Notes on Sketching for Affine Subspace Embeddings -I

The notation affine subspace. Let A be an $m \times n$ matrix and we can view it as a linear transformation mapping $\mathbb{R}^n \to \mathbb{R}^n$. In linear algebra, if T is a transformation from a vector space V to W, $T : V \to W$, then T is said to be linear if the following conditions are satisfied:

1. $$T(x + y) = T(x) + T(y), \qquad \text{for all } x, y \in V$$

2. $$T(\alpha x) = \alpha T(x), \qquad \text{for all } \alpha \in \mathbb{F}, x \in V$$

   where, $\mathbb{F}$ is the underlying field of the vector spaces V and W.

From the definition, it follows that $T(0) = T(0 + 0) = T(0) + T(0)$ and hence $T(0) = 0$. In our discussion, we restrict V and W to be $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively.

Given a matrix $A \in \mathbb{R}^{m \times n}$, the transformation $L : \mathbb{R}^n \to \mathbb{R}^m$ defined as $L(x) = Ax$, is a linear transformation. Define a transformation

$$T(x) = Ax + b$$

where, $b \in \mathbb{R}^m$ and not necessarily 0. If $b \neq 0$, then, $T(0) = b$ and does not satisfy the linearity axiom. However, for any x, $T(x)$ equals Ax shifted by b, that is, it is a *linear transformation that is shifted* by a constant vector b. Such transformations are often called *affine transformations* or *affine maps*.

Example 1. An example of affine transformations is the solution to the system of equations

$$Ax = b$$

where, $b \neq 0$. If b lies in the column space of A, then, the space of solutions to this system is given as

$$x_p + N = \{x_p + x_n \mid x_n \in N\}$$

where, N is the null space of A, or, equivalently, the space of solutions to the homogeneous equations $Ax = 0$; and x_p is *any* particular solution $Ax_p = b$. If x and x' are any two particular solutions satisfying $Ax = b$ and $Ax' = b$, then, necessarily, $A(x - x') = 0$, or, that $x - x' \in N$. Since N is a vector space, $x + N = x' + N$ (since, for any $n \in N$, $x + n = x' + ((x - x') + n) = x' + n' \in N$, as $(x - x') \in N$ and the sum $(x - x') + n \in N$ from vector subspace closure properties under addition).

Let A be an $m \times n$ matrix over reals and let $p \geq 1$ be an integer. The mapping

$$T(X) = AX$$

where, $X \in \mathbb{R}^{m\times p}$ can be viewed as a linear transformation $T : \mathbb{R}^{n\times p} \to \mathbb{R}^{m\times p}$. From the above discussion, it then follows that

$$T(X) = AX + B$$

is an affine transformation, where, B is a fixed $m \times p$ matrix over reals.

Sketching for Approximate Preservation of Affine Transformations

Approximate subspace norm preserving embeddings. We have previously seen how randomized sketching can be used to approximately preserve the norms of all vectors in the range of a linear transformation from $\mathbb{R}^d \to \mathbb{R}^n$. Equivalently, this can be viewed as the approximate preservation of norms of all vectors in a d-dimensional subspace V of $\mathbb{R}^n$. Since a basis $v_1, v_2, \ldots, v_d$ for a given d-dimensional subspace V of $\mathbb{R}^n$ can be placed as columns of a matrix $A = \begin{bmatrix} v_1 & v_2 & \ldots & v_d \end{bmatrix}$, the subspace V is the column space of A. The basis can be alternately replaced by an orthonormal basis, while preserving the equivalence of the column space. So without loss of generality, we can assume that A has orthonormal columns. With this view, a distribution $\mathcal{D}$ over matrices $\mathbb{R}^{k\times n}$ is said to approximately preserve the norms under the embedding of the subspace V defined by a randomly chosen matrix S from the distribution $\mathcal{D}$ if

$$\mathrm{P}_{S\sim\mathcal{D}}\Big[\text{ for all } x \in \mathbb{R}^d, \|SAx\|_2 \in (1\pm\epsilon)\|Ax\|_2 \Big] \geq 1-\delta$$

where, $0 < \epsilon < 1$ and $0 < \delta < 1$ are the parameters of this approximate subspace norm preserving embedding.

Approximate norm preserving embeddings of affine transformations. Suppose we define a transformation $T(x) = Ax + b$, where, $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$ and we wish to approximately preserve the norm $\|T(x)\|$ under the transformation S, that is, $\|S(T(x))\|_2 \in (1\pm\epsilon)\|Tx\|_2$, for all $x \in \mathbb{R}^d$. To achieve this, it suffices to choose S randomly from a distribution that approximately preserves norms of the column space defined by $\begin{bmatrix} A & b \end{bmatrix}$. That is, suppose the distribution $\mathcal{D}$ satisfies the property that

$$\mathrm{P}_{S\sim\mathcal{D}}\Big[\text{ for all } x \in \mathbb{R}^d, \|SAx + Sb\|_2 \in (1\pm\epsilon)\|Ax+b\|_2 \Big] \geq 1-\delta\ .$$

Written equivalently, this is,

$$\mathrm{P}_{S\sim\mathcal{D}}\left[\text{ for all } x \in \mathbb{R}^d, \left\| S \begin{bmatrix} A & b \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix} \right\| \in (1\pm\epsilon) \left\| \begin{bmatrix} A & b \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix} \right\| \right] \geq 1-\delta\ .$$

Since S preserves the norms of the column space of $\begin{bmatrix} A & b \end{bmatrix}$, the above statement follows.

Norm preservation in more general affine subspace embeddings. Suppose we define an affine embedding as

$$T(X) = AX - B$$

where, $A \in \mathbb{R}^{n\times d}$ and $B \in \mathbb{R}^{n\times p}$ and $T : \mathbb{R}^{d\times p} \to \mathbb{R}^{n\times p}$. We wish to consider random matrices S drawn from a distribution $\mathcal{D}$ so that

$$\text{for any } X,\ \|SAX - SB\|_F \in (1\pm\epsilon)\|AX-B\|_2$$

with probability $1-\delta$ under the distribution $\mathcal{D}$.

The problem diverges from subspace embedding if p is significantly larger than d. The transformation T is an affine mapping, that is, to AX, a fixed matrix B is added. Hence, it is not necessary to preserve the entire column space $\begin{bmatrix} A & B \end{bmatrix}$, since preserving $\|AX - BY\|_F$ for all X, Y is not required. In particular, $Y = I$, making it an affine transformation. We present an analysis below.

Given $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^d$, let $A = U_r \Sigma_r V_r^T$ denote the thin SVD of A, where, r is assumed to be the rank of A. Recall from the definition of pseudo-inverse that

$$A^- = V_r \Sigma_r^{-1} U_r^T \ .$$

The projection matrix on the column space of A is

$$U_r U_r^T = AA^-$$

. The projection matrix on the orthogonal complement of the column space of A is $I - U_r U_r^T = I - AA^-$. Any vector $b \in \mathbb{R}^m$ can be written as the sum of orthogonal components,

$$b = U_r U_r^T b + (I - U_r U_r^T) b = AA^- b + (I - AA^-) b \ .$$

For any $x \in \mathbb{R}^d$,

$$\|Ax - b\|_2^2 = \|(Ax - AA^- b) + (-I + AA^-) b\|_2^2$$

The vector $Ax - AA^- b$ is in the column space of A and $(I - AA^-)(-b)$ is in the orthogonal complement space of the column space of A. Therefore,

$$\|Ax - b\|_2^2 = \|Ax - AA^- b\| + \|(I - AA^-) b\|_2^2 \ .$$

The solution to the linear regression problem

$$\text{Min}_{x \in \mathbb{R}^d} \|Ax - b\|_2$$

is obtained equivalently from the solution to the problem

$$\text{Min}_{x \in \mathbb{R}^d} \|Ax - b\|_2^2$$

which from above is equivalent to

$$\|(I - AA^-) b\|_2^2 + \text{Min}_{x \in \mathbb{R}^d} \|Ax - AA^- b\|_2^2$$

The expression $\|Ax - AA^- b\|_2$ is minimized over x by setting $x = A^- b$ and obtaining $\|Ax - AA^- b\| = 0$. The class of all solutions for $Ax = AA^- b$ is $x = A^- b + N = \{A^- b + n \mid n \in N\}$, where, N is the nullspace of A. Since $A^- = V_r \Sigma^{-1} U_r^T$, $A^- b$ is in the row space of A and therefore orthogonal to all $n \in N$. Therefore, for any general solution to $Ax = AA^- b$, $x = A^- b + n$,

$$\|x\|_2^2 = \|A^- b + n\|_2^2 = \|A^- b\|_2^2 + \|n\|_2^2 \ .$$

The special solution $x^* = A^- b$ has the smallest $\|\cdot\|_2$ among all x that have the same value of $\min_{x \in \mathbb{R}^d} \|Ax - b\|_2 = \|(I - AA^-) b\|_2$. That is,

$$\operatorname*{argmin}_{\substack{x \in \mathbb{R}^d \\ \|x\|_2 \text{ is minimum}}} \|Ax - b\|_2 = A^- b = x^* \ .$$

Notation. Consider the generalized linear regression problem of

$$\min_{X \in \mathbb{R}^{d \times p}} \|AX - B\|_F \ .$$

Equivalently, we can study the problem

$$\min_{X \in \mathbb{R}^{d \times p}} \|AX - B\|_F^2$$

which has the same optimal solutions. We have,

$$\|AX - B\|_F^2 = \sum_{j=1}^{p} \|AX_j - B_j\|_2^2 \ .$$

The minimization problem of $\|AX - B\|_F^2$ is therefore the sum of p independent optimization problems, one corresponding to each column index $j = 1, 2, \ldots, p$, that is.

$$\min_X \|AX - B\|_F^2 = \min_X \sum_{j=1}^{p} \|AX_j - B_j\|_2^2 = \sum_{j=1}^{p} \min_{X_j \in \mathbb{R}^d} \|AX_j - B_j\|^2$$

The optimal solution to the jth minimization problem is denoted by $X_j^* = A^- B_j$. These columns are placed in order to give the optimal solution X^* to the affine regression problem $\min_{X \in \mathbb{R}^{d \times p}} \|AX - B\|_F^2$ as

$$X^* = \begin{bmatrix} X_1^* & X_2^* & \cdots & X_p^* \end{bmatrix} \ .$$

Thus,

$$X^* = \begin{bmatrix} A^- B_1 & A^- B_2 & \cdots & A^- B_p \end{bmatrix} = A^- B \ .$$

For each column index j, $AX_j^* - B_j = -(-AA^- + I_m)B_j$ which is the component of B_j in the orthogonal complement of the column space of A in $\mathbb{R}^n$. Thus, $AX^* - B = -(I_m - AA^-)B$, each column of $AX^* - B$ lies in the orthogonal complement of the column space of A in $\mathbb{R}^n$. Thus, the column space of $AX^* - B$ is a subspace of the orthogonal complement of the column space of A in $\mathbb{R}^n$. It follows therefore that for any $X \in \mathbb{R}^{d \times n}$,

$$\begin{aligned}
\|AX - B\|_F^2 &= \|AX - AX^* + AX^* - B\|_F^2 \\
&= \sum_{j=1}^{p} \left[\|A(X_j - X_j^*) + (Ax_j^* - B_j)\|_2^2 \right] \\
&= \sum_{j=1}^{p} \left[\|A(X_j - X_j^*)\|_2^2 + \|Ax_j^* - B_j\|_2^2 + 2(A(X_j - X_j^*))^T (AX^* j - B_j) \right] \\
&= \sum_{j=1}^{p} \left[\|A(X_j - X_j^*)\|_2^2 + \|Ax_j^* - B_j\|_2^2 \right] \ .
\end{aligned}$$

since, $A(X_j - X_j^*)$ lies in the column space of A and $AX_j^* - B_j$ lies in the orthogonal complement of the column space of A, their inner product is 0. The last step therefore simplifies to

$$\begin{aligned}
\|AX - B\|_F^2 &= \sum_{j=1}^{p} \|A(X_j - X_j^*)\|_2^2 + \sum_{j=1}^{p} \|Ax_j^* - B_j\|_2^2 \\
&= \|A(X - X^*)\|_F^2 + \|AX^* - B\|_F^2 \ .
\end{aligned}$$

Frobenius norm squared of sum of matrices. There is a simple expression for $\|C + D\|_F^2$, given two $m \times n$ real valued matrices C and D. This is used in the equations above for the special case when the column spaces of C and D are orthogonal.

$$\begin{aligned}
\|C + D\|_F^2 &= \sum_{j=1}^{\|} (C + D)_j\|_2^2 = \sum_{j=1}^{n} \|C_j + D_j\|_2^2 \\
&= \sum_{j=1}^{n} \left[\|C_j\|^2 + \|D_j\|^2 + 2C_j^T D_j\right] \\
&= \sum_{j=1}^{n} \|C_j\|^2 + \sum_{j=1}^{n} \|D_j\|^2 + 2\sum_{j=1}^{n} C_j^T D_j \\
&= \|C\|_F^2 + \|D\|_F^2 + 2\mathrm{tr}\ C^T D \ \ .
\end{aligned}$$

An inequality via trace of matrix product. We review a simple inequality of tr $C^T D$, where, C and D are any two given $m \times n$ matrices. We have,

$$\begin{aligned}
\mathrm{tr}\ C^T D &= \sum_{j=1}^{n} C_j^T D_j \\
&\leq \sum_{j=1}^{n} \|C_j\|_2 \|D_j\|_2 && \text{by Cauchy-Schwarz inequality} \\
&\leq \left(\sum_{j=1}^{n} \|C_j\|_2^2\right)^{1/2} \left(\sum_{j=1}^{n} \|D_j\|_2^2\right)^{1/2} && \text{Cauchy-Schwarz inequality second time} \\
&= \|C\|_F \|D\|_F \ \ .
\end{aligned}$$

The second application of the Cauchy-Schwarz inequality is as follows. Consider two vectors $a = \begin{bmatrix}\|C_1\| & \|C_2\| & \cdots & \|C_n\|\end{bmatrix}^T$ and $b = \begin{bmatrix}\|D_1\| & \|D_2\| & \cdots & \|D_n\|\end{bmatrix}^T$. By Cauchy-Schwarz inequality, the inner product of a and b is bounded above by the product of their norms. This would give,

$$a^T b \leq \left|a^T b\right| \leq \|a\|\|b\| = \|C\|_F \|D\|_F \ \ .$$

Norm Preserving Affine Transformation

We would like to find conditions so that

$$\|SAX - SB\|_F \in (1 \pm \epsilon)\|AX - B\|_F$$

for all $X \in \mathbb{R}^{d \times p}$, and with probability $1 - \delta$.

Without loss of generality, we assume that A has orthonormal columns.

(*1) To begin with, we will assume that S preserves the norms in the column space of A to within factors of $1 \pm \epsilon$. That is,

$$\|SAx\|_2 \in (1 \pm \epsilon)\|x\|_2, \qquad \text{for all } x \in \mathbb{R}^d \ \ .$$

Further conditions are derived as we proceed with the analysis.

Denote the optimal solution to $\text{Min}_{X\in\mathbb{R}^{d\times p}}\|AX-B\|_F$ by X^*. For ease of notation, let B^* denote $B^* = AX^* - B = -(I - AA^-)B$. By orthogonality of the column spaces of A and B^*, we have for any $X \in \mathbb{R}^{d\times n}$,

$$\|AX - B\|_F^2 = \|AX - AX^*\|_F^2 + \|B^*\|_F^2 \ .$$

We now consider $\|SAX - SB\|_F^2$ to see the conditions under which it is within $(1 \pm \epsilon)$ factors of $\|AX - B\|_F^2$.

$$\begin{aligned}\|SAX - SB\|_F^2 &= \|SA(X - X^*) + S(AX^* - B)\|_F^2 = \|SA(X - X^*) + SB^*\|_F^2 \\ &= \|SA(X - X^*)\|_F^2 + \|SB^*\|_F^2 + 2\text{tr } (SA(X - X^*))^T SB^* \\ &= (1 \pm \epsilon)\|A(X - X^*)\|_F^2 + (1 \pm \epsilon)\|B^*\|_F^2 + 2\text{tr } (SA(X - X^*))^T SB^* \end{aligned} \tag{1}$$

where, the last step makes the following assumption.
(*2) The random matrix S chosen from the distribution $\mathcal{D}$ approximately preserves the Frobenius norm of the matrix B^*

$$\|SB^*\|_F^2 \in (1 \pm \epsilon)\|B^*\|_F^2 \qquad\qquad \text{Assumption 2} \ .$$

We now consider the cross term $2\text{tr } (SA(X - X^*))^T SB^*$. This can be written as

$$2\text{tr } (SA(X - X^*))^T SB^* = 2\text{tr } (X - X^*)^T (SA)^T SB^* \leq 2\|X - X^*\|_F \|(SA)^T SB^*\|_F \ .$$

Suppose we make the following third assumption about the random matrix S.
(*3) S approximately preserves the inner product of A_i with B_j^*, for each $i, j = 1, 2, \ldots, p$, namely,

$$\big|\|(SA)^T SB^*\|_F - \|A^T B^*\|_F\big| \leq \frac{\epsilon}{\sqrt{d}}\|A\|_F\|B^*\|_F \qquad\qquad \text{Assumption 3.}$$

Since the columns of B^* are orthogonal to the columns of A, $A^T B = 0_{n\times n}$ and therefore $\|A^T B\|_F = 0$. Further, since A has orthonormal columns, $\|A\|_F^2 = d$. Therefore, Assumption 3 implies that

$$2\|(SA)^T SB^*\|_F \leq 2\epsilon\|B^*\|_F \ . \tag{2}$$

So the expression for $2\text{tr } (SA(X - X^*))^T SB^*$ is given by

$$\begin{aligned} 2|\text{tr } (SA(X - X^*))^T SB^*| &= 2|\text{tr } (X - X^*)^T (SA)^T SB^*| \\ &\leq \|X - X^*\|_F \|(SA)^T SB^*\|_F \\ &\leq 2\epsilon\|X - X^*\|_F\|B\|_F \ . \end{aligned} \tag{3}$$

Substituting in Eqn (1) and rearranging the terms, we have,

$$\begin{aligned} &\big|\|SAX - SB\|_F^2 - \|AX^* - B\|_F^2 - \|B^*\|_F^2\big| \\ &\leq \epsilon\|A(X - X^*)\|_F^2 + \epsilon\|B^*\|_F^2 + 2\epsilon\|X - X^*\|_F\|B^*\|_F \\ &\leq \epsilon\left[\|A(X - X^*)\|_F^2 + \|B^*\|_F^2 + (\|X - X^*\|_F^2 + \|B^*\|_F^2)\right] \qquad \text{by AM} \geq \text{GM, } 2\alpha\beta \leq \alpha^2 + \beta^2. \\ &= 2\epsilon\left[\|A(X - X^*)\|_F^2 + \|B^*\|_F^2\right] \ . \end{aligned}$$

The last step uses the fact that since A has orthonormal columns, $\|A(X - X^*)\|_F = \|X - X^*\|_F$. Noting that $\|AX - B\|_F^2 = \|A(X - X^*)\|_F^2 + \|B^*\|_F^2$, we have,

$$\big|\|SAX - SB\|_F^2 - \|AX - B\|_F^2\big| \leq 2\epsilon\|AX - B\|_F^2 \ . \tag{4}$$

In other words, under conditions of Assumptions 1,2 and 3, we have,

$$\|SAX - SB\|_F^2 \in (1 \pm 2\epsilon)\|AX - B\|_F^2$$

which would satisfy the notion of approximation preservation of norms under affine space embedding.

Conditions for satisfying assumptions 1,2 and 3

These are posed as exercises. The conditions for the number of rows m for the random matrix S depends on the distribution $\mathcal{D}$ from which it is drawn.