Using the Johnson-Lindenstrauss lemma in linear and integer programming

Vu Khac Ky[1], Pierre-Louis Poirion, Leo Liberti

LIX, École Polytechnique, F-91128 Palaiseau, France
Email:{vu,poirion,liberti}@lix.polytechnique.fr

July 2, 2015

Abstract

The Johnson-Lindenstrauss lemma allows dimension reduction on real vectors with low distortion on their pairwise Euclidean distances. This result is often used in algorithms such as k-means or k nearest neighbours since they only use Euclidean distances, and has sometimes been used in optimization algorithms involving the minimization of Euclidean distances. In this paper we introduce a first attempt at using this lemma in the context of feasibility problems in linear and integer programming, which cannot be expressed only in function of Euclidean distances.

1 Introduction

In machine learning theory there is a wonderful and deep result from functional analysis and geometric convexity called the *Johnson-Lindenstrauss Lemma* (JLL) [3]. Intuitively, this lemma employs a concentration of measure [4] argument to prove that a "cloud" of high-dimensional points can be projected in a much lower dimensional space whilst keeping Euclidean distances approximately the same. Although this result was previously exploited in purely Euclidean distance based algorithms such as k-means [1] and k nearest neighbours [2] (among others), it has not often been applied to optimization problems. There are a few exceptions, namely high dimensional linear regression [6], where the application of the JLL is reasonably natural.

In this paper we present some new results on the application of the JLL to establish the feasibility of Linear Programs (LP) and Integer Linear Programs (ILP). We consider problems with m constraints (where m is large), and reduce m by projection on a random subspace, while ensuring that, with high probability, the reformulated problem is feasible if and only if the original problem is feasible.

The geometrical intuition underlining our idea stems from the cone interpretation of LP feasibility. Let P be a feasibility-only LP in standard form (i.e. all inequalities have been turned into equations by the introduction of m additional non-negative variables), written as $Ax = b$ with $x \geq 0$, where A is an $m \times n$ rational matrix, $b \in \mathbb{R}^m$ is a rational vector, and x is a vector of n decision variables (which might be continuous or integer). Then P can be interpreted as the following geometric decision problem: given a cone spanned by the columns of A, is $b \in \mathsf{cone}(A)$ or not? In this setting, the role of the JLL is seen to be the following: if we project A and b in a smaller dimensional space, the JLL assures us that the "shape" of the projected cone and of the ray b are approximately the same, and hence that the answer to the problem will be the same most of the times.

In Section 2, we formally define the problem. In Section 3, we recall the Johnson-Lindenstrauss Lemma and prove some results linked to its application to the ILP case. In Section 4, we derive results concerning LP feasibility when the cone generated by the matrix A is pointed. In Section 5, we generalize the previous results, proving that the distance between a point and a closed set should be approximately preserved. Finally in Section 6, we present some computational results.

[1]Supported by a Microsft Research Ph.D. fellowship.

2 Linear programming and the cone membership problem

It is well-known that any linear program can be reduced (via an easy bisection argument) to LP feasibility, defined as follows:

> LINEAR FEASIBILITY PROBLEM (LFP). Given $b \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m\times n}$. Decide whether there exists $x \in \mathbb{R}^n$ such that $Ax = b \wedge x \geq 0$.

We assume that m and n are very large integer numbers. We also assume that A is full row-rank. In particular, we have $m \leq n$, since otherwise we can find x uniquely from $Ax = b$ by taking the left inverse of A.

LFP problems can obviously be solved using the simplex method. Despite the fact that simplex methods are often very efficient in practice, there are instances for which the methods run in exponential time. On the other hand, polynomial time algorithms such as interior point methods are known to scale poorly, in practice, on several classes of instances.

If $a_1, \ldots, a_n$ are the column vectors of A, then the LFP is equivalent to finding $x \geq 0$ such that b is a non-negative linear combination of $a_1, \ldots, a_n$. In other words, the LFP is equivalent to the following cone membership problem:

> CONE MEMBERSHIP (CM). Given $b, a_1, \ldots, a_n \in \mathbb{R}^m$, decide whether $b \in \mathsf{cone}\{a_1, \ldots, a_n\}$.

This problem can be viewed as a special case of the *restricted linear membership problem*, which is defined as follows:

> RESTRICTED LINEAR MEMBERSHIP (RLM). Given $b, a_1, \ldots, a_n \in \mathbb{R}^m$ and $X \subseteq \mathbb{R}^n$, decide whether $b \in \mathsf{lin}_X(a_1, \ldots, a_n)$, i.e. whether $\exists \lambda \in X$ s.t. $b = \sum\limits_{i=1}^{n} \lambda_i a_i$.

For example, when $X = \mathbb{R}^n_+$, we have the cone membership problem; and when $X = \mathbb{Z}^n$ (or $\{0,1\}^n$) we have the integer (binary) cone membership problem.

It is known from the JLL (see below for an exact statement) that there is a (linear) mapping $T : \mathbb{R}^m \to \mathbb{R}^k$, where $k \ll m$, such that the pairwise distances between all vector pairs (a_i, a_j) undergo low distortion. In other words, the complete graph on $\{a_1, \ldots, a_n\}$ weighted with the pairwise Euclidean distances realized in $\mathbb{R}^m$ can also be approximately realized in $\mathbb{R}^k$. We are now stipulating that such a graph is a reasonable representation of the intuitive notion of "shape". Under this hypothesis, it is reasonable to expect that the image of $C = \mathsf{cone}(a_1, \ldots, a_n)$ under T has approximately the same shape as C. Thus, given an instance of CM, we expect to be able to "approximately solve" a much smaller (randomly projected) instance instead. Notice that since CM is a decision problem, "approximately" really refers to a randomized algorithm which is successful with high probability.

Notationwise, every norm $\|\cdot\|$ is Euclidean unless otherwise specified, and we shall denote by E^{c} the complement of an event E.

3 Random projections and RLM problems

The JLL is stated as follows:

3.1 Theorem (Johnson-Lindenstrauss Lemma [3])
Given $\varepsilon \in (0,1)$ and $A = \{a_1, \ldots, a_n\}$ be a set of n points in $\mathbb{R}^m$. Then there exists a mapping $T : \mathbb{R}^m \to \mathbb{R}^k$ such that

$$(1-\varepsilon)\|a_i - a_j\| \leq \|T(a_i) - T(a_j)\| \leq (1+\varepsilon)\|a_i - a_j\| \tag{1}$$

for all $1 \leq i, j \leq n$, in which k is $O(\varepsilon^{-2} \log n)$.

Thus, all sets of n points can be projected to a subspace having dimension logarithmic in n (and, surprisingly, independent of m), such that no distance is distorted by more than $1 + 2\varepsilon$. The JLL is a consequence of a general property (see Lemma 3.2 below) of *sub-Gaussian* random mappings $T(x) = \sqrt{\frac{1}{k}} P\, x$ where P is an appropriately chosen matrix. Some of the most popular are:

- orthogonal projections on a random k-dimensional linear subspace of $\mathbb{R}^m$;
- random $k \times m$ matrices with each entry independently drawn from the standard normal distribution $\mathcal{N}(0,1)$;
- random $k \times m$ matrices with each entry independently taking values $+1$ and -1, each with probability $\frac{1}{2}$;
- random $k \times m$ matrices with entries independently taking values $+1$, 0, -1, respectively with probability $\frac{1}{6}$, $\frac{2}{3}$, $\frac{1}{6}$.

3.2 Lemma (Random projection lemma)
Let $T : \mathbb{R}^m \to \mathbb{R}^k$ be one of the above random mappings. Then for all $\varepsilon \in (0,1)$ and all vector $x \in \mathbb{R}^m$, we have:

$$\mathsf{Prob}(\,(1-\varepsilon)\|x\| \leq \|T(x)\| \leq (1+\varepsilon)\|x\|\,) \geq 1 - 2e^{-\mathcal{C}\varepsilon^2 k} \tag{2}$$

for some constant $\mathcal{C} > 0$ (independent of m, k, ε).

Not only can this lemma prove the existence of a mapping satisfying conditions in the Johnson-Lindenstrauss lemma, but also it implies that the probability of finding such a mapping is very large. Indeed, from the random projection lemma, the probability that Eq. (1) holds for all $i \neq j \leq m$ is at least

$$1 - 2\binom{m}{2} e^{-\mathcal{C}\varepsilon^2 k} = 1 - m(m-1)e^{-\mathcal{C}\varepsilon^2 k}. \tag{3}$$

Therefore, if we want this probability to be large than, say 99.9%, then simply choose any k such that $\frac{1}{100m(m-1)} > e^{-\mathcal{C}\varepsilon^2 k}$. This means k can be chosen to be $k = \lceil \frac{\ln(1000) + 2\ln(m)}{\mathcal{C}\varepsilon^2} \rceil$, which is $O(\,\varepsilon^{-2}(\ln(m) + 3.5)\,)$.

We shall also need a squared version of the random projection lemma.

3.3 Lemma (Random projection lemma, squared version)
Let $T : \mathbb{R}^m \to \mathbb{R}^k$ be one of the random mappings in Lemma 3.2. Then for all $\varepsilon \in (0,1)$ and all vector $x \in \mathbb{R}^m$, we have:

$$\mathsf{Prob}(\,(1-\varepsilon)\|x\|^2 \leq \|T(x)\|^2 \leq (1+\varepsilon)\|x\|^2\,) \geq 1 - 2e^{-\mathcal{C}(\varepsilon^2 - \varepsilon^3)k} \tag{4}$$

for some constant $\mathcal{C} > 0$ (independent of m, k, ε).

Another direct consequence of the random projection lemma is the concentration around zero of the involved random linear projection kernel.

3.4 Corollary
Let $T : \mathbb{R}^m \to \mathbb{R}^k$ be one of the random mappings as in Lemma 3.2 and $0 \neq x \in \mathbb{R}^m$. Then we have

$$\mathsf{Prob}(T(x) \neq 0) \geq 1 - 2e^{-\mathcal{C}k}. \tag{5}$$

for some constant $\mathcal{C} > 0$ (independent of n, k).

Proof. For any $\varepsilon \in (0, 1)$, we define the following events:

$$\begin{aligned} \mathcal{A} &= \big\{T(x) \neq 0\big\} \\ \mathcal{B} &= \big\{(1-\varepsilon)\|x\| \leq \|T(x)\| \leq (1+\varepsilon)\|x\|\big\}. \end{aligned}$$

By Lemma 3.2 it follows that $\mathsf{Prob}(\mathcal{B}) \geq 1 - 2e^{-c\varepsilon^2 k}$ for some constant $\mathcal{C} > 0$ independent of m, k, ε. On the other hand, $\mathcal{A}^{\mathsf{c}} \cap \mathcal{B} = \emptyset$, since otherwise, there is a mapping T_1 such that $T_1(x) = 0$ and $(1-\varepsilon)\|x\| \leq \|T_1(x)\|$, which altogether imply that $x = 0$ (a contradiction). Therefore, $\mathcal{B} \subseteq \mathcal{A}$, and we have $\mathsf{Prob}(\mathcal{A}) \geq \mathsf{Prob}(\mathcal{B}) \geq 1 - 2e^{-\mathcal{C}\varepsilon^2 k}$. This holds for all $0 < \varepsilon < 1$, so $\mathsf{Prob}(\mathcal{A}) \geq 1 - 2e^{\mathcal{C}k}$. □ □

3.5 Lemma
Let $T : \mathbb{R}^m \to \mathbb{R}^k$ be one of the random mappings as in Lemma 3.2 and $b, a_1, \ldots, a_n \in \mathbb{R}^m$. Then for any given vector $x \in \mathbb{R}^n$, we have:

(i) If $b = \sum\limits_{i=1}^{n} x_i a_i$ then $T(b) = \sum\limits_{i=1}^{n} x_i T(a_i)$;

(ii) If $b \neq \sum_{i=1}^n x_i a_i$ then $\mathsf{Prob}\left[T(b) \neq \sum_{i=1}^n x_i T(a_i)\right] \geq 1 - 2e^{-\mathcal{C}k}$;

(iii) If $b \neq \sum_{i=1}^n y_i a_i$ for all $y \in X \subseteq \mathbb{R}^n$, where $|X|$ is finite, then

$$\mathsf{Prob}\left[\forall y \in X \;\; T(b) \neq \sum_{i=1}^{n} y_i T(a_i)\right] \geq 1 - 2|X|e^{-\mathcal{C}k};$$

for some constant $\mathcal{C} > 0$ (independent of n, k).

Proof. Point (i) follows by linearity of T, and (ii) by applying Cor. 3.4 to $Ax - b$. For (iii), we have

$$\begin{aligned} \mathsf{Prob}\left[\forall y \in X \;\; T(b) \neq \sum_{i=1}^{n} y_i T(a_i)\right] &= \mathsf{Prob}\left[\bigcap_{y \in X} \big\{T(b) \neq \sum_{i=1}^{n} y_i T(a_i)\big\}\right] \\ = 1 - \mathsf{Prob}\left[\bigcup_{y \in X} \big\{T(b) \neq \sum_{i=1}^{n} y_i T(a_i)\big\}^{\mathsf{c}}\right] &\geq 1 - \sum_{y \in X} \mathsf{Prob}\left[\big\{T(b) \neq \sum_{i=1}^{n} y_i T(a_i)\big\}^{\mathsf{c}}\right] \\ \text{[by (ii)]} \qquad &\geq 1 - \sum_{y \in X} 2e^{-\mathcal{C}k} = 1 - 2|X|e^{-\mathcal{C}k}, \end{aligned}$$

as claimed. □ □

This lemma can be used to solve the RLM problem when the cardinality of the restricted set X is bounded by a polynomial in n. In particular, if $|X| < n^d$, where d is small w.r.t. n, then

$$\mathsf{Prob}\big[T(b) \notin \mathsf{Lin}_X \{T(a_1), \ldots, T(a_n)\}\big] \geq 1 - 2n^d e^{-\mathcal{C}k}. \tag{6}$$

Then by taking any k such that $k \geq \frac{1}{\mathcal{C}} \ln(\frac{2}{\delta}) + \frac{d}{\mathcal{C}} \ln n$, we obtain a probability of success of at least $1 - \delta$. We give an example to illustrate that such a bound for $|X|$ is natural in many different settings.

3.6 Example
If $X = \{x \in \{0,1\}^n | \sum_{i=1}^n \alpha_i x_i \leq d\}$ *for some* d*, where* $0 < \alpha_i$ *for all* $1 \leq i \leq n$*, then* $|X| < n^{\overline{d}}$*, where* $\overline{d} = \max_{1 \leq i \leq n} \lfloor \frac{d}{\alpha_i} \rfloor$*. To see this, let* $\underline{\alpha} = \min_{1 \leq i \leq n} \alpha_i$*; then* $\sum_{i=1}^{n} x_i \leq \sum_{i=1}^{n} \frac{\alpha_i}{\underline{\alpha}} x_i \leq \frac{d}{\underline{\alpha}}$*, which implies* $\sum_{i=1}^{n} x_i \leq \overline{d}$*. Therefore* $|X| \leq \binom{n}{0} + \binom{n}{1} + \ldots \binom{n}{\overline{d}} < n^{\overline{d}}$*, as claimed.* □

Lemma 3.5 also gives us an indication as to why estimating the probability that

$$T(b) \notin \mathsf{cone}\{T(a_1), \ldots, T(a_n)\}$$

is not straightforward. This event can be written as an intersection of infinitely many sub-events

$$\{T(b) \neq \sum_{i=1}^{n} y_i T(a_i)\}$$

where $y \in \mathbb{R}_+^n$; even if each of these occurs with high probability, their intersection might still be small. As these events are dependent, however, we still hope to find to find a useful estimation for this probability.

4 Projections of separating hyperplanes

In this section we show that if a hyperplane separates a point x from a closed and convex set C, then its image under a random projection T is also likely to separate $T(x)$ from $T(C)$. The separating hyperplane theorem applied to cones can be stated as follows.

4.1 Theorem (Separating hyperplane theorem)
Given $b \notin \mathsf{cone}\{a_1, \ldots, a_n\}$ *where* $b, a_1, \ldots, a_n \in \mathbb{R}^m$*. Then there is* $c \in \mathbb{R}^m$ *such that* $c^T b < 0$ *and* $c^T a_i \geq 0$ *for all* $i = 1, \ldots, n$*.*

For simplicity, we will first work with *pointed cone*. Recall that a cone C is called pointed if and only if $C \cap -C = \{0\}$. The associated separating hyperplane theorem is obtained by replacing all $\geq$ inequalities by strict ones. Without loss of generality, we can assume that $\|c\| = 1$. From this theorem, it immediately follows that there is a positive ε_0 such that $c^T b < -\varepsilon_0$ and $c^T a_i > \varepsilon_0$ for all $1 \leq i \leq n$.

4.2 Proposition
Given $b, a_1, \ldots, a_n \in \mathbb{R}^m$ *of norms* 1 *such that* $b \notin \mathsf{cone}\{a_1, \ldots, a_n\}$*,* $\varepsilon > 0$*, and* $c \in \mathbb{R}^m$ *with* $\|c\| = 1$ *be such that* $c^T b < -\varepsilon$ *and* $c^T a_i \geq \varepsilon$ *for all* $1 \leq i \leq n$*. Let* $T : \mathbb{R}^m \to \mathbb{R}^k$ *be one of the random mappings as in Lemma 3.3, then*

$$\mathsf{Prob}\big[T(b) \notin \mathsf{cone}\{T(a_1), \ldots, T(a_n)\}\big] \geq 1 - 4(n+1)e^{-\mathcal{C}(\varepsilon^2 - \varepsilon^3)k}$$

for some constant $\mathcal{C}$ *(independent of* m, n, k, ε*).*

Proof. Let A be the event that both $(1-\varepsilon)\|c-x\|^2 \leq \|T(c-x)\|^2 \leq (1+\varepsilon)\|c-x\|^2$ and $(1-\varepsilon)\|c+x\|^2 \leq \|T(c+x)\|^2 \leq (1+\varepsilon)\|c+x\|^2$ hold for all $x \in \{b, a_1, \ldots, a_n\}$. By Lemma 3.3, we have $\mathsf{Prob}(A) \geq 1 - 4(n+1)e^{-c(\varepsilon^2-\varepsilon^3)k}$. For any random mapping T such that A occurs, we have

$$\begin{aligned} \langle T(c), T(b) \rangle &= \frac{1}{4}(\|T(c+b)\|^2 - \|T(c-b\|^2) \\ &\leq \frac{1}{4}(\|c+b\|^2 - \|c-b\|^2) + \frac{\varepsilon}{4}(\|c+b\|^2 + \|c-b\|^2) \\ &= c^T b + \varepsilon < 0 \end{aligned}$$

and, for all $i = 1, \dots, n$, we can similarly derive $c^T a_i - \varepsilon \geq 0$ from $\langle T(c), T(a_i) \rangle$. Therefore, by Thm. 4.1, $T(b) \notin \mathsf{cone}\{T(a_1), \dots, T(a_n)\}$. □ □

From this proposition, it follows that the larger ε will provide us a better probability. The largest ε can be found by solving the following optimization problem.

Separating Coefficient Problem (SCP).
Given $b \notin \mathsf{cone}\,\{a_1, \dots, a_n\}$, find $\varepsilon = \max\limits_{c,\varepsilon} \{\varepsilon | \, \varepsilon \geq 0, c^T b \leq -\varepsilon, c^T a_i \geq \varepsilon\}$.

Note that ε can be extremely small when the cone C generated by $a_1, \dots, a_n$ is almost non-pointed, i.e. the convex hull of $a_1, \dots, a_n$ contains a point close to 0. Indeed, for any convex combination $x = \sum_i \lambda_i a_i$ with $\sum_i \lambda_i = 1$ of a_i's, we have:

$$\|x\| = \|x\| \, \|c\| \geq c^T x = \sum_{i=1}^{n} \lambda_i c^T a_i \geq \sum_{i=1}^{n} \lambda_i \varepsilon = \varepsilon.$$

Therefore, $\varepsilon \leq \min\{\|x\| \mid x \in \mathsf{conv}\{a_1, \dots, a_n\}\}$.

5 Projection of minimum distance

In this section we show that if the distance between a point x and a closed set is positive, it remains positive with high probability after applying a random projection. First, we consider the following problem.

Convex Hull Membership (CHM).
Given $b, a_1, \dots, a_n \in \mathbb{R}^m$, decide whether $b \in \mathsf{conv}\{a_1, \dots, a_n\}$.

5.1 Proposition
Given $a_1, \dots, a_n \in \mathbb{R}^m$, let $C = \mathsf{conv}\{a_1, \dots, a_n\}$, $b \in \mathbb{R}^m$ such that $b \notin C$, $d = \min\limits_{x \in C} \|b - x\|$ and $D = \max\limits_{1 \leq i \leq n} \|b - a_i\|$. Let $T : \mathbb{R}^m \to \mathbb{R}^k$ be a random mapping as in Lemma 3.2. Then

$$\mathsf{Prob}\big[T(b) \notin T(C)\big] \geq 1 - 2n^2 e^{-\mathcal{C}(\varepsilon^2 - \varepsilon^3)k} \tag{7}$$

for some constant $\mathcal{C}$ (independent of m, n, k, d, D) and $\varepsilon < \frac{d^2}{D^2}$.

Proof. Let S_ε be the event that both $(1-\varepsilon)\|x-y\|^2 \leq \|T(x-y)\|^2 \leq (1+\varepsilon)\|x-y\|^2$ and $(1-\varepsilon)\|x+y\|^2 \leq \|T(x+y)\|^2 \leq (1+\varepsilon)\|x+y\|^2$ hold for all $x, y \in \{0, b - a_1, \dots, b - a_n\}$. Assume S_ε occurs. Then for all real $\lambda_i \geq 0$ with $\sum\limits_{i=1}^{n} \lambda_i = 1$, we have:

$$
\begin{aligned}
& \|T(b) - \sum_{i=1}^{n} \lambda_i T(a_i)\|^2 = \|\sum_{i=1}^{n} \lambda_i T(b - a_i)\|^2 \\
= \; & \sum_{i=1}^{n} \lambda_i^2 \|T(b-a_i)\|^2 + 2 \sum_{1 \leq i < j \leq n} \lambda_i \lambda_j \langle T(b-a_i), T(b-a_j) \rangle \\
= \; & \sum_{i=1}^{n} \lambda_i^2 \|T(b-a_i)\|^2 + \frac{1}{2} \sum_{1 \leq i < j \leq n} \lambda_i \lambda_j \Big(\|T(b-a_i+b-a_j)\|^2 - \|T(a_i - a_j)\|^2 \Big) \\
\geq \; & (1-\varepsilon) \sum_{i=1}^{n} \lambda_i^2 \|b-a_i\|^2 + \frac{1}{2} \sum_{1 \leq i < j \leq n} \lambda_i \lambda_j \Big((1-\varepsilon) \big\|b-a_i+b-a_j\big\|^2 - (1+\varepsilon)\|a_i - a_j\|^2 \Big) \\
= \; & \|b - \sum_{i=1}^{n} \lambda_i a_i\|^2 - \varepsilon \Big(\sum_{i=1}^{n} \lambda_i^2 \|b-a_i\|^2 + \frac{1}{2} \sum_{1 \leq i < j \leq n} \lambda_i \lambda_j (\|b-a_i+b-a_j\|^2 + \|a_i - a_j\|^2) \Big) \\
= \; & \|b - \sum_{i=1}^{n} \lambda_i a_i\|^2 - \varepsilon \Big(\sum_{i=1}^{n} \lambda_i^2 \|b-a_i\|^2 + \sum_{1 \leq i < j \leq n} \lambda_i \lambda_j (\|b-a_i\|^2 + \|b-a_j\|^2) \Big).
\end{aligned}
$$

From the definitions of d and D, we have:

$$
\|T(b) - \sum_{i=1}^{n} \lambda_i T(a_i)\|^2 \geq d^2 - \varepsilon D^2 \Big(\sum_{i=1}^{n} \lambda_i^2 + 2 \sum_{1 \leq i < j \leq n} \lambda_i \lambda_j \Big) = d^2 - \varepsilon D^2 \Big(\sum_{i=1}^{n} \lambda_i \Big)^2 =
$$

$= d^2 - \varepsilon D^2 > 0$ due to the choice of $\varepsilon < \frac{d^2}{D^2}$. Now, since $\|T(b) - \sum\limits_{i=1}^{n} \lambda_i T(a_i)\|^2 > 0$ for all choices of λ, it follows that $T(b) \notin \mathsf{conv}\{T(a_1), \ldots, T(a_n)\}$. In summary, if S_ε occurs, then $T(b) \notin \mathsf{conv}\{T(a_1), \ldots, T(a_n)\}$. Thus, by Lemma 3.3 and the union bound,

$$
\mathsf{Prob}(T(b) \notin T(C)) \geq \mathsf{Prob}(S_\varepsilon) \geq 1 - 2\big(n + 2\binom{n}{2}\big) e^{-\mathcal{C}(\varepsilon^2 - \varepsilon^3)k} = 1 - 2n^2 e^{-\mathcal{C}(\varepsilon^2 - \varepsilon^3)k}
$$

for some constant $\mathcal{C} > 0$. □ □

In order to deal with the CM problem, we consider the A-norm of $x \in \mathsf{cone}\{a_1, \ldots, a_n\}$ as $\|x\|_A = \min\big\{ \sum\limits_{i=1}^{n} \lambda_i \;\big|\; \lambda \geq 0 \wedge x = \sum\limits_{i=1}^{n} \lambda_i a_i \big\}$. For each $x \in \mathsf{cone}\{a_1, \ldots, a_n\}$, we say that $\lambda \in \mathbb{R}^n_+$ yields a *minimal A-representation* of x if and only if $\sum\limits_{i=1}^{n} \lambda_i = \|x\|_A$. We define $\mu_A = \max\{\|x\|_A \mid x \in \mathsf{cone}\{a_1, \ldots, a_n\} \wedge \|x\| \leq 1\}$; then, for all $x \in \mathsf{cone}\{a_1, \ldots, a_n\}$, $\|x\| \leq \|x\|_A \leq \mu_A \|x\|$. In particular $\mu_A \geq 1$. Note that μ_A serves as a measure of worst-case distortion when we move from Euclidean to $\|\cdot\|_A$ norm.

5.2 Theorem
Given $b, a_1, \ldots, a_n \in \mathbb{R}^m$ of norms 1 such that $b \notin C = \mathsf{cone}\{a_1, \ldots, a_n\}$, let $d = \min\limits_{x \in C} \|b - x\|$ and $T : \mathbb{R}^m \to \mathbb{R}^k$ be one of the random mappings in Lemma 3.3. Then:

$$
\mathsf{Prob}(\, T(b) \notin \mathsf{cone}\{T(a_1), \ldots, T(a_n)\} \,) \geq 1 - 2n(n+1) e^{-\mathcal{C}(\varepsilon^2 - \varepsilon^3)k} \tag{8}
$$

for some constant $\mathcal{C}$ (independent of m, n, k, d), in which $\varepsilon = \frac{d^2}{\mu_A^2 + 2\|p\|\mu_A + 1}$.

Proof. For any $0 < \varepsilon < 1$, let S_ε be the event that both $(1-\varepsilon)\|x-y\|^2 \leq \|T(x-y)\|^2 \leq (1+\varepsilon)\|x-y\|^2$ and $(1-\varepsilon)\|x+y\|^2 \leq \|T(x+y)\|^2 \leq (1+\varepsilon)\|x+y\|^2$ hold for all $x, y \in \{b, a_1, \ldots, a_n\}$. By Lemma 3.3, we have

$$
\mathsf{Prob}(S_\varepsilon) \geq 1 - 4\binom{n+1}{2} e^{-\mathcal{C}(\varepsilon^2 - \varepsilon^3)k} = 1 - 2n(n+1) e^{-\mathcal{C}(\varepsilon^2 - \varepsilon^3)k}
$$

for some constant $\mathcal{C}$ (independent of m,n,k,d). We will prove that if S_ε occurs, then we have $T(b) \notin \mathsf{cone}\{T(a_1),\dots,T(a_n)\}$. Assume that S_ε occurs. Consider an arbitrary $x \in \mathsf{cone}\{a_1,\dots,a_n\}$ and let $\sum\limits_{i=1}^{n} \lambda_i a_i$ be the minimal A-representation of x. Then we have:

$$\begin{aligned}
&\|T(b)-T(x)\|^2 = \|T(b) - \sum_{i=1}^{n}\lambda_i T(a_i)\|^2 \\
= \;& \|T(b)\|^2 + \sum_{i=1}^{n}\lambda_i^2\|T(a_i)\|^2 - 2\sum_{i=1}^{n}\lambda_i\langle T(b),T(a_i)\rangle + 2\sum_{1\le i<j\le n}\lambda_i\lambda_j\langle T(a_i),T(a_j)\rangle \\
= \;& \|T(b)\|^2 + \sum_{i=1}^{n}\lambda_i^2\|T(a_i)\|^2 + \sum_{i=1}^{n}\frac{\lambda_i}{2}(\|T(b-a_i)\|^2 - \|T(b+a_i)\|^2) + \sum_{1\le i<j\le n}\frac{\lambda_i\lambda_j}{2}(\|T(a_i+a_j)\|^2 - \|T(a_i-a_j)\|^2) \\
\ge \;& (1-\varepsilon)\|b\|^2 + (1-\varepsilon)\sum_{i=1}^{n}\lambda_i^2\|a_i\|^2 + \sum_{i=1}^{n}\frac{\lambda_i}{2}((1-\varepsilon)\|b-a_i\|^2 - (1+\varepsilon)\|b+a_i\|^2) \\
&\qquad + \sum_{1\le i<j\le n}\frac{\lambda_i\lambda_j}{2}((1-\varepsilon)\|a_i+a_j\|^2 - (1+\varepsilon)\|a_i-a_j\|^2),
\end{aligned}$$

because of the assumption that S_ε occurs. Since $\|b\| = \|a_1\| = \dots \|a_n\| = 1$, the RHS can be written as

$$\begin{aligned}
&\|b - \sum_{i=1}^{n}\lambda_i a_i\|^2 - \varepsilon\Big(1 + \sum_{i=1}^{n}\lambda_i^2 + 2\sum_{i=1}^{n}\lambda_i + 2\sum_{j\ne i}\lambda_i\lambda_j\Big) \\
= \;& \|b - \sum_{i=1}^{n}\lambda_i a_i\|^2 - \varepsilon\big(1 + \sum_{i=1}^{n}\lambda_i\big)^2 \\
= \;& \|b - x\|^2 - \varepsilon\big(1 + \|x\|_A\big)^2
\end{aligned}$$

Denote by $\alpha = \|x\|$ and let p be the projection of b to $\mathsf{cone}\{a_1,\dots,a_n\}$, which implies $\|b-p\| = \min\{\|b-x\| \mid x \in \mathsf{cone}\{a_1,\dots,a_n\}\}$.

Claim. For all b,x,α,p given above, we have $\|b-x\|^2 \ge \alpha^2 - 2\alpha\|p\| + 1$.

By this claim (proved later), we have:

$$\begin{aligned}
\|T(b)-T(x)\|^2 \;&>\; \alpha^2 - 2\alpha\|p\| + 1 - \varepsilon\big(1+\|x\|_A\big)^2 \\
\ge \alpha^2 - 2\alpha\|p\| + 1 - \varepsilon\big(1+\mu_A\alpha\big)^2 \;&=\; \big(1-\varepsilon\mu_A^2\big)\alpha^2 - 2\big(\|p\| + \varepsilon\mu_A\big)\alpha + (1-\varepsilon).
\end{aligned}$$

The last expression can be viewed as a quadratic function with respect to α. We will prove this function is nonnegative for all $\alpha \in \mathbb{R}$. This is equivalent to

$$\begin{aligned}
&\big(\|p\| + \varepsilon\mu_A\big)^2 - \big(1-\varepsilon\mu_A^2\big)(1-\varepsilon) \le 0 \\
\Leftrightarrow\;& \big(\mu_A^2 + 2\|p\|\mu_A + 1\big)\varepsilon \le 1 - \|p\|^2 \\
\Leftrightarrow\;& \varepsilon \le \frac{1-\|p\|^2}{\mu_A^2 + 2\|p\|\mu_A + 1} = \frac{d^2}{\mu_A^2 + 2\|p\|\mu_A + 1},
\end{aligned}$$

which holds for the choice of ε as in the hypothesis. In summary, if the event S_ε occurs, then $\|T(b) - T(x)\|^2 > 0$ for all $x \in \mathsf{cone}\{a_1,\dots,a_n\}$, i.e. $T(x) \notin \mathsf{cone}\{T(a_1),\dots,T(a_n)\}$. Thus,

$$\mathsf{Prob}(T(b) \notin TC) \ge \mathsf{Prob}(S_\varepsilon) \ge 1 - 2n(n+1)e^{-c(\varepsilon^2-\varepsilon^3)k}$$

as claimed. □

Proof of the claim. If $x = 0$ then the claim is trivially true, since $\|b-x\|^2 = \|b\|^2 = 1 = \alpha^2 - 2\alpha\|p\| + 1$. Hence we assume $x \ne 0$. First consider the case $p \ne 0$. By Pythagoras' theorem, we must have

$d^2 = 1 - \|p\|^2$. We denote by $z = \frac{\|p\|}{\alpha}x$, then $\|z\| = \|p\|$. Set $\delta = \frac{\alpha}{\|p\|}$, we have

$$\begin{aligned}\|b-x\|^2 &= \|b-\delta z\|^2\\ &= (1-\delta)\|b\|^2 + (\delta^2-\delta)\|z\|^2 + \delta\|b-z\|^2\\ &= (1-\delta) + (\delta^2-\delta)\|p\|^2 + \delta\|b-z\|^2\\ &\geq (1-\delta) + (\delta^2-\delta)\|p\|^2 + \delta d^2\\ &= (1-\delta) + (\delta^2-\delta)\|p\|^2 + \delta(1-\|p\|^2)\\ &= \delta^2\|p\|^2 - 2\delta\|p\|^2 + 1 = \alpha^2 - 2\alpha\|p\| + 1.\end{aligned}$$

Next, we consider the case $p = 0$. In this case we have $b^T(x) \leq 0$ for all $x \in \mathsf{cone}\{a_1, \ldots, a_n\}$. Indeed, for an arbitrary $\delta > 0$,

$$0 \leq \frac{1}{\delta}(\|b-\delta x\|^2 - 1) = \frac{1}{\delta}(1 + \delta^2\|x\|^2 - 2\delta b^T x - 1) = \delta\|x\|^2 - 2b^T x$$

which tends to $-2b^T x$ when $\delta \to 0^+$. Therefore

$$\|b-x\|^2 \quad = \quad 1 - 2b^T x + \|x\|^2 \geq \|x\|^2 + 1 = \alpha^2 - 2\alpha\|p\| + 1,$$

which proves the claim. □ □

6 Computational results

Let T be the random projector, A the constraint matrix, b the RHS vector, and X either $\mathbb{R}^n_+$ in the case of LP and $\mathbb{Z}^n_+$ in the case of IP. We solve $Ax = b \wedge x \in X$ and $T(A)x = T(b) \wedge x \in X$ to compare accuracy and performance. In these results, A is dense. We generate (A, b) componentwise from three distributions: uniform on $[0, 1]$, exponential, gamma. For T, we only test the best-known type of projector matrix $T(y) = Py$, namely P is a random $k \times m$ matrix each component of which is independently drawn from a normal $\mathcal{N}(0, \frac{1}{\sqrt{k}})$ distribution. All problems were solved using CPLEX 12.6 on an Intel i7 2.70GHz CPU with 16.0 GB RAM. All the computational experiments were carried out in JuMP [5].

Because accuracy is guaranteed for feasible instances by Lemma 3.5 (i), we only test *infeasible* LP and IP feasibility instances. For every given size $m \times n$ of the constraint matrix, we generate 10 different instances, each of which is projected using 100 randomly generated projectors P. For each size, we compute the percentage of success, defined as an infeasible original problem being reduced to an infeasible projected problem. Performance is evaluated by recording the average user CPU time taken by CPLEX to solve the original problem, and, comparatively, the time taken to perform the matrix multiplication PA plus the time taken by CPLEX to solve the projected problem.

In the above computational results, we only report the actual solver execution time (in the case of the original problem) and matrix multiplication plus solver execution (in the case of the projected problem). Lastly, although Tables 1 tell a successful story, we obtained less satisfactory results with other distributions. Sparse instances yielded accurate but poorly performant results. So far, this seems to be a good practical method for dense LP/IP.

References

[1] Christos Boutsidis, Anastasios Zouzias, and Petros Drineas. Random projections for k-means clustering. In *Advances in Neural Information Processing Systems*, pages 298–306, 2010.

Table 1: LP: above, IP: below. *Acc.*: accuracy (% feas./infas. agreement), *Orig.*: original (CPU), *Proj.*: projected instances (CPU).

		Uniform			Exponential			Gamma		
m	n	Acc.	Orig.	Proj.	Acc.	Orig.	Proj.	Acc.	Orig.	Proj.
600	1000	99.5%	1.57s	0.12s	93.7%	1.66s	0.12s	94.6%	1.64s	0.11s
700	1000	99.5%	2.39s	0.12s	92.8%	2.19s	0.12s	93.1%	2.15s	0.11s
800	1000	99.5%	2.55s	0.11s	95.0%	2.91s	0.11s	97.3%	2.78 s	0.11s
900	1000	99.6%	3.49s	0.12s	96.1%	3.65s	0.13s	97.0%	3.57s	0.13s
1000	1500	99.5%	16.54s	0.20s	93.0%	18.10s	0.20s	91.2%	17.58s	0.20s
1200	1500	99.6%	22.46s	0.23s	95.7%	22.46s	0.20s	95.7%	22.58s	0.22s
1400	1500	100%	31.08s	0.24s	93.2%	35.24s	0.26s	95.0%	31.06s	0.23s
1500	2000	99.4%	48.89s	0.30s	91.3%	51.23s	0.31s	90.1%	51.08s	0.31
1600	2000	99.8%	58.36s	0.34s	93.8%	58.87s	0.34s	95.9%	60.35s	0.358s
500	800	100%	20.32s	4.15s	100%	4.69s	10.56s	100%	4.25s	8.11s
600	800	100%	26.41s	4.22s	100%	6.08s	10.45s	100%	5.96s	8.27s
700	800	100%	38.68s	4.19s	100%	8.25s	10.67s	100%	7.93s	10.28s
600	1000	100%	51.20s	7.84s	100%	10.31s	8.47s	100%	8.78s	6.90s
700	1000	100%	73.73s	7.86s	100%	12.56s	10.91s	100%	9.29s	8.43s
800	1000	100%	117.8s	8.74s	100%	14.11s	10.71s	100%	12.29s	7.58s
900	1000	100%	130.1s	7.50s	100%	15.58s	10.75s	100%	15.06s	7.65s
1000	1500	100%	275.8s	8.84s	100%	38.57s	22.62s	100%	35.70s	8.74s

[2] P. Indyk and A. Naor. Nearest neighbor preserving embeddings. *ACM Transactions on Algorithms*, 3(3):Art. 31, 2007.

[3] W. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In G. Hedlund, editor, *Conference in Modern Analysis and Probability*, volume 26 of *Contemporary Mathematics*, pages 189–206, Providence, 1984. American Mathematical Society.

[4] M. Ledoux. *The concentration of measure phenomenon.* Number 89 in Mathematical Surveys and Monographs. American Mathematical Society, Providence, 2005.

[5] M. Lubin and I. Dunning. Computing in operations research using julia. *INFORMS Journal on Computing*, 27(2):238–248, 2015.

[6] Mert Pilanci and Martin J Wainwright. Randomized sketches of convex programs with sharp guarantees. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 921–925. IEEE, 2014.