, Department of Electrical and Computer Engineering, Tennessee Tech. University, TN, USA (2017)